Information Sciences 408 (2017) 129–145

Contents lists available at ScienceDirect

Information Sciences

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

Axiomatic generalization of the membership degree weighting function for fuzzy C means clustering: Theoretical development and convergence analysis

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

[a] Stat-Math Unit, Indian Statistical Institute, 203 B.T.Road, Kolkata 700108, West Bengal, India
[b] Electronics and Communication Sciences Unit, Indian Statistical Institute, 203 B.T.Road, Kolkata 700108, West Bengal, India

ARTICLE INFO

Article history:
Received 11 April 2016
Revised 1 March 2017
Accepted 21 March 2017
Available online 10 April 2017

Keywords:
Fuzzy clustering
Generalized weighting function
CMWF
Convergence analysis

ABSTRACT

For decades practitioners have been using the center-based partitional clustering algorithms like Fuzzy C Means (FCM), which rely on minimizing an objective function, comprising of an appropriately weighted sum of distances of each data point from the cluster representatives. Numerous generalizations in terms of choice of the distance function have been introduced for FCM since its inception. In a stark contrast to this fact, to the best of our knowledge, there has not been any significant effort to address the issue of convergence of the algorithm under the structured generalization of the weighting function. Here, by generalization we mean replacing the conventional weighting function ${u_{ij}}^m$ (where u_{ij} indicates the membership of data $\mathbf{x}_i$ to cluster C_j and m is the real-valued fuzzifier with $1 \leq m < \infty$) with a more general function $g(u_{ij})$ which can assume a wide variety of forms under some mild assumptions. In this article, for the first time, we present a novel axiomatic development of the general weighting function based on the membership degrees. We formulate the fuzzy clustering problem along with the intuitive justification of the technicalities behind the axiomatic approach. We develop an Alternative Optimization (AO) procedure to solve the main clustering problem by solving the partial optimization problems in an iterative way. The novelty of the article lies in the development of an axiomatic generalization of the weighting function of FCM, formulation of an AO algorithm for the fuzzy clustering problem with the extension to this general class of weighting functions, and a detailed convergence analysis of the proposed algorithm.

1. Introduction

Clustering is the unsupervised grouping or partitioning of a collection of patterns based on some similarity or dissimilarity measure and without any prior knowledge of the class labels. Each pattern or data point is represented as a vector of measurements or a point in some multi-dimensional space. Most often the unlabeled patterns are classified in such a way, that the patterns belonging to the same cluster (same label) are more similar to each other than they are to the patterns belonging to the other clusters [47]. A clustering algorithm is expected to learn the correct labels of the patterns for well-defined subgroups in the data. The absence of any kind of training dataset makes this method different from that of

[*] Corresponding author.
E-mail addresses: arkajyotisaha93@gmail.com (A. Saha), swagatam.das@isical.ac.in, swagatamdas19@yahoo.co.in (S. Das).

http://dx.doi.org/10.1016/j.ins.2017.03.024

supervised learning. Clustering has been extensively studied across varied disciplines like decision-making, exploratory data analysis, taxonomy, data mining, information retrieval, social network analysis, and image interpretation [26,27].

Data clustering algorithms come in two primary categories: hierarchical and partitional [27]. This paper focuses on the partitional clustering algorithms which directly organize the data into a set of clusters by optimizing some criterion function(e.g. a squared-error function indicating the intra-cluster spread). Both the local and global structures in the data can be reflected by the criterion function. The local structures may be emphasized by assigning clusters to the weighted mean of the data points. On the other hand, the global criteria usually involve minimizing some measure of dissimilarity in the objects within each cluster while maximizing the dissimilarity across different clusters. Partitional clustering algorithms can be either hard or fuzzy. In fuzzy clustering, clusters are treated as fuzzy sets and each pattern has a membership (in [0, 1]) of belonging to each cluster, rather than completely belonging to any one cluster as is the case for hard clustering. A cluster is represented by a vector prototype or quantizer, also commonly known as a cluster representative in the literature. Each data point is assigned a membership degree in each cluster, based on their similarity with the corresponding cluster representatives. FCM and its variants try to shift the representatives towards the actual centers of the regions (clusters) with a high aggregation of data points by minimizing the appropriately defined cost functions [46].

The formulation of the objective function of a fuzzy clustering algorithm consists of two parts. First one is the choice of the distance function, i.e. a way of measuring the level of dissimilarity between a pair of points. The second one is the weight or the importance of the dissimilarity between the ith point and the cluster representative of the jth cluster in the objective function. This weight or importance is expressed in terms of a suitable weighting function of the membership degree of the ith data point to the jth cluster; the higher the membership, higher is the weight.

Fuzzy C-Means (FCM) is the most representative fuzzy partitional clustering algorithm till date. It was first introduced by Dunn [14] and subsequently generalized by Bezdek [6] by generalizing the fuzzifier value in the range $[1,\infty)$. Due to its natural ability to handle overlapping clusters, a lot of research efforts have been put into FCM and similar clustering algorithms. There have been a plethora of literature on the generalization of the conventional FCM algorithm in terms of the dissimilarity measure or the distance function used [4,9,22,30,34,35,42,45]. Convergence properties of the Alternative Optimization (AO) of FCM [7] and its variants have been extensively investigated for various distance functions including the conventional Euclidean distance. Some notable works in this direction can be found in [19,21,23]. Recently Fazendeiro and de Oliveira [15] reported the convergence analysis of an FCM variant (called FCM with the focal point) for an observer-biased clustering framework which aims at generating a reasonable set of clusters corresponding to different levels of granularity in different regions of the data space. Chaomurilige et al. [11] analytically investigated the optimal parameter selection of the GK (Gustafson Kessel) clustering algorithm, which is a variant of FCM derived by replacing the Euclidean distance with the Mahalanobis distance. Several cluster validity functions have also been investigated in conjunction with FCM to estimate the number of clusters, see, for example, recent works like [31,44]. Recently Saha and Das [41] reformulated the FCM algorithm with a separable geometric distance measure and theoretically demonstrated the robustness of the same towards noise feature perturbations.

In contrast to the volume of analytical studies on the properties of FCM with various distance functions, there exists very few articles addressing the generalization of the membership degree weighting function. Generalization, in this context, essentially means using a more flexible function $g\ (u_{ij})$ of the membership degree u_{ij} of the ith data point x_i to belong to the jth cluster C_j. The weighting function $g(u_{ij})$ is used to find out the weight of the distance of data point $\mathbf{x}_i$ from cluster prototype $\mathbf{z}_j$ in the objective function of FCM. We will present this notion more formally in Section 2. Note that the conventional form of FCM assumes $g(u_{ij}) = {u_{ij}}^m$, where m is the so-called fuzzifier and $1 \leq m < \infty$. m determines the degree of overlap among the clusters and also helps in convexifying the objective function. There are some works proposing and justifying a few intuitive variations of the membership degree weighting function in FCM [28,29], but as far as our knowledge is concerned, no convergence analysis is available for those variations of FCM. Yu et al. [49] took an analytical approach to select the proper fuzzifier m for conventional FCM by exploiting a relation between the fixed points of FCM and the dataset itself. Ganguly et al. [18] used a multi-dimensional membership vector for each data point replacing the traditional, scalar membership value defined in the original FCM. However, they preserved the conventional form of the membership degree weighting function.

Thus, while use of the generalized dissimilarity measures in FCM and the corresponding convergence analyses attracted so much attention in recent past, why is it exactly the opposite case, regarding the weighting function? What might be the possible reasons behind the fact that, after Bezdek [6] proposed a generalization of [14] in terms of the fuzzifier, there does not exist any significant literature regarding further generalization of the weighing function while preserving the convergence properties of FCM? We point out a couple of reasons, that may provide some insight into the matter:

1. First, development of various distance measures (between two points in a space or between two probability measures), is a question of independent interest and has a plethora of literature of its own [10,12,13,33]. Hence, a generalization of clustering process in terms of dissimilarity measure is always backed by the theory behind the development of that distance measure in the first place. On the other hand, there exists no such literature regarding the generalization of the weighting function, because, this is an issue, only relevant for the fuzzy clustering problems. Hence, there does not exist axiomatic approach towards the development of generalized membership degree weighting functions.
2. As far as the convergence analysis is concerned, there is a major road block in the generalization of a fuzzy clustering algorithm in terms of the weighting function. Whenever a new distance function is developed, development of its center

or *population minimizer* [39] i.e. the minimizer of the weighted sum of a group of points from a particular point, is an intrinsic part of it [3,5,37,38]. This plays a pivotal role in the convergence analysis of fuzzy clustering algorithms with general distance measures. For the reason mentioned above, there is no such existing literature regarding the weighting function, as the optimization problem is not even defined in any framework, other than fuzzy clustering.

In this work, we analytically investigate the use other permissible forms of the membership degree weighting function so that the algorithm may possess some additional degrees of freedom (in terms of choosing this function) besides enjoying the strong local convergence properties of FCM. The contributions of the article can be summarized in the following way:

1. We develop an axiomatic approach of generalization of the membership degree weighting function for FCM. We provide intuitive justification of the axiomatic development. We also discuss some common examples of this general class of weighting functions, which includes the common function (u_{ij}^m) already existing in the literature.
2. We propose a general class of fuzzy clustering problems with the extension of the generalized weighting functions and present an AO based clustering algorithm for the clustering purpose. Here, we also demonstrate that the previous attempts of intuitive generalization of this function [28,29] can be derived as special cases of the proposed generalization.
3. We discuss the issues like the existence and the uniqueness of solutions of the optimization problems underlying the AO algorithm and carry out a full-fledged convergence analysis to obtain a similar convergence property as that of the conventional FCM.

Organization of the paper is in order. In Section 2, we develop an axiomatic approach of extension of the weighting function, along with the motivation of such generalization. We provide some generic examples of the generalized membership degree weight functions. We also formulate the clustering problem along with intuitive justification of the axiomatic development of the general class of weighting functions and develop an AO algorithm to solve it. In Section 3, we carry out a full-fledged convergence analysis of the proposed algorithm. In Section 4, we conclude by presenting a brief summary of the findings and discussing the probable future extensions of the development presented in this article.

2. Clustering with generalized weighting function

In this section, we present the motivation of the generalization and the fundamentals of the axiomatic approach. Next we develop the fuzzy clustering procedure with the general class of membership degree weighting functions.

