



Sub-One Quasi-Norm-Based k -Means Clustering Algorithm and Analyses

Qi An¹ · Shan Jiang²

Accepted: 6 April 2024
© The Author(s) 2024

Abstract

Recognizing the pivotal role of choosing an appropriate distance metric in designing the clustering algorithm, our focus is on innovating the k -means method by redefining the distance metric in its distortion. In this study, we introduce a novel k -means clustering algorithm utilizing a distance metric derived from the ℓ_p quasi-norm with $p \in (0, 1)$. Through an illustrative example, we showcase the advantageous properties of the proposed distance metric compared to commonly used alternatives for revealing natural groupings in data. Subsequently, we present a novel k -means type heuristic by integrating this sub-one quasi-norm-based distance, offer a step-by-step iterative relocation scheme, and prove the convergence to the Kuhn-Tucker point. Finally, we empirically validate the effectiveness of our clustering method through experiments on synthetic and real-life datasets, both in their original form and with additional noise introduced. We also investigate the performance of the proposed method as a subroutine in a deep learning clustering algorithm. Our results demonstrate the efficacy of the proposed k -means algorithm in capturing distinctive patterns exhibited by certain data types.

Keywords Sub-one quasi-norm · k -means clustering · Lloyd's heuristics

1 Introduction

Partitional clustering involves the segmentation of a set of samples into clusters based on defined criteria such that samples in the same cluster exhibit greater similarity than those in different clusters. The k -means algorithm, a prominent unsupervised partitional clustering scheme [26], accomplishes this by categorizing data points into k disjoint clusters while minimizing the total distortion function. Distortion, in this context, is the sum of distances between data points and their assigned cluster center. This algorithm has gained widespread acclaim due to its simplicity, efficiency, ease of implementation, and empirical success in various fields such as data mining, pattern recognition, marketing research, and social network

✉ Shan Jiang
shanjiang@xmu.edu.cn

Qi An
anq@ouchn.edu.cn

¹ Department of Computer Science, The Open University of China, 100039 Beijing, China

² School of Management, Xiamen University, 361005 Xiamen, China

analysis [12, 20, 37]. Jain and Kar provided a comprehensive overview of clustering methods, addressing major challenges and key issues in the algorithm design [22]. The most commonly used heuristic for k -means clustering is the Lloyd's method. It commences with a set of randomly chosen initial cluster centers, and in each iteration, assigns each data point to the cluster with the closest center, determined by a distance metric. The cluster representatives are then updated to the center of the newly formed clusters. This iterative process is repeated until a predefined stopping rule is met.

One of the major decisions in the design of the k -means method is the choice of an appropriate distance metric. The sum of squared Euclidean distances is the conventional measure of dissimilarity between data and has been widely adopted in studies involving k -means and deep learning based clustering [35]. In this paper, we propose a k -means clustering algorithm with the usage of a distance metric derived from the sub-one quasi-norm (ℓ_p quasi-norm with $p \in (0, 1)$). In contrast to the Euclidean distance, this metric leverages similar data-items more effectively while assigning less weight to dissimilar ones. To illustrate the desired properties inherent in the proposed distance metric, we use the Iris dataset to exemplify its ability to yield more unbiased similarity values in classification tasks. Computational experiments on both synthetic and real-life datasets corroborate the enhanced performance of our proposed k -means method in achieving effective classification. Additional experiments were conducted on datasets with introduced random noise to underscore the resilience of the proposed k -means method to noise. Moreover, we also integrated our novel k -means method into a deep learning clustering model to further investigate its efficacy in handling more intricate tasks and real-world applications. These findings validate the context-dependent effectiveness of our proposed sub-one quasi-norm based k -means algorithm over standard k -means algorithms using other common distance metrics.

The rest of this paper is organized as follows. Section 2 reviews the search for an appropriate distance metric in the k -means clustering literature. Section 3 introduces the sub-one quasi-norm-based distance and analyzes its desired properties. In Sect. 4, we incorporate the sub-one quasi-norm-based distance metric into a k -means type heuristic and provide a step-by-step iterative relocation scheme. Section 5 presents numerical results to compare the performance of k -means algorithms with the proposed distance against other commonly used distance metrics. Section 6 concludes the paper.