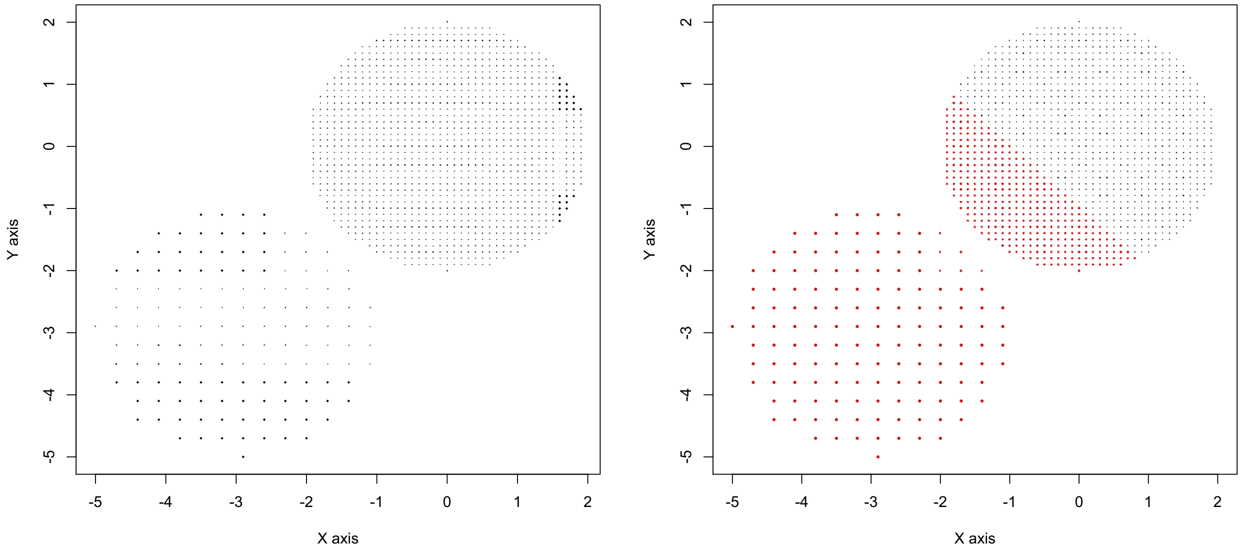
2.1. Motivation for generalization

The choice of distance measure in the FCM determines the shapes of the clusters that will be detected by the FCM algorithm (e.g. squared Euclidean - spherical, Mahalanobis distance - elliptical etc.). In this paper, we are concerned with the contribution of a specific data point to the objective function, given the membership degrees u_{ij}. In the classical k-means algorithm, when we optimize the objective function with respect to the membership matrix (keeping the cluster prototypes fixed), we basically solve a locally defined optimization problem. Hence, the solutions are obtained at the boundary points i.e. $\{0, 1\}$. In presence of ambiguous data and in-distinctive cluster boundaries, assignment of a data point completely to one single cluster is not a very realistic choice to make. In that case, a point is assigned to one or more clusters with varying degree of membership. In order to achieve that, the optimization task with respect to the membership matrix was turned into a convex optimization problem by introducing the notion of a fuzzifier. Here, the contribution of a data point (with membership u_{ij}) in the objective function, was related to the membership degree, through the function $u_{ij} \to {u_{ij}}^m, m > 1$.

Here, to the best of our knowledge, no justification was provided for the choice of ${u_{ij}}^m$ as the weighting function to make the optimization problem convex, when it was first proposed [6,14]. A detailed exploration of the significance of fuzzifier and its implications was carried out in [28,29]. The authors established a relationship between the value of the fuzzifier m and the smoothness of transition from a high membership degree in one cluster to the other. Higher the fuzzifier, smoother the transition i.e. as $m \to 1$, we obtain crispier partition and as $m \to \infty$, we get equal membership degrees to all clusters with all cluster representatives or centers converging to the global mean of the data set.

From the membership update scheme of the conventional FCM algorithm, it is clearly evident that a data point has non-zero membership in *all* clusters, unless a data point actually coincides with one of the cluster representatives. This implication produces somehow counter-intuitive results, as this will indicate that, no matter how far a data is from a cluster representative and how well it is represented by other clusters, it will have non-zero membership in all the clusters (except for the rare case, where the data point actually coincides with a cluster representative) [29]. The disadvantages of such membership allocation can lead to the following undesired results:

1. In the case of clustering a data where clusters have unequal data densities, the cluster(s) with the higher data density attracts the cluster representative(s) of the other cluster(s) with lower data density. This leads to counter-intuitive clustering performances in the conventional FCM framework. In Fig. 1, it is demonstrated that in the aforementioned scenario, the representative of the cluster with lower data density is attracted towards the representative of the cluster with higher density, which leads to dubious results.



(a) (b)

Fig. 1. (a) Original Dataset, (b) Clustering performance with u_{ij}^m as the membership degree weighting function.

2. We consider a situation, where, we have already clustered a data set. Now we add an outlier in the data and recluster the extended data set with one more cluster as the original dataset. In an ideal scenario, the outlier should be clustered separately in the new cluster, and the earlier cluster structure should have remained as it was. But this is not the actual scenario, that takes place in case of clustering with conventional FCM [29].

Hence, the choice of ${u_{ij}}^m$ as the weighting function does fail in certain cases, the aforementioned two being the most prominent of them.

With the exploration of the shortcomings of the conventional fuzzifier based FCM, there were some attempts to address the problem by proposing some specific weighting functions [28,29]. This approach had the following problems:

1. To the best of our knowledge, no convergence analysis corresponding to the proposed membership degree weighting functions, are available in the literature.
2. Though some justifications were provided to chalk out the basic properties (increasing and convex) of any such function, but the choice was one single function, that satisfied the necessary condition. It hardly represented the class of all possible weighting functions.

In the modern day of huge data influx, where, the existence of outliers in the data or unevenly distributed data is more of a norm than an exception, it can be effective to generalize the notion of the weighting function in an axiomatic approach as well as to investigate the corresponding convergence properties. The present article attempts to make a humble contribution in this context.

2.2. Generalized membership-degree weighting function

We define a general class of weighting functions, based on which we will develop the fuzzy clustering algorithm. We call this general class of functions as Consistent Membership-degree Weighting Function (CMWF). For the sake of notational clarity, we will denote the membership of any ith data point to cluster C_j or u_{ij} simply by a subscript-free scalar variable u. The CMWF class of functions can be formally defined in the following way.

Definition 1. A function $g:(0,\,1)\to[0,\,1]$ is called a consistent membership degree weighting function, if it satisfies the following criteria:

1. g is strictly increasing on (0, 1).
2. $$\lim_{u\to 0^+} g(u)=0 \tag{1a}$$
$$\lim_{u\to 1^-} g(u)=1 \tag{1b}$$
3. g is strictly convex function on (0, 1).

4. g is differentiable on (0, 1).

The significance and intuitions behind the technical conditions on CMWF, will be discussed in detail in Section 2.5. As a brief summary of the motivation, it can be said, the first condition arises from the fundamental notion of the weight function, whereas the second condition is necessary for the sake of a well-defined clustering problem, maintaining consistency with the general notion of weighting functions. The third condition also arises from the basic notion but is instrumental in convergence analysis. Here, we emphasis that the first assumption implies that, g is differentiable almost on everywhere on (0, 1) (i.e. non-differentiable at almost countably many points). The fourth condition strengthens the differentiability property of g by making it everywhere differentiable on (0, 1), which plays a pivotal role in the development of the convergence analysis.

2.3. Generic examples of CMWF

In this subsection, we discuss some generic examples of members of interest in the CMWF family.

Example 1. The first example and the corresponding clustering algorithm (will be defined in Section 2.5) was first proposed in [14] with the weight function $u \to u^2$. Later [6] generalized the weighting function by using the mapping $u \to u^m; m > 1$, where the m is conventionally called the fuzzifier. To the best of our knowledge, the class of weighting functions generated by varying the value of m is the only such class available in literature and used for fuzzy clustering till date. Recently, a theoretical approach towards the choice of fuzzifier m for a particular dataset under consideration was developed in literature [24]. This general class of weighting functions is generated by the following member of CMWF:

$$g_{1,m}(u) = u^m, \ m > 1.$$

Now, we will show that this belongs to the class of CMWFs. First, this is a strictly increasing, strictly convex function, which is obvious from considering the first ($g'(u) = mu^{m-1} > 0;\ \forall u \in (0,1)$) and the second derivative ($g''(u) = m(m-1)u^{m-2} > 0;\ \forall u \in (0,1)$). Now, the second condition is also trivially satisfied, as $\lim_{u\to 0^+} u^m = 0$ and $\lim_{u\to 1^-} u^m = 1$

Example 2. A second generic example of CMWF is given by the following function:

$$g_2(u) = \frac{e^u - 1}{e - 1}.$$

The aforementioned function is strictly increasing, strictly convex function from the properties of exponential function (the first derivative is positive and second derivative is always positive). The second condition also follows trivially as $\lim_{u\to 0^+} e^u = 1$ and $\lim_{u\to 1^-} e^u = e$.

Example 3. Another generic example of CMWF is given by the following function:

$$g_{3,m}(u) = \frac{e^{u^m} - 1}{e - 1}, \ m > 1.$$

The aforementioned function is strictly increasing, strictly convex function from the fact that composition of two strictly increasing strictly convex functions is again strictly convex and strictly increasing ($g_{3,m} = g_2 \circ g_{1,m}$). The second condition also follows trivially as $\lim_{x\to 0^+} e^{u^m} = 1$ and $\lim_{u\to 1^-} e^{u^m} = e$.

As far as kernel functions are concerned, they are generally centered around the origin. From the fundamental nature of kernel functions, they are usually decreasing in [0, 1]. Moreover, the kernel functions are not strictly convex, if we consider their respective whole domains. Nonetheless, we can derive a CMWF from kernel functions or shift them to make it strictly increasing and strictly convex in [0, 1]. We present an example of each of the cases in the perspective of the Gaussian kernel functions :

Example 4. First of all, we look at the following function :

$$g_4(u) = \frac{e^{\frac{u^2}{2}} - 1}{e^{\frac{1}{2}} - 1}.$$

In the aforementioned function, $e^{\frac{u^2}{2}}$ plays the functional part (which is obtained from the inverse of the functional part of a standard Gaussian density function), while the other constants are used for normalization. By construction, g_4 is differentiable on (0, 1) and satisfies the limit conditions. We also check that,

$$g_4'(u) = \frac{ue^{\frac{u^2}{2}}}{e^{\frac{1}{2}} - 1}, \Rightarrow g_4'(u) > 0; \forall u \in (0,1).$$

This proves that the function is strictly increasing on (0, 1).

$$g_4''(u) = \frac{e^{\frac{u^2}{2}} + u^2 e^{\frac{u^2}{2}}}{e^{\frac{1}{2}} - 1}, \Rightarrow g_4''(u) > 0; \forall u \in (0,1).$$

This proves that the function is strictly convex on (0, 1), which also concludes the proof of the fact that the function under consideration is a valid CMWF.

Example 5. On the other hand, we can obtain a CMWF from the Gaussian kernel, just by making a location shift as follows:

$$g_5(u) = \frac{e^{-\frac{(u-2)^2}{2}} - e^{-2}}{e^{-\frac{1}{2}} - e^{-2}}.$$

In the aforementioned function, $e^{-\frac{(u-2)^2}{2}}$ plays the functional part (which is obtained from the inverse of the functional part of a standard Gaussian density function), while the other constants are used for normalization. By construction, g_5 is differentiable on (0, 1) and satisfies the limit conditions. We also check that,

$$g_5'(u) = \frac{-(u-2)e^{\frac{(u-2)^2}{2}}}{e^{-\frac{1}{2}} - e^{-2}}, \Rightarrow g_5'(u) > 0; \forall u \in (0, 1).$$

This proves that the function is strictly increasing on (0, 1).

$$g_5''(u) = \frac{-e^{\frac{(u-2)^2}{2}} + (u-2)^2 e^{\frac{(u-2)^2}{2}}}{e^{-\frac{1}{2}} - e^{-2}}, \Rightarrow g_5''(u) > 0 \forall u \in (0, 1).$$

The positivity of $g_5''(u)$ follows from the fact that $(u-2)^2 - 1 > 0, \forall u \in (0, 1)$. This proves that the function is strictly convex on (0, 1), which also concludes the proof of the fact that the function under consideration is a valid CMWF.

2.4. Relationship with the existing generalized weighting functions

Here, we demonstrate that the previous attempts of intuitive generalization of the membership degree weighting function, can be obtained as a special case of the proposed CMWF family .

1. The weighting function proposed in [29] can be obtained as a special case of the CMWF family, with the following choice of the weighting function
$$g_{6,\beta}(u) = \frac{1-\beta}{1+\beta}u^2 + \frac{2\beta}{1+\beta}u, \ \beta > 0.$$
Now, we show that $g_{6,\beta}$ is in CMWF. First, this is a strictly increasing, strictly convex function, which is obvious from considering the first ($g'_{6,\beta}(u) = 2\frac{1-\beta}{1+\beta}u + \frac{2\beta}{1+\beta} > 0; \ \forall u \in (0, 1)$) and the second derivative ($g''_{6,\beta}(u) = 2\frac{1-\beta}{1+\beta} > 0; \ \forall u \in (0, 1)$). Now, the second condition is also trivially satisfied, as $\lim_{u \to 0^+} g_{6,\beta}(u) = 0$ and $\lim_{u \to 1^-} g_{6,\beta}(u) = \frac{1-\beta}{1+\beta} + \frac{2\beta}{1+\beta} = 1$
2. The weighting function proposed in [28] can be obtained as a special case of the CMWF family, with the following choice of the function
$$g_{7,\beta}(u) = \frac{e^{\beta u} - 1}{e^{\beta} - 1}, \ \beta > 0$$
Now, we show that $g_{7,\beta}$ is in CMWF. First, this is a strictly increasing, strictly convex function, which is obvious from considering the first ($g'_{7,\beta}(u) = \beta\frac{e^{\beta u}-1}{e^{\beta}-1} > 0; \ \forall u \in (0, 1)$) and the second derivative ($g''_{7,\beta}(u) = \beta^2\frac{e^{\beta u}-1}{e^{\beta}-1} > 0; \ \forall u \in (0, 1)$). Now, the second condition is also trivially satisfied, as $\lim_{u \to 0^+} g_{7,\beta}(u) = 0$ and $\lim_{u \to 1^-} g_{7,\beta}(u) = 1$.

2.5. Clustering problem formulation and algorithm development

In this section, we present the general class of clustering problems with CMWF. We first formulate the problem, next by means of the iterative sequence of solutions of several optimization problems, we develop an AO algorithm for solving the clustering problem.

2.5.1. Problem formulation

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}, \mathbf{x}_i \in \mathbb{R}^d, \forall i = 1, 2, \cdots, n$ be the given set of patterns, which we want to partition into c (prefixed) clusters with $2 \leq c \leq n$. Let $\mathcal{B} \subset \mathbb{R}^d$ be the convex hull of $\mathcal{X}$. The general clustering problem with any member of CMWF can be formulated as follows:

$$\mathbf{P}: \ \ \textit{minimize} \ f_g(\mathbf{U}, \mathcal{Z}) = \sum_{i=1}^{n}\sum_{j=1}^{c} \tilde{g}(u_{ij}) d(\mathbf{z}_j, \mathbf{x}_i), \tag{2}$$

where

$$d(\mathbf{z}_j, \mathbf{x}_i) = \sum_{l=1}^{d}(z_{j,l} - x_{i,l})^2, \tag{3}$$

$$\tilde{g}(u) = \begin{cases} g(u) & \text{if } u \in (0,1), \\ 0, & \text{if } u = 0, \\ 1, & \text{otherwise,} \end{cases} \tag{4}$$

subject to

$$\sum_{j=1}^{c} u_{ij} = 1, \forall i = 1, 2, \cdots, n, \tag{5a}$$

$$0 < \sum_{i=1}^{n} u_{ij} < n, \forall j = 1, 2, \cdots, c, \tag{5b}$$

$$u_{ij} \in [0, 1], \forall i = 1, 2, \cdots, n; \forall j = 1, 2, \cdots, c, \tag{5c}$$

$$\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_c\}, \mathbf{z}_j \in \mathcal{B} \subseteq \mathbb{R}^d, \forall j = 1, 2, \cdots, c; \mathcal{Z} \in \mathcal{B}^c. \tag{5d}$$

Here, the first three restrictions (5a)–(5c) are conventional restrictions on the membership matrix. The fourth one (5d) restricts the optimization to the convex hull of the patterns, which we will later show is equivalent to optimizing in the whole $\mathbb{R}^d$ corresponding to each $\mathbf{z}_j$. The distance function under consideration (3) is squared Euclidean distance. $\tilde{g}(u)$ in (4) is just the canonical extension of g from (0, 1) to [0, 1], such that $\tilde{g}$ is continuous at 0 and 1.

We impose four restrictions (strictly increasing, existence of limits at boundary points, strictly convex, and differentiable) on the weighting function. In what follows, we explain in detail, why these conditions can be obtained from intuitive notion of the weighting function. We present the mathematical and intuitive justification of them as follows:

1. **Increasing function:** Here, $h(u_{ij})$ denotes the weight of contribution of the distance between the ith data point and the cluster representative corresponding to the jth cluster. Quite naturally the higher the membership of a data point in a particular cluster, higher should be its contribution. In order to express this mathematically, we consider the problem of optimizing the objective function with respect to the membership matrix, with fixed cluster representatives. Now, if g is non increasing, in order to minimize the objective function, we will assign highest membership in the furthest cluster, which is completely counter-intuitive. Hence, we need g to be an increasing function.
2. **Convexity:** For a particular point and a specific cluster with high membership degree for the concerned point, change of weight of the corresponding distance due to a small change in membership degree, should be higher than that corresponding to a cluster, in which, the point has lower membership degree. This justifies the condition of convexity of the membership degree weighting function . In order to mathematically observe this, we consider the problem of optimizing the objective function with respect to the membership matrix, with fixed cluster representatives. We start from the membership allocation given by $\mathbf{u}_i$ (denoting the ith row of the membership matrix), where $u_{ij_0} \geqslant u_{ij}$, if $d_{i,j_0} \leqslant d_{i,j}$, $\forall j = 1, 2, \cdots, c$; and $d_{i,\,j}$ is the distance between ith data point and representative of the jth cluster. In order to further reduce the objective function value, let us assume that we decrease the u_{i,j_0} value by ϵ and increase another u_{ij} by the same amount. In that case, the change in the objective function will be,

   $$\triangle = (g(u_{ij} + \epsilon) - g(u_{ij}))d_{i,j} - (g(u_{ij_0}) - g(u_{ij_0} - \epsilon))d_{i,j_0}$$

   Since, $d_{i,j_0} \leqslant d_{i,j}$, a necessary condition under which the objective function will decrease is given by,

   $$g(u_{ij} + \epsilon) - g(u_{ij}) < g(u_{ij_0}) - g(u_{ij_0} - \epsilon).$$

   Dividing both sides by ϵ will imply that the slope is increasing and this will lead to the convexity of the weighting function. This strictly convex property of the weighting function will also play pivotal role in the convergence analysis of the proposed algorithm.
3. **Existence of limits at boundary points:** As the membership value tends to zero, the weight of the corresponding distance should tend to zero, which is ensured by (1a). On the other hand, if the membership approaches its highest value (i.e. 1), the weight of the corresponding distance in objective function should approach the maximum value, i.e. 1. The existence of these two limits are technical assumptions, needed to extend g to the closed interval [0, 1], which makes $g(0)$, $g(1)$ well defined and the extended function $\tilde{g}$ continuous on [0, 1].
4. **Differentiability:** g being convex, it is differentiable almost everywhere on (0, 1) (i.e. non-differentiable at countably many points). The technical assumption of everywhere differentiability plays pivotal role in convergence analysis.

2.5.2. The AO based algorithm development

In this section, we develop an AO based optimization procedure for the proposed clustering problem. We start with a prefixed permissible fractional error and an initial estimate of the cluster representatives. We then optimize the objective function (2) with respect to membership matrix, such that it satisfies the conditions (5a)–(5c). For the sake of notational simplicity, we define the class of the membership matrices as follows:

$$\mathcal{U}_{c,n} = \{\mathbf{U} \mid \mathbf{U} \,\text{is an}\, n \times c \,\text{real matrix and satisfies}\, (5a) - (5c)\},$$

Next, we optimize the objective function (2) with respect to cluster representatives such that the restriction (5d) is maintained. We perform this iterative process, until the difference between the values of the objective function corresponding to two consecutive iterations become smaller than the permissible fractional error. Algorithm 1 describes the fuzzy clustering with CMWF.