2 Literature Review

The widely recognized k -means algorithm seeks to minimize the sum of squared Euclidean distances from each data point to its nearest cluster center [14, 21, 26]. Nonetheless, the Euclidean distance metric may not consistently produce high-quality cluster centers especially when the dataset contains noise or outliers. Numerous studies have proposed alternative mathematical metrics to improve distance calculations. In order to increase resistance to noise and outliers, Kersten suggested employing Manhattan distance instead of Euclidean distance in the k -means clustering algorithm [25]. Singh et al. explored the implementation of the k -means algorithm using Minkowski distance metric [33]. Recently, there has been a trend towards exploring various nonlinear distance metrics as an alternative to Euclidean distance [31]. Bobrowski and Bezdek extended the k -means clustering algorithm by exploring two similarity measures with nonhyperelliptical topologies: the square of the l_1 norm and the square of the supremum norm l_∞ [5]. Hathaway et al. examined clustering with general l_p norm distances, with p varying from 0 to 1, and evaluated clustering performance with arti-

ficial datasets [18]. Aside from members of the Minkowski or p -norm family, Filippone et al. surveyed modified k -means clustering algorithms incorporating kernel-based similarity measures that produce nonlinear separating hypersurfaces between clusters [15]. de Amorim and Mirkin extended the existing k -means method with an automated feature weighting scheme using the corresponding Minkowski metric to improve resilience to irrelevant or noisy features [11].

Researchers have also suggested replacing Euclidean distance with various divergence measures that violate some metric properties for the k -means clustering problem. Dhillon et al. proposed a divisive algorithm based on Kullback–Leibler divergences, which are not symmetric and do not obey the triangle inequality [13]. Banerjee et al. presented an alternative formulation of the k -means clustering scheme using the general Bregman divergence that unifies all divergence measures [4]. Nielsen et al. investigated a parametric family of divergence measures for k -means clustering [29]. Chakraborty and Das derived the S -distance from the S -divergence defined on the open cone of positive definite matrices and proved the convergence of the induced k -means algorithm to a local optimum [7]. In fuzzy partitioning clustering, where a data sample can belong to multiple clusters with varying degrees of fuzzy membership, Saha and Das explored using geometric divergence to enhance the immunity of fuzzy k -means type algorithms towards noisy features [30]. Seal et al. proposed a Jeffrey-divergence based distance metric and developed a fuzzy c -means (FCM) clustering algorithm that guarantees convergence to a local minimum [31].

Recognizing the sensitivity of the commonly used Euclidean distance to the distortion of heavy outliers and its biased impact on the overall similarity values [10], alternative measures of dissimilarity have been studied in the machine learning literature. The non-convex ℓ_p quasi-norm with $p \in (0, 1)$ has gained research interest due to its capacity for promoting sparsity in model parameters. This sub-one quasi-norm has found various applications in matrix completion [28], proximal support vector machine [9], vector reconstruction [24] in the fields of machine learning and compressed sensing. It has been shown to be more robust than the classical sparse-inducing term ℓ_1 norm, which induces the Manhattan distance [34]. The advantages of the sub-one quasi-norm in handling approximately sparse data, noisy measurements, and heavy outliers have been demonstrated. In light of these advantages, Jiang et al. [23] first proposed a sub-one quasi-norm-based similarity measure for collaborative filtering in a recommender system. This similarity measure has proven superior to commonly used distance metrics, as it makes better use of rating values and avoids overweighing dissimilar ratings between users, whether in fully co-rated, sparsely co-rated or cold-start scenarios. Our work represents another step towards exploiting the desired properties of ℓ_p quasi-norm in learning models by extending the sub-one quasi-norm-based similarity measure to clustering analysis.

3 Sub-One Quasi-Norm-Based Distance Metric

This section describes the distance metric induced by the ℓ_p quasi-norm with $p \in (0, 1)$ and its desired properties.