ALGORITHM 1: Clustering with CMWF.

```
Data: Data points, number of clusters, permissible fractional error, choice of CMWF, i.e. $g$.
Result: Membership matrix of objects in clusters, cluster representatives.
initialization;
$c \leftarrow$ number of clusters;
$\epsilon \leftarrow$ permissible fractional error;
$\mathcal{Z}^{(0)} \leftarrow$ initial clusters satisfying (5d);
$t \leftarrow 0$;
while $f_g(\mathbf{U}^{(t-1)}, \mathcal{Z}^{(t-1)}) - f_g(\mathbf{U}^{(t)}, \mathcal{Z}^{(t)}) \geqslant \epsilon f_g(\mathbf{U}^{(t-1)}, \mathcal{Z}^{(t-1)})$ do
    $\mathbf{U}^{(t+1)} = \mathrm{argmin}_{\mathbf{U} \in \mathcal{U}_{c,n}} \; f_g(\mathbf{U}, \mathcal{Z}^{(t)})$;
    $\mathcal{Z}^{(t+1)} = \mathrm{argmin}_{\mathcal{Z}_j \in \mathcal{B}^c} \; f_g(\mathbf{U}^{(t+1)}, \mathcal{Z})$;
    $t \leftarrow t + 1$;
end
```

3. Convergence analysis of clustering with CMWF

We carry out a complete convergence analysis of the fuzzy clustering with CMWF, proposed in Algorithm 1. We first address the existence and uniqueness of the solutions of partial optimization problems with respect to the membership matrix and cluster representatives. Next we develop the convergence properties of the proposed algorithm.

3.1. Existence and uniqueness of the solutions of partial optimization problems

We address the partial optimization problems in the proposed algorithm (Algorithm 1) starting with the following theorem.

Theorem 1. *For fixed $\mathbf{U}^* \in \mathcal{U}_{c,n}$, the problem $\mathbf{P}_1$: minimize $f_g(\mathbf{U}^*, \mathcal{Z})$, $\mathcal{Z} \in \mathcal{B}^c$, has a unique solution.*

Proof. The function to be minimized in problem $\mathbf{P}_1$ is a strictly convex function with respect to $\mathcal{Z}$ (from the fact that the squared Euclidean distance $d(\mathbf{y}, \mathbf{x})$ is a strictly convex function, with respect to $\mathbf{y}$) and the optimization task is carried out on a convex set, hence, there exists at most one solution.

Now, the function under consideration is also a continuous function with respect to $\mathcal{Z}$ (from the fact that the squared Euclidean distance $d(\mathbf{y}, \mathbf{x})$ is a continuous function, with respect to $\mathbf{y}$), hence, attains its maxima and minima in a compact set, which is indeed the case here ($\mathcal{B}$ being the convex hull of $\mathcal{X}$, is closed and bounded, i.e. compact). Hence, the minimization task under consideration, has atleast one solution in the feasible region.

Employing the two aforementioned statements, we guarantee the existence of unique solution of the optimization task $\mathbf{P}_1$ in the feasible region. □

Theorem 2. *Let $J_1 : \mathcal{B}^c \to \mathbb{R}, J_1(\mathcal{Z}) = f_g(\mathbf{U}^*, \mathcal{Z}); \mathbf{z}_j \in \mathcal{B}, \forall j = 1, 2, \cdots, c$, where $\mathbf{U}^* \in \mathcal{U}_{c,n}$ is fixed. Then $\mathcal{Z}^*$ is a global minimum of J_1 if and only if $\mathbf{z}_j^*$ is given by:*

$$\mathbf{z}_j^* = \left[\sum_{i=1}^{n} \tilde{g}(u_{ij})\mathbf{x}_i\right] \Big/ \left[\sum_{i=1}^{n} \tilde{g}(u_{ij})\right]; j = 1, 2, \cdots, c. \tag{6}$$

Proof. The condition in the theorem is derived as a first order necessary condition by setting the partial derivative of the objective function with respect to $\mathbf{z}_j$ equal to zero. From the strict convexity of the objective function with respect to $\mathbf{z}_j$, it follows that those conditions are also sufficient conditions to uniquely determine the minimizer. □

Theorem 3. *For fixed $\mathcal{Z}^* = \{\mathbf{z}_1^*, \mathbf{z}_2^*, \cdots, \mathbf{z}_c\}, \mathbf{z}_j^* \in \mathcal{B}^c$, the problem $\mathbf{P}_2$: minimize $f_g(\mathbf{U}, \mathcal{Z}^*), \mathbf{U} \in \mathcal{U}_{c,n}$, has a unique solution under the assumption $\Psi_i^* = \{j \mid d(\mathbf{z}_j^*, \mathbf{x}_i) = 0\} = \phi$; $\forall i = 1, 2, \cdots, n$.*

Proof. Under the assumption, the function to be minimized in problem $\mathbf{P}_2$ is a strictly convex function with respect to $\mathbf{U}$ (from the first condition of CMWF) and the optimization task is carried out on a convex set ($\mathcal{U}_{c,n}$ is a convex set), there exists at most one solution.

Now, the function under consideration is also a continuous function with respect to $\mathbf{U}$ (from the fact that being strictly increasing and strictly convex, g is continuous on (0, 1), so is $\tilde{g}$; moreover, from the way it is defined (4), $\tilde{g}$ is continuous at 0 and 1). Hence, it attains its maxima and minima in a compact set, which is indeed the case here ($\bar{\mathcal{U}}_{c,n} = \mathcal{U}_{c,n}$, hence, $\mathcal{U}_{c,n}$ is closed. Being a bounded set, it is also compact). Thus, the minimization task under consideration, has at least one solution in the feasible region.

Employing the two aforementioned statements, we guarantee the existence of unique solution of the optimization task $\mathbf{P}_2$ in the feasible region. □

Theorem 4. *Let $J_2 : \mathcal{U}_{c,n} \to \mathbb{R}, J_2(\mathbf{U}) = f_g(\mathbf{U}, \mathcal{Z}^*)$; $\mathbf{U}^* \in \mathcal{U}_{c,n}$, where $\mathbf{z}_j^* \in \mathcal{B}; \forall j = 1, 2, \cdots, c$ is fixed. Then $\mathbf{U}^*$ is a global minimum of J_2 if and only if $\mathbf{U}^*$ satisfies the following equations, $\forall i = 1, 2, \cdots, n$;*

$$\Psi_i^* = \left\{ j \mid d(\mathbf{z}_j^*, \mathbf{x}_i) = 0 \right\};$$

if $\Psi_i^* = \phi$,

$$u_{ij}^* = {v_{ij}^*}^2, \tag{7a}$$

$$\tilde{g}'({v_{ij}^*}^2) d(\mathbf{z}_j^*, \mathbf{x}_i) = \tilde{g}'({v_{ij'}^*}^2) d(\mathbf{z}_{j'}^*, \mathbf{x}_i); \ \forall j \neq j' \in \left\{1, 2, \cdots, c\right\}, \tag{7b}$$

$$\sum_{j=1}^{c} {v_{ij}^*}^2 = 1. \tag{7c}$$

If $\Psi_i^* \neq \phi$

$$u_{ij}^* = \begin{cases} \geqslant 0 \text{ with } \sum\limits_{\mathbf{z}_k^* = \mathbf{x}_i} u_{ik}^* = 1 & \text{if } j \in \Psi_i^*, \\ 0 & \text{otherwise}. \end{cases} \tag{7d}$$

Proof. Minimization of J_2 over $\mathcal{U}_{c,n}$ is a Kuhn–Tucker problem with $cn + c$ many inequality constraints given by (5a) and (5c) as well as n many equality constraints given by (5b). We can merge (5a) and (5c) by making the substitution, $u_{ij} = v_{ij}^2$, and in that case, we have the following optimization task at hand:

$$\mathbf{RP}_2 \ : \ \textit{minimize } RJ_2(\mathbf{V}) = \sum_{i=1}^{n} \sum_{j=1}^{c} \tilde{g}(v_{ij}^2) d(\mathbf{z}_j^*, \mathbf{x}_i),$$

such that

$$\sum_{j=1}^{c} v_{ij}^2 = 1 \ i = 1, 2, \cdots, n.$$

Let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_n)$ be the vector of Lagrange multipliers corresponding to the n equality constraints. Now, the Lagrangian can be written in the following way:

$$LJ_2(\mathbf{V}, \boldsymbol{\lambda}) = RJ_2(\mathbf{V}) + \sum_{i=1}^{n} \lambda_i \left[\sum_{j=1}^{c} v_{ij}^2 - 1 \right].$$

First, from the strict convexity of $\tilde{g}(x)$ on (0, 1), it follows that under the linear constraints of (5b) on $\mathbf{U}$, for some fixed i, $u_{ij}^* = 0$ for some j, if and only if $|\Psi_i^*| \neq \phi$. Hence, under the assumption $|\Psi_i^*| = \phi$, we get the desired conditions by setting the first order condition equal to 0 and by employing Theorem 3.

If $\Psi_i^* \neq \phi$, (i.e. the data point under consideration actually coincides with a cluster representative) with the membership assignment in (7d), the contribution of the ith point in the objective function will be 0. It can be trivially seen that, any other membership assignment will lead to a positive contribution, contradicting its candidature for an optimizer of the objective function.

Now, we will show that the optimizer thus obtained, actually satisfies the inequality constraints given in (5b). Assume $\exists j_0 \in$ {1, 2, ⋯, c}, such that $\sum_{i=1}^{n} u_{ij_0}^* = 0$, i.e. $u_{ij_0} = 0$; $i = 1, 2, \cdots, n$, i.e. $\Psi_i^* \neq \phi$; $\forall i$ and $j_0 \notin \cup_{i=1}^{n} \Psi_i^*$. Hence, $| \cup_{i=1}^{n} \Psi_i^* | \leqslant c - 1 < n$, but $\Psi_i^* \cap \Psi_j^* = \phi, \forall i \neq j$ (if $\Psi_i^* \cap \Psi_j^* \neq \phi$, then $d(\mathbf{x}_i, \mathbf{x}_j) = 0$, a contradiction!). Hence $| \cup_{i=1}^{n} \Psi_i^* | = \sum_{i=1}^{n} |\Psi_i^*| \geqslant n$, which is a contradiction. Hence, there does not exist such j_0, which completes the proof of the theorem. □

3.2. Convergence properties of CMWF based clustering

In this section, we use Zangwill's global convergence theorem [50] to derive the convergence results regarding the AO based clustering with CMWF, proposed in Algorithm 1. In order to update the membership matrix and cluster representatives, we define the following operators:

Definition 2. $T_{memb} : \mathcal{B}^c \to \mathcal{U}_{c,n}$, $T_{memb}(\mathcal{Z}) = \mathbf{U} = [u_{ij}]$, where, u_{ij} is given by the following rule (Theorem 4):

$$\Psi_i = \left\{ j \mid d(\mathbf{z}_j, \mathbf{x}_i) = 0 \right\};$$

if $\Psi_i = \phi$,

$$u_{ij} = {v_{ij}}^2, \tag{8a}$$

$$g'({v_{ij}}^2) d(\mathbf{z}_j, \mathbf{x}_i) = g'({v_{ij'}}^2) d(\mathbf{z}_{j'}, \mathbf{x}_i);\ \forall j \neq j' \in \left\{1, 2, \cdots, c\right\}, \tag{8b}$$

$$\sum_{j=1}^{c} {v_{ij}}^2 = 1. \tag{8c}$$

If $\Psi_i \neq \phi$

$$u_{ij} = \begin{cases} \geqslant 0 \text{ with } \sum\limits_{\mathbf{z}_k = \mathbf{x}_i} u_{ik} = 1 & \text{if } j \in \Psi_i, \\ 0 & \text{otherwise.} \end{cases} \tag{8d}$$

Definition 3.

$$T_{cent} : \mathcal{U}_{c,n} \to \mathcal{B}^c, T_{cent}(\mathbf{U}) = \mathcal{Z} = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_c),$$

where the vectors $(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_c), \mathbf{z}_j \in \mathcal{B}, j = 1, 2, \cdots, c$ are calculated as:

$$\mathbf{z}_j = \left[\sum_{i=1}^{n} \tilde{g}(u_{ij}) \mathbf{x}_i \right] \Bigg/ \left[\sum_{i=1}^{n} \tilde{g}(u_{ij}) \right]. \tag{9}$$

We first define the clustering operator in the following way:

Definition 4.

$$J : (\mathcal{U}_{c,n} \times \mathcal{B}^c) \to (\mathcal{U}_{c,n} \times \mathcal{B}^c),\ J = O_{cent} \circ O_{memb},$$

where

$$O_{memb} : \mathcal{U}_{c,n} \times \mathcal{B}^c \to \mathcal{U}_{c,n};\ O_{memb}(\mathbf{U}, \mathcal{Z}) = T_{memb}(\mathcal{Z}),$$

$$O_{cent} : \mathcal{B}^c \to \mathcal{U}_{c,n} \times \mathcal{B}^c;\ O_{cent}(\mathbf{U}) = (\mathbf{U}, T_{cent}(\mathbf{U})),$$

$$J(\mathbf{U}, \mathcal{Z}) = (T_{memb}(\mathcal{Z}), T_{cent}(T_{memb}(\mathcal{Z}))).$$

Lemma 1. *$\mathcal{U}_{c,n} \times \mathcal{B}^c$ is compact.*

Proof. $\mathcal{U}_{c,n}$ and $\mathcal{B}^c$ were proved to be compact in Theorems 1 and 3 respectively. Hence, the lemma follows. □

Definition 5. We define a subset $\mathcal{T}$ of optimal points in $\mathcal{U}_{c,n} \times \mathcal{B}^c$ as follows

$$\mathcal{T} = \left\{ \begin{array}{l} (\mathbf{U}^*, \mathcal{Z}^*) \in \mathcal{U}_{c,n} \times \mathcal{B}^c \quad | \\ f_g(\mathbf{U}^*, \mathcal{Z}^*) \leqslant f_g(\mathbf{U}, \mathcal{Z}^*), \forall \mathbf{U} \in \mathcal{U}_{c,n}, \mathbf{U} \neq \mathbf{U}^*, \\ f_g(\mathbf{U}^*, \mathcal{Z}^*) < f_g(\mathbf{U}^*, \mathcal{Z}), \forall \mathcal{Z} \in \mathcal{B}^c, \mathcal{Z} \neq \mathcal{Z}^*. \end{array} \right\} \tag{10}$$

Lemma 2. *The set defined in (10) satisfies the following two conditions:*

1. *If $\mathbf{h} \notin \mathcal{T}$, then $f_g(\mathbf{h}^*) < f_g(\mathbf{h}), \forall \mathbf{h}^* \in J(\mathbf{h}), \mathbf{h}, \mathbf{h}^* \in \mathcal{U}_{c,n} \times \mathcal{B}^c$.*
2. *If $\mathbf{h} \in \mathcal{T}$, then $f_g(\mathbf{h}^*) \leqslant f_g(\mathbf{h}), \forall \mathbf{h}^* \in J(\mathbf{h}), \mathbf{h}, \mathbf{h}^* \in \mathcal{U}_{c,n} \times \mathcal{B}^c$.*

Proof. For any $(\mathbf{U}*, \mathcal{Z}*) \in \mathcal{U}_{c,n} \times \mathcal{B}^c$, the following chain of inequalities hold true:

$$\begin{aligned} & f_g(J(\mathbf{U}^*, \mathcal{Z}^*)) = f_g(T_{memb}(\mathcal{Z}^*), T_{cent}(T_{memb}(\mathcal{Z}^*))) \\ & \leqslant f_g(T_{memb}(\mathcal{Z}^*), \mathcal{Z}^*) \leqslant f_g(\mathbf{U}^*, \mathcal{Z}^*). \end{aligned}$$

Equality holds if and only if,

$$\mathbf{U}^* \in T_{memb}(\mathcal{Z}^*);$$
$$\text{and } \mathcal{Z}^* = T_{cent}(\mathbf{U}^*),$$

which implies that $(\mathbf{U}^*, \mathcal{Z}^*)$ is in $\mathcal{T}$. □

Lemma 3. *Under the assumption, $\Psi_i = \phi, \forall i = 1, 2, \cdots, n$; the map, T_{memb} is a point-to-point continuous map.*

Proof. Theorem 3 implies that the map under consideration is a point-to-point map. In order to address the issue of continuity, we proceed as follows.

We define the following set of functions:

$$B_{i,j_1,j_2} : \mathcal{B}^c \times \mathcal{U}_{c,n} \to \mathbb{R};$$

$$B_{i,j_1,j_2}(\mathcal{Z}, \mathbf{U}) = \tilde{g}'(u_{ij_1})d(\mathbf{z}_{j_1}, \mathbf{x}_i) - \tilde{g}'(u_{ij_2})d(\mathbf{z}_{j_2}, \mathbf{x}_i); \tag{11a}$$

$$D_i : \mathcal{B}^c \times \mathcal{U}_{c,n} \to \mathbb{R};$$

$$D_i(\mathcal{Z}, \mathbf{U}) = \sum_{j=1}^{c} u_{ij} - 1, \ i = 1, 2, \cdots, n. \tag{11b}$$

Let us define the kernel of a function $f : \mathcal{X} \to \mathbb{R}$ as follows:

$$\phi(f) = \left\{\mathbf{x} \in \mathcal{X} \mid f(\mathbf{x}) = 0\right\}.$$

Now, $\mathcal{B}_{i,j_1,j_2}(\forall i = 1, 2, \cdots n;\ j_1, j_2 \in \left\{1, 2, \cdots, c; \right\}, j_1 \neq j_2)$ is a continuous function ($\tilde{g}$ being convex and differentiable, is also continuously differentiable, hence $\tilde{g}'$ is continuous), hence, their kernel is closed. $D_i(\forall i = 1, 2, \cdots, n)$ is also continuous, hence has a closed kernel. Being intersection of closed sets, the set given by:

$$\mathcal{S} = \left\{ \cap_{i=1}^{n} \cap_{j_1=1}^{c-1} \cap_{j_2=j_1+1}^{c} \phi(B_{i,j_1,j_2}) \right\} \cap \left\{ \cap_{i=1}^{n} \phi(D_i) \right\}$$

is also closed.

Now, by definition, $\mathcal{S} \subseteq \mathcal{B}^c \times \mathcal{U}_{c,n}$. From the very definition of T_{memb}, we can rewrite $\mathcal{R}$ as follows:

$$\mathcal{R} = \left\{\mathcal{Z}, T_{memb}(\mathcal{Z}) \mid \mathcal{Z} \in \mathcal{B}^c\right\}$$

Hence $\mathcal{R}$ is also the graph of the function T_{memb}. Now, we apply the closed graph theorem to prove the continuity of T_{memb}.

Closed graph theorem: ([36], p. 171) The function from a topological space to a compact Hausdorff space is continuous, if and only if the graph of the function (the graph of $T : \mathcal{P} \to \mathcal{Y}$ is the set $\{((x, y) \in \mathcal{P} \times \mathcal{Y} \mid T(x) = y)\}$) is closed.

Using the fact that $\mathcal{U}_{c,n}$ is a compact set and $\mathcal{S}$ is a closed set, from the closed graph theorem it follows that T_{memb} is continuous. □

Lemma 4. *The map T_{memb} is closed at $\mathcal{Z}^{R_1^*}$ if $(\mathbf{U}, \mathcal{Z}^{R_1^*}) \notin \mathcal{T}$ for some $\mathbf{U} \in \mathcal{U}_{c,n}$.*

Proof. We have to prove the following: for all sequences $\{\mathcal{Z}^{R_1^{(t)}}\}_{t=0}^{\infty}(\mathcal{Z}^{R_1^{(t)}} \in \mathcal{B}^c)$ converging to $\mathcal{Z}^{R_1^*}$ and $\{\mathbf{U}^{R_1^{(t)}}\}_{t=0}^{\infty}$ $[\in T_{memb}(\mathcal{Z}^{R_1^{(t)}})]$ converging to $\mathbf{U}^{R_1^*}$; we have that, $\mathbf{U}^{R_1^*} \in T_{memb}(\mathcal{Z}^{R_1^*})$.

We shall show that the closedness property holds true individually for membership vector corresponding to each of the patterns $\mathbf{x}_i, \forall i = 1, 2, \cdots, n$. In order to show that, we define the following:

$$\Psi_i^{(t)} = \{j \mid d(\mathbf{z}_j^{R_1^{(t)}}, \mathbf{x}_i) = 0\};\ \ \Psi_i^* = \{j \mid d(\mathbf{z}_j^{R_1^*}, \mathbf{x}_i) = 0\}.$$

If $|\Psi_i^*| = 0$, using the convergence of $\{\mathbf{z}_j^{R_1^{(t)}}\}_{t=0}^{\infty}$ to $\mathbf{z}_j^{R_1^*}$ and the continuity of dissimilarity measure, we can find M_{fuzz1} such that, $\forall t > M_{fuzz1}$, $|\Psi_i^{(t)}| = 0$, implying the lemma from Lemma 3. If $|\Psi_i^*| > 0$, $\forall c > 1$, we can find $\forall c(> 0) M_{fuzz2}$ such that $\forall t > M_{fuzz2}$, $\max_{j \in \Psi_i^*} d(\mathbf{z}_j^{R_1^{(t)}}, \mathbf{x}_i) < c \min_{j \notin \Psi_i^*}\ dist_g(\mathbf{z}_j^{R_1^{(t)}}, \mathbf{x}_i)$, which implies the lemma. □

Lemma 5. *The map J is closed at $(\mathbf{U}^{R_2}, \mathcal{Z}^{R_2})$ if $(\mathbf{U}^{R_2}, \mathcal{Z}^{R_2}) \notin \mathcal{T}$.*

Proof. Lemma 4 and the continuity of T_{cent} implies this. □

Theorem 5. *$\forall\ \mathcal{Z}^{(0)} \in \mathcal{B}^c$, the sequence, $\{J(T_{memb}(\mathcal{Z}^{(0)}), \mathcal{Z}^{(0)})\}_{t=1}^{\infty}$ either terminates at a point in $\mathcal{T}$ (as defined in (10)) or has a subsequence that converges to a point in $\mathcal{T}$.*

Proof. We first restate below the Zangwill's global convergence theorem which is used to prove Theorem 5.