For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, a well-defined function $d : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}_+$ that measures the distance between \mathbf{x} and \mathbf{y} has to satisfy the following three requirements:

- Identity of indiscernibles: $d(\mathbf{x}, \mathbf{y}) = 0 \iff \mathbf{x} = \mathbf{y}$;
- Symmetry: $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$;
- Triangle inequality: $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$.

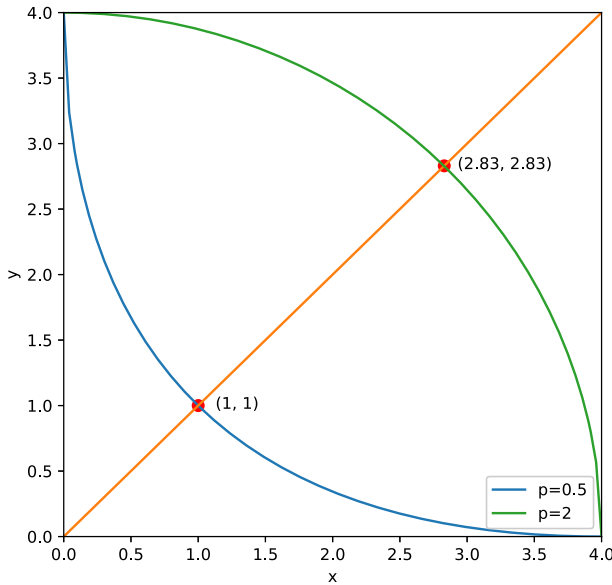


Fig. 1 Equal distance contour plot of $l_{0.5}$ -distance and Euclidean distance

Minkowski distance $(\sum_{i=1}^m |x_i - y_i|^p)^{1/p}$ constitutes a family of classical distance metrics derived from the ℓ_p norm $\|\mathbf{x}\|_p = (\sum_{i=1}^m |x_i|^p)^{1/p}$ where $p \geq 1$. Familiar examples within this family include the Euclidean distance (induced by $p = 2$) and the Manhattan distance (induced by $p = 1$). In this research, we focus on $p \in (0, 1)$ for which $\|\cdot\|_p$ is no longer subadditive and forms a quasi-norm. For $p \in (0, 1)$, it holds that

$$\|\mathbf{x}\|_p^p + \|\mathbf{y}\|_p^p \leq \|\mathbf{x} + \mathbf{y}\|_p^p$$

for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$. Hence, the function $\|\cdot\|_p^p$ induced by the ℓ_p quasi-norm becomes subadditive and defines a distance function expressed as

$$d_p(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^m |x_i - y_i|^p.$$

The Minkowski distance and the ℓ_p quasi-norm-based distance metric exhibit distinct behaviors in similarity evaluation. As the value of p diminishes below 1, there is a heightened sensitivity of the ℓ_p quasi-norm-based distance to no change or minor variations in a single dimension when assessing similarity. To illustrate this characteristic, we depict the contour of the $l_{0.5}$ quasi-norm-based distance function alongside the Euclidean distance function in a two-dimensional space in Fig. 1. Notably, the point $(0, 4)$ has the same Euclidean distance from the origin $(0, 0)$ as the point $(2.83, 2.83)$. However, it is as distant from the origin as the point $(1, 1)$ in terms of the $l_{0.5}$ quasi-norm-based distance. Therefore, the point $(0, 4)$ is deemed more similar to the origin when $l_{0.5}$ quasi-norm-based distance is adopted as a similarity measure, compared to the Euclidean distance. In other words, the ℓ_p quasi-norm-based distance metric with $p \in (0, 1)$ emphasizes similarity along the x -axis between the points $(0, 0)$ and $(0, 4)$, while Euclidean distance emphasizes variations along the y -axis.

In the context of data classification tasks, leveraging ℓ_p quasi-norm-based distance as a similarity measure offers two advantages. First, assigning greater importance to similar data-

Table 1 Example of six iris plants

Attribute in cm	A	B	C	D	E	F
Sepal length	6.3	5.0	5.8	6.7	5.8	5.9
Sepal width	2.3	2.3	2.7	2.5	2.7	3.0
Petal length	4.4	3.3	5.1	5.8	5.1	5.1
Petal width	1.3	1.0	1.9	1.8	1.9	1.8

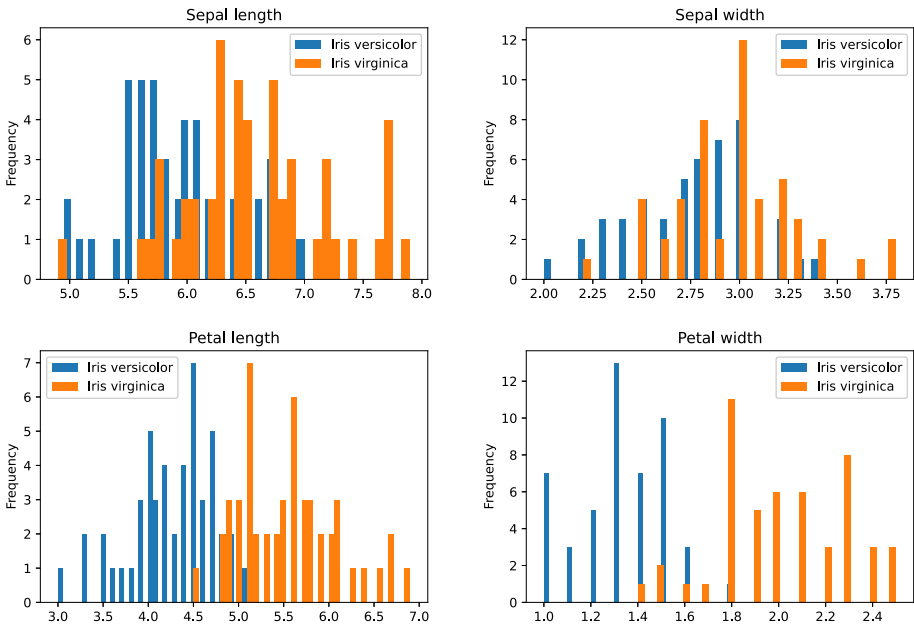


Fig. 2 Class-stratified histogram of attributes in Iris dataset

items helps mitigate the impact of substantial variations in few dimensions, which can aid in the identification of data samples belonging to the same category. Second, it enhances the robustness of clustering algorithms against noise and outliers by avoiding excessive emphasis on dissimilar data-items.

We demonstrate the effectiveness of the proposed ℓ_p quasi-norm-based distance metric using an illustrative example involving six iris plants from the well-known Iris dataset [16], which is widely recognized in the pattern recognition literature. Table 1 presents the sepal and petal length and width measurements (in centimeters) for plants A, B, C, D, E, and F. In this context, plants A and B belong to the Versicolor type, while plants C, D, E and F belong to the Virginica type. From Table 1, we can make the following observations: (i) Regarding the sepal length variable, the measurement of plant A falls in the range of plants C, D, E, and F, while the measurement of plant B stands out. (ii) Regarding the sepal width, plant A has the same measurement as plant B, but differs from plants C, D, E, and F. (iii) Regarding the petal length and width, the measurement of plant A is similar to that of plant B, but deviates from plants C, D, E, and F. To gain a deeper understanding of attribute value distribution for the Versicolor and Virginica iris species, we present histograms in Fig. 2 examining each attribute individually across all instances in the Iris dataset. The figure reveals that petal length has the widest distribution, whereas petal width exhibits the lowest mean among the four attributes.

Table 2 Distance to plant A under different distance metrics

Distance metric	B	C	D	E	F
$\ell_{0.5}$ quasi-norm-based distance	2.7367	2.9508	2.9700	2.9508	3.0129
Manhattan distance (ℓ_1 norm-based distance)	2.7000	2.2000	2.5000	2.2000	2.3000
Euclidean distance (ℓ_2 norm-based distance)	1.7292	1.2225	1.5524	1.2225	1.1790
Chebyshev distance	1.3	0.7	1.4	0.7	0.7

Although there is a notable overlap in the distributions of Versicolor and Virginica for sepal length and width, the measurements for petal length and width form fairly distinct groups. Thereby, distinguishing between the two species requires recognizing the similar petal length and width measurements within the same iris type.