Zangwill's global convergence theorem [50]: Let $\mathcal{R}$ be a set of minimizers of a continuous objective function Ω on $\mathcal{H}$. Let $A : \mathcal{H} \to \mathcal{H}$ be a point-to-set map which determines an algorithm that given a point $\mathbf{s}_0 \in \mathcal{H}$, generates a sequence $\{\mathbf{s}_t\}_{t=0}^{\infty}$ through the iteration $\mathbf{s}_{t+1} \in A(\mathbf{s}_t)$. We further assume

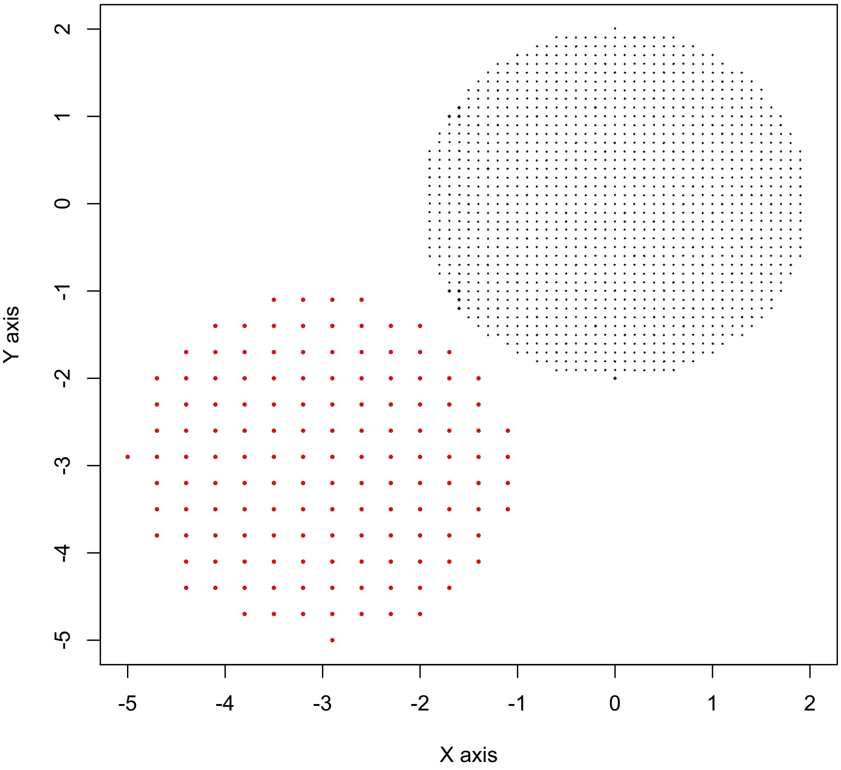


Fig. 2. Clustering performance of the FCM with CMWF of Example 3 (Section 2.3) as the weighting function.

1. The sequence $\{\mathbf{s}_t\}_{t=0}^{\infty} \in \mathcal{C} \subseteq \mathcal{H}$, where $\mathcal{C}$ is a compact set.
2. The continuous objective function Ω on $\mathcal{H}$ satisfies the following:
   (a) If $\mathbf{s} \notin \mathcal{R}$, then $\Omega(\mathbf{s}') < \Omega(\mathbf{s}), \forall \mathbf{s}' \in A(\mathcal{H})$,
   (b) If $\mathbf{s} \in \mathcal{R}$, then $\Omega(\mathbf{s}') \leqslant \Omega(\mathbf{s}), \forall \mathbf{s}' \in A(\mathcal{H})$.
3. The map A is closed at $\mathbf{s}$ if $\mathbf{s} \notin \mathcal{R}$ (if A is actually a point-to-point map instead of a point-to-set map, the condition (c) of the theorem turns out to be simply the continuity of A.)

Then the limit of any convergent subsequence of $\{\mathbf{s}_t\}_{t=0}^{\infty}$ is in $\mathcal{R}$.

We take A to be J, $\mathcal{H}$ to be $\mathcal{U}_{c,n} \times \mathcal{B}^c$; $\mathbf{s}_0$ to be $(\mathbf{U}^{(0)}, \mathcal{Z}^{(0)})$; $\mathcal{C}$ to be the whole of $\mathcal{H}$ (compactness of $\mathcal{H}$ is guaranteed by Lemma 1); Ω to be f_g (being the sum of continuous functions, f_g is a continuous function), $\mathcal{R}$ to be $\mathcal{T}$ (10) (Lemma 2 justifies the choice). By Lemma 5, J is closed on this particular choice of $\mathcal{H}$. Thus, from Zangwill's convergence theorem, the limit of any convergent subsequence of $\{\mathbf{U}^{(t)}, \mathcal{Z}^{(t)}\}_{t=0}^{\infty}$ has a limit in $\mathcal{T}$. Next $\forall(\mathcal{Z}^{(0)}) \in \mathcal{B}^c$, we consider the sequence $\{J^{(t)}(T_{memb}(\mathcal{Z}^{(0)}), \mathcal{Z}^{(0)}\}_{t=0}^{\infty}$ which is contained in the compact set given by $\mathcal{H}$. Hence by Bolzano–Weierstrass theorem [16] it has a convergent subsequence. These statements imply that the sequence given by $\{J^{(t)}(T_{memb}(\mathcal{Z}^{(0)}), \mathcal{Z}^{(0)})\}_{t=0}^{\infty}$, $\forall \mathcal{Z}^{(0)} \in \mathcal{B}^c$ either terminates at a point in $\mathcal{T}$ given by (10) or has a subsequence that converges to a point in $\mathcal{T}$. □

This provides us a complete convergence analysis of the proposed general class of algorithms with CMWF.

The developed convergence property is exactly same as that corresponding to the classical FCM with squared Euclidean distance. With the choice of the CMWF, that generates the conventional FCM weighting function (discussed in Example 1, Section 2), we obtain the convergence result corresponding to classical FCM [6,8]. Hence, this article presents a novel generalization (with respect to the weighting functions) of the classical FCm along with its convergence properties.

4. Experimental results

In this section, we present a sample performance comparison (on several simulated and real life datasets) of a specific member of the proposed generalized algorithm with the classical FCM, just to highlight the usefulness of the proposed CMWF family. Before discussing the main comparative study, we illustrate a simple proof of concept result with the dataset shown in Fig. 1(a). Fig. 2 shows the clustering obtained by FCM with the weighting function of Example 3 in Section 2.3. It is evident that the dataset is perfectly clustered with the choice of this CMWF.

Table 1
Summary of the used datasets (here, n, d and c stand for the no. of data points, no. of features, and actual clusters respectively).

Data	n	d	c
Spherical 5_2 [1]	250	2	5
Spherical 6_2 [1]	300	2	6
st900 [2]	900	2	9
elliptical_10_2 [2]	500	2	10
R15 [43]	5000	2	15
S1 [17]	5000	2	15
S3 [17]	5000	2	15
Iris Data [32]	150	4	3
Seed Data [32]	210	8	3
Wine Data [32]	178	13	3
Wisconsin Breast Cancer Data [32]	699	10	2
Wisconsin Diagnostic Breast Cancer Data [32]	569	32	2

Table 2
Comparison of ARI values. Best total and mean values are marked in boldface.

Data	FCM with CMWF	Classical FCM
Spherical 5_2	0.922	0.877 (0.012)
Spherical 6_2	0.962	0.8898 (0.01)
st900	0.839	0.839 (1)
elliptical_10_2	0.903	0.841 (0.001)
R15	0.933	0.8906 (0.15)
S1	0.962	0.925 (0.05)
S3	0.701	0.691 (0.45)
Iris Data	0.801	0.729 (0.001)
Seed Data	0.716	0.716 (1)
Wine Data	0.4136	0.3539 (0.172)
Wisconsin Breast Cancer Data	0.863	0.79 (0.05)
Wisconsin Diagnostic Breast Cancer Data	0.55	0.491 (0.001)
Total	**9.575**	9.0313
Mean	**0.798**	0.753

4.1. Benchmark dataset

Here, we consider a total of 12 datasets of which 7 are synthetic and 5 from real world. Summary of the datasets along with the relevant citations are provided in Table 1. Description of the synthetic datasets can be found in the respective references. In particular, the datasets R15, S1, and S3 can be downloaded from http://cs.joensuu.fi/sipu/datasets/. Each of S1 and S3 contain 15 Gaussian clusters with different degrees of overlap among the clusters. The dataset R15 was generated using 15 similar two-dimensional Gaussian distributions.

4.2. Performance measure

Here we employ hard partition based validity functions. In order to achieve hard partition from soft partition, we assign the point to the cluster with the highest membership. In case of a tie, it is assigned to any cluster with each candidate having equal probability.

Let $\mathcal{T} = \left\{t_1, t_2, \cdots, t_R\right\}$ and $\mathcal{S} = \left\{s_1, s_2, \cdots, s_c\right\}$ be two valid partitions of the given data. Let $\mathcal{T}$ be the actual partition and $\mathcal{S}$ be the obtained partition in some clustering algorithm, Now, we wish to evaluate the goodness of $\mathcal{S}$. n_{ij} is the number of objects present in both cluster t_i and s_j; $n_{i.}$ is the number of objects present in cluster t_i; $n_{.j}$ is the number of objects present in cluster s_j.

In order to compare the clustering performance of the different algorithms, we use a well accepted measure called Adjusted Rand Index (ARI) [25,48]. ARI can be formally expressed in the following way:

$$ARI(\mathcal{T}, \mathcal{S}) = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}$$

4.3. Computational protocols

The computational protocols that we followed throughout the simulations are in order.