An ideal distance metric should produce a smaller distance between plants A and B compared to that between plant A and other plants. To achieve this, this distance metric should not be excessively influenced by significant variations in sepal length but should recognize similar petal length and width measurements between iris plants of the sane type. Table 2 presents distances between plant A and other plants under ℓ_p quasi-norm-based and other commonly used metrics (Manhattan, Euclidean and Chebyshev) for comparison. Larger distances indicate greater dissimilarity from plant A. It is evident that, except for the $\ell_{0.5}$ quasi-norm-based distance, other metrics were not able to identify plant B as the most similar to plant A, consequently failing to classify them under the same iris type. These metrics tended to overemphasize dissimilarities in sepal length, a variable with a broad distribution, and failed to capture the shared sepal width as well as the close petal length and width measurements between plants A and B. Due to the concave nature of the contour plots of the l_p quasi-norm function, l_p quasi-norm-based distance makes better use of identical data-items while avoiding giving excessive weights to dissimilar data-items. This property is particularly advantageous when handling data with slight variations in a single variable, especially one possessing significant discriminatory power for classification, as demonstrated in our experimental results in Sect. 5.

4 Sub-One Quasi-Norm-Based k -Means Algorithm

Consider a set of vectors $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$ with $\mathbf{x}_i \in \mathbb{R}^m, i = 1, \dots, n$. We want to partition \mathcal{X} into k many exhaustive and mutually exclusive classes ($2 \leq k \leq n$). Adopting the ℓ_p quasi-norm-based distance metric, we formalize the clustering problem in the following mathematical form:

$$\begin{aligned}
 \text{(P): } \quad & \min f(\mathbf{W}, \mathbf{Z}) = \sum_{i=1}^n \sum_{j=1}^k w_{ij} d_p(\mathbf{x}_i, \mathbf{z}_j) \\
 & \text{s.t. } \sum_{j=1}^k w_{ij} = 1, \quad i = 1, \dots, n, \\
 & \quad w_{ij} \in \{0, 1\}, \quad i = 1, \dots, n, \quad j = 1, \dots, k,
 \end{aligned}$$

where $\mathbf{W} = [w_{ij}] \in \mathbb{Z}_2^{n \times k}, \mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_k]^T \in \mathbb{R}^{k \times m}$.

Regardless of the choice of the distance metric, Problem (P) is an NP-hard non-convex programming problem [2]. By relaxing the binary requirements of w_{ij} , this problem can be reformulated as a reduced problem defined as:

$$\begin{aligned}
 \text{(RP): } \quad & \min F(\mathbf{W}) := \inf_{\mathbf{Z}} f(\mathbf{W}, \mathbf{Z}) \\
 \text{s.t. } \quad & \mathbf{W} \in \mathcal{C} := \{\mathbf{W} \in \mathbb{R}^{n \times k} \mid \sum_{j=1}^k w_{ij} = 1, w_{ij} \geq 0, i = 1, \dots, n, j = 1, \dots, k\}.
 \end{aligned}$$

Problems (P) and (RP) are equivalent because, due to the concavity of $F(\cdot)$, the optimal solutions for Problem (RP) precisely correspond to the extreme points of \mathcal{C} . Meanwhile, the extreme points of \mathcal{C} inherently satisfy the constraints of Problem (P). Problem (RP) is typically addressed through a greedy alternating minimization algorithm. Here, we introduce a sub-one quasi-norm-based k -means clustering algorithm for solving (RP) based on the well-established Lloyd’s k -means method. The following iterative heuristics provide a detailed account of its implementation.

Algorithm 1 .

Input: Parameters n, m, k , dataset $\{\mathbf{x}_i\}_{i=1}^n$ with $\mathbf{x}_i \in \mathbb{R}^m, i = 1, \dots, n$.