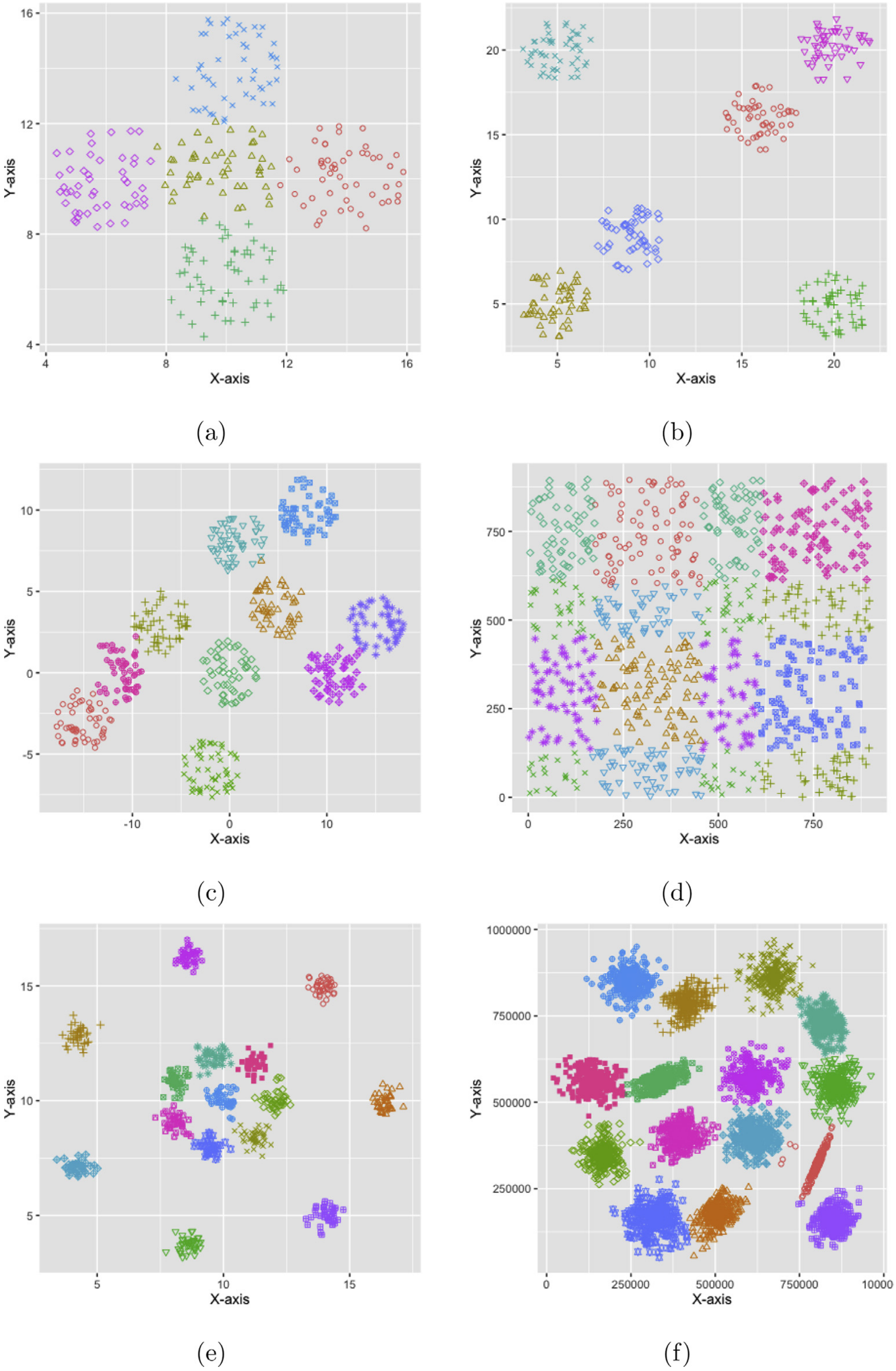
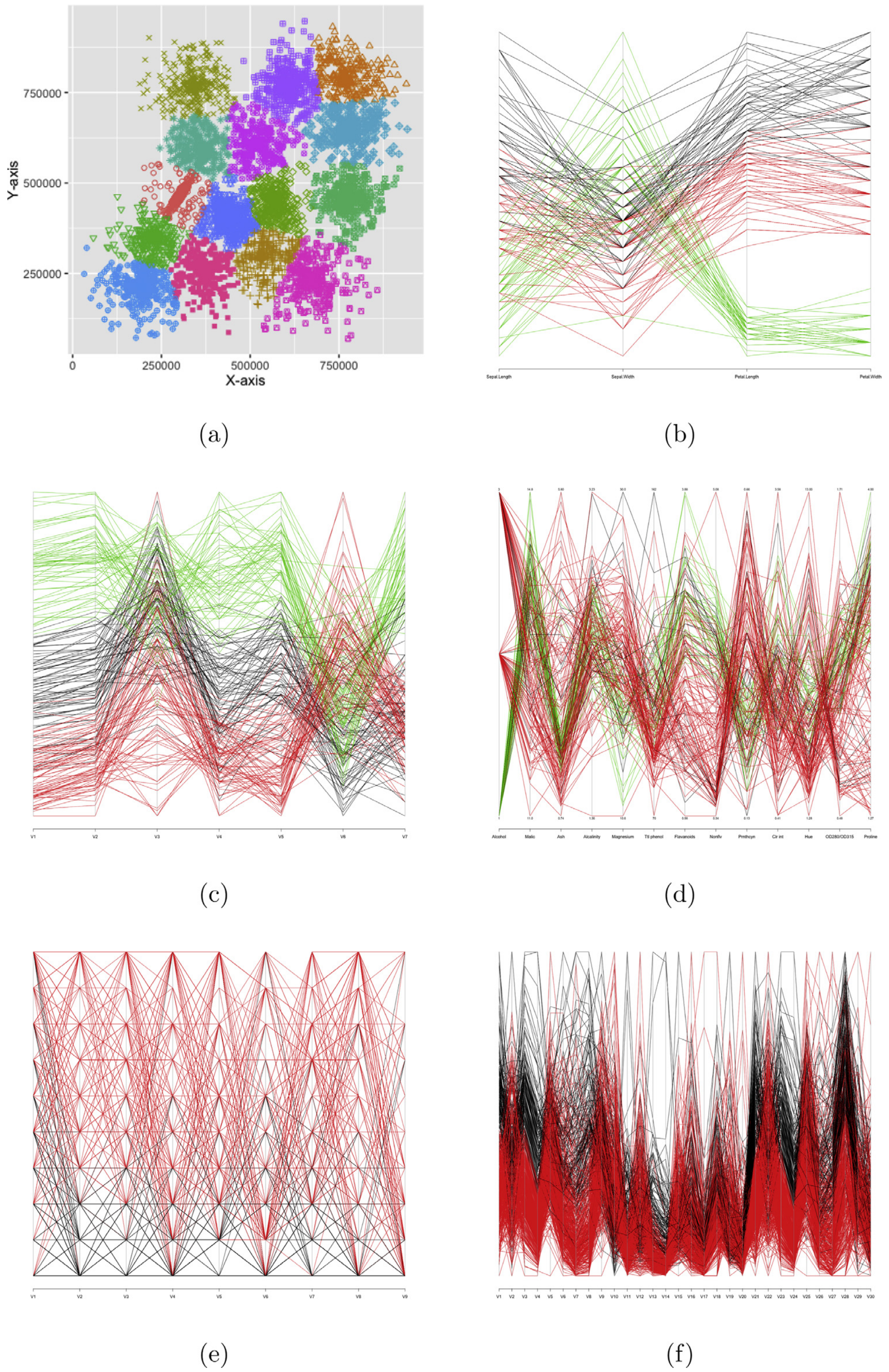


Fig. 3. The best clustering performance of our algorithm on (a) Spherical 5_2, (b) Spherical 6_2, (c) st900, (d) elliptical_10_2, (e) R15, (f) S1. (For interpretation of the references to colour in the text, the reader is referred to the web version of this article.)



(a) (b)

(c) (d)

(e) (f)

Fig. 4. The best clustering performance of our algorithm on (a) S3, (b) Iris, (c) Seed, (d) Wine, (e) Wisconsin Breast Cancer, (f) Wisconsin Diagnostic Breast Cancer Data. (For interpretation of the references to colour in the text, the reader is referred to the web version of this article.)

Choice of CMWF : From the very definition of CMWF, it is obvious that mathematically, there are infinitely many candidates for them. Due to the impossibility of demonstration of a complete performance comparison among all possible CMWFs, we choose and fix g_4 (Example 4, Section 2.3) as our choice of CMWF. We again draw readers' attention to the fact that, a performance comparison among different choices of CMWF is not the focus or the purpose of the present theoretical article. This can be taken as an interesting future research topic.

Algorithms under consideration : Proposed FCM with the specific choice of CMWF (A1), FCM with the conventional weighting function (A2). Both the algorithms used the conventional squared Euclidean distance function.

4.4. Results and statistical comparison

For each of the 12 datasets under consideration, we carry out the Wilcoxon's rank sum test (paired) between the performance by the algorithms (over 30 independent runs) under consideration, to see if we have a statistically significant improvement in our algorithm from the others. Here best performance is reported, based on mean ARI value achieved in 30 runs. For each run, algorithms A1 and A2 started with the same initial cluster representatives so that any performance difference between them may be attributed to their internal operators only.

For a specific dataset $\mathcal{R}$, let the median of the difference of the ARI corresponding to algorithms ([ARI of A1 - ARI of A2]; corresponding to each of the 30 runs) A1 and A2 be denoted by m. We perform paired Wilcoxon's rank-sum test on the following hypothesis testing setup:

$$\mathbf{H}_0 : m = 0 \text{ vs. } \mathbf{H}_1 : m > 0$$

In Table 2, we report the best performance, based upon mean value in 30 runs. In parenthesis, we report the corresponding P-values in parenthesis in the column corresponding to Classical FCM.

From Table 2, we observe that our algorithm achieved best average ARI value in all the datasets, which give us some experimental evidence on the practical usefulness of the proposed algorithm.

We graphically present the best clustering performances (Figs. 3 and 4) corresponding to our proposed generalized algorithms. In case of 2D simulated datasets, different color along with a different sign is used to identify different clusters in the scatter plot. In case of the real world datasets, with more than two features, we present the parallel coordinate plot of the clustered datasets, where each cluster is denoted by a different color.

From the *P*-values obtained in Table 2, we see that 7 out of the 12 tests have a *P*-value less than or equal to 0.05. Hence, there is enough statistically significant evidence (significance level 0.05) in the data to conclude that the ARI obtained using FCM with CMWF is bigger than that obtained with classical FCM. In the rest of the cases, the average value of the obtained ARI using FCM with CMWF is better than that corresponding to the classical FCM .

5. Conclusion

Starting with a general class of membership-based weighting functions, in this article we presented a novel class of the generalized FCM algorithms. We undertook a detailed analysis of the mathematical properties of the sub-optimization problems involved in the FCM formulation and also investigated the convergence properties (corresponding to a suitable set of optimal points) of the proposed clustering algorithm. It is evident that the CMWF-based clustering scheme can also be generalized to any FCM variant with other distance measures.

After the classical FCM algorithm [6] was introduced in literature, to the best of our knowledge, this is the first time, an article has directly addressed the issue of structured axiomatic generalization of the weighting function and has established its convergence properties which are comparable to that of the classical FCM [6]. The theoretical development of the new generalized CMWF based clustering algorithm provides us with a flexible class of weighting functions which can be used for the development of data-specific clustering algorithms with improved performances. Members from the CMWF class can be integrated with other advanced variants of FCM like the neutrosophic c-means clustering algorithm [20], interval type-2 FCM [40] etc.

What is the best way to choose an appropriate weighting function given the data? How can we approximate the updating rule corresponding to the weighting function, when a closed form expression is not readily available? Even partial answers to such theoretical questions would have a significant practical impact and deserve further investigations. We wish to continue our research work in this direction.

References

[1] S. Bandyopadhyay, U. Maulik, Genetic clustering for automatic evolution of clusters and application to image classification, Pattern Recognit. 35 (6) (2002) 1197–1208.

[2] S. Bandyopadhyay, S.K. Pal, Classification and Learning using Genetic Algorithms: Applications in Bioinformatics and Web Intelligence, Springer Science & Business Media, 2007.

[3] A. Banerjee, X. Guo, H. Wang, On the optimality of conditional expectation as a Bregman predictor, Inf. Theory IEEE Trans. 51 (7) (2005) 2664–2669.

[4] A. Banerjee, S. Merugu, I.S. Dhillon, J. Ghosh, Clustering with Bregman divergences, J. Mach. Learn. Res. 6 (2005) 1705–1749.

[5] A. Ben-Tal, A. Charnes, M. Teboulle, Entropic means, J. Math. Anal. Appl. 139 (2) (1989) 537–551.

[6] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Kluwer Academic Publishers, Norwell, MA, USA, 1981.

[7] J.C. Bezdek, R.J. Hathaway, Convergence of alternating optimization, Neural Parallel Sci. Comput. 11 (4) (2003) 351–368.

[8] J.C. Bezdek, R.J. Hathaway, M.J. Sabin, W.T. Tucker, Convergence theory for fuzzy c-means: counterexamples and repairs, Syst. Man Cybern. IEEE Trans. 17 (5) (1987) 873–877.

[9] L. Bobrowski, J.C. Bezdek, c-means clustering with the l_1 and l_∞ norms, Syst. Man Cybern. IEEE Trans. 21 (3) (1991) 545–554.

[10] L.M. Bregman, The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, USSR Comput. Math. Math. Phys. 7 (3) (1967) 200–217.

[11] Chaomurilige, J. Yu, M.-S. Yang, Analysis of parameter selection for gustafson–kessel fuzzy clustering using jacobian matrix, IEEE Trans. Fuzzy Syst. 23 (6) (2015) 2329–2342.

[12] I. Csisz, et al., On topological properties of f-divergences, Studia Sci. Math. Hungar. 2 (1967) 329–339.

[13] I. Csiszár, I-divergence geometry of probability distributions and minimization problems, Ann. Probab. (1975) 146–158.

[14] J.C. Dunn, A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters, Cybern. Syst. (1973) 32–57.

[15] P. Fazendeiro, J.V. de Oliveira, Observer-biased fuzzy clustering, IEEE Trans. Fuzzy Syst. 23 (1) (2015) 85–97.

[16] P. Fitzpatrick, Advanced Calculus, vol. 5, American Mathematical Soc., 2006.

[17] P. Fränti, O. Virmajoki, Iterative shrinking method for clustering problems, Pattern Recognit. 39 (5) (2006) 761–775.

[18] S. Ganguly, D. Bose, A. Konar, Clustering using vector membership: an extension of the fuzzy c-means algorithm, in: 2013 Fifth International Conference on Advanced Computing (ICoAC), 2013, pp. 27–32, doi:10.1109/ICoAC.2013.6921922.

[19] L. Groll, J. Jakel, A new convergence proof of fuzzy c-means, IEEE Trans. Fuzzy Syst. 13 (5) (2005) 717–720, doi:10.1109/TFUZZ.2005.856560.

[20] Y. Guo, A. Sengur, Ncm: neutrosophic c-means clustering algorithm, Pattern Recognit. 48 (8) (2015) 2710–2724. http://dx.doi.org/10.1016/j.patcog.2015.02.018.

[21] L.O. Hall, D.B. Goldgof, Convergence of the single-pass and online fuzzy c-means algorithms, IEEE Trans. Fuzzy Syst. 19 (4) (2011) 792–794, doi:10.1109/TFUZZ.2011.2143418.

[22] R.J. Hathaway, J.C. Bezdek, Y. Hu, Generalized fuzzy c-means clustering strategies using l_p norm distances, Fuzzy Syst. IEEE Trans. 8 (5) (2000) 576–582.

[23] F. Hoppner, F. Klawonn, A contribution to convergence theory of fuzzy c-means and derivatives, IEEE Trans. Fuzzy Syst. 11 (5) (2003) 682–694, doi:10.1109/TFUZZ.2003.817858.

[24] M. Huang, Z. Xia, H. Wang, Q. Zeng, Q. Wang, The range of the value for the fuzzifier of the fuzzy c-means algorithm, Pattern Recognit. Lett. 33 (16) (2012) 2280–2284.

[25] L. Hubert, P. Arabie, Comparing partitions, J. classification 2 (1) (1985) 193–218.

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

[27] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264–323.

[28] F. Klawonn, F. Höppner, An alternative approach to the fuzzifier in fuzzy clustering to obtain better clustering results, in: Proceedings of the 3rd Eusflat Conference, 2003, pp. 730–734.

[29] F. Klawonn, F. Höppner, What is fuzzy about fuzzy clustering? Understanding and improving the concept of the fuzzifier, in: Advances in Intelligent Data Analysis V, Springer, 2003, pp. 254–264.

[30] F. Klawonn, A. Keller, Fuzzy clustering based on modified distance measures, in: Advances in Intelligent Data Analysis, Springer, 1999, pp. 291–301.

[31] Y. Lei, J. Bezdek, J. Chan, N. Vinh, S. Romano, J. Bailey, Extending information-theoretic validity indices for fuzzy clustering, IEEE Trans. Fuzzy Syst. PP (99) (2016), doi:10.1109/TFUZZ.2016.2584644. 1–1

[32] M. Lichman, UCI machine learning repository, 2013.

[33] J. Lin, Divergence measures based on the shannon entropy, Inf. Theory IEEE Trans. 37 (1) (1991) 145–151.

[34] S. Miyamoto, Y. Agusta, An efficient algorithm for l_1 fuzzy c-means and its termination, Control Cybern. 24 (1995) 421–436.

[35] S. Miyamoto, Y. Agusta, Algorithms for L_1 and L_p fuzzy c-means and their convergence, in: Data Science, Classification, and Related Methods, Springer, 1998, pp. 295–302.

[36] J.R. Munkres, Topology, second ed., Prentice Hall, Prentice Hall, Upper Saddle River, NJ 07458, 2000.

[37] F. Nielsen, S. Boltz, The burbea-rao and bhattacharyya centroids, Inf. Theory IEEE Trans. 57 (8) (2011) 5455–5466.

[38] F. Nielsen, R. Nock, Sided and symmetrized Bregman centroids, Inf. Theory IEEE Trans. 55 (6) (2009) 2882–2904.

[39] R. Nock, F. Nielsen, S.I. Amari, On conformal divergences and their population minimizers, IEEE Trans. Inf. Theory 62 (1) (2016) 527–538, doi:10.1109/TIT.2015.2448072.

[40] C. Qiu, J. Xiao, L. Han, M.N. Iqbal, Enhanced interval type-2 fuzzy c-means algorithm with improved initial center, Pattern Recognit. Lett. 38 (2014) 86–92.

[41] A. Saha, S. Das, Geometric divergence based fuzzy clustering with strong resilience to noise features, Pattern Recognit. Lett. 79 (2016) 60–67.

[42] M. Teboulle, A unified continuous optimization framework for center-based clustering methods, J. Mach. Learn. Res. 8 (2007) 65–102.

[43] C.J. Veenman, M.J. Reinders, E. Backer, A maximum variance cluster algorithm, Pattern Anal. Mach. Intell. IEEE Trans. 24 (9) (2002) 1273–1280.

[44] C.H. Wu, C.S. Ouyang, L.W. Chen, L.W. Lu, A new fuzzy clustering validity index with a median factor for centroid-based clustering, IEEE Trans. Fuzzy Syst. 23 (3) (2015) 701–718, doi:10.1109/TFUZZ.2014.2322495.

[45] J. Wu, H. Xiong, C. Liu, J. Chen, A generalization of distance functions for fuzzy-means clustering with centroids of arithmetic means, Fuzzy Syst. IEEE Trans. 20 (3) (2012) 557–571.

[46] S.D. Xenaki, K.D. Koutroumbas, A.A. Rontogiannis, A novel adaptive possibilistic clustering algorithm, IEEE Trans. Fuzzy Syst. 24 (4) (2016) 791–810, doi:10.1109/TFUZZ.2015.2486806.

[47] R. Xu, D. Wunsch, et al., Survey of clustering algorithms, Neural Netw. IEEE Trans. 16 (3) (2005) 645–678.

[48] K.Y. Yeung, W.L. Ruzzo, Details of the adjusted rand index and clustering algorithms, supplement to the paper “an empirical study on principal component analysis for clustering gene expression data”, Bioinformatics 17 (9) (2001) 763–774.

[49] J. Yu, Q. Cheng, H. Huang, Analysis of the weighting exponent in the fcm, IEEE Trans. Syst. Man Cybern. Part B 34 (1) (2004) 634–639, doi:10.1109/TSMCB.2003.810951.

[50] W.I. Zangwill, Nonlinear Programming: A Unified Approach, vol. 196, Prentice-Hall Englewood Cliffs, NJ, 1969.