Initialize: A set of randomly selected k vectors in \mathbb{R}^m such that $\mathbf{Z}^{(0)} = [z_1^0, \dots, z_k^0]^\top$. $\mathbf{W}^{(0)} = \operatorname{argmin}_{\mathbf{W} \in \mathcal{C}} f(\mathbf{W}, \mathbf{Z}^{(0)})$. Set $t = 0$;

- 1: Compute $\mathbf{Z}^{(t+1)} = \operatorname{argmin}_{\mathbf{Z}} f(\mathbf{W}^{(t)}, \mathbf{Z})$, i.e.. update the cluster centers based on the most recent class assignment. If $f(\mathbf{W}^{(t)}, \mathbf{Z}^{(t+1)}) = f(\mathbf{W}^{(t)}, \mathbf{Z}^{(t)})$, terminate the algorithm; otherwise, continue.
- 2: Compute $\mathbf{W}^{(t+1)} = \operatorname{argmin}_{\mathbf{W} \in \mathcal{C}} f(\mathbf{W}, \mathbf{Z}^{(t+1)})$, i.e. fix the cluster centers and assign each vector to its nearest center. If $f(\mathbf{W}^{(t+1)}, \mathbf{Z}^{(t+1)}) = f(\mathbf{W}^{(t)}, \mathbf{Z}^{(t+1)})$, terminate the algorithm; otherwise, continue.
- 3: Update $t \rightarrow t + 1$, return to Step 1.

Output: Current cluster assignment and cluster centers.

In Algorithm 1, Problem (P) is solved by alternating separate optimizations over the following two problems:

- (P1): For fixed $\mathbf{Z} \in \mathbb{R}^{k \times m}$, minimize $f(\mathbf{W}, \mathbf{Z})$ subject to $\mathbf{W} \in \mathcal{C}$;
- (P2): For fixed $\mathbf{W} \in \mathbb{Z}_2^{n \times k}$, minimize $f(\mathbf{W}, \mathbf{Z})$ subject to $\mathbf{Z} \in \mathbb{R}^{k \times m}$.

The solution to Problem (P1) is straightforward. Given any $\mathbf{Z} \in \mathbb{R}^{k \times m}$, for each $i = 1, \dots, n$, we set $w_{ir} = 1$ for the smallest r satisfying $d_p(\mathbf{x}_i, \mathbf{z}_r) \leq d_p(\mathbf{x}_i, \mathbf{z}_j)$ for all $j = 1, \dots, k$. Subsequently, we set $w_{ij} = 0$ for all $j \neq r$. However, the solution to Problem (P2) is not as straightforward, especially when the ℓ_p quasi-norm-based distance function d_p with $p \in (0, 1)$ is adopted. Problem (P2) is an ℓ_p quasi-norm minimization problem that has been proved to be strongly NP-hard [17]. Due to the non-convexity of ℓ_p quasi-norm, a local optimum of Problem (P2) may not be a global optimal solution.

For a fixed $\mathbf{W} \in \mathbb{Z}_2^{n \times k}$, as the distance function d_p is separable by variable, Problem (P2) can be decomposed into independent univariate minimization problems:

$$\text{(P2-}jl\text{)} : \quad \min f_{jl}(z_{jl}) := \sum_{i=1}^n w_{ij} |z_{jl} - x_{il}|^p$$

for all $j = 1, \dots, k$ and $l = 1, \dots, m$. \mathbf{Z}^* is an optimal solution to Problem (P2) if and only if z_{jl}^* solves Problem (P2- jl) for all $j = 1, \dots, k$ and $l = 1, \dots, m$. We note that the local minimizer of Problem (P2- jl) is easy to locate.

Lemma 1 For any $j = 1, \dots, k$ and $l = 1, \dots, m$, Problem (P2- jl) attains its local minimum values at x_{il} , where $i \in \{1, \dots, n\}$ satisfies $w_{ij} \neq 0$.

Proof Consider any j and l . The function $f_{jl}(\cdot)$ is continuous and differentiable everywhere except at $z_{il} = x_{il}$ for any $i = 1, \dots, n$ with $w_{ij} \neq 0$. Define $n_j := \sum_{i=1}^n w_{ij}$. Let the sequence i_1, \dots, i_{n_j} be an ordered list of the set $\{i \mid w_{ij} \neq 0, i = 1, \dots, n\}$ with $x_{i_1 l} \leq \dots \leq x_{i_{n_j} l}$. On any open interval among $(-\infty, x_{i_1 l})$, $(x_{i_1 l}, x_{i_2 l})$, \dots , $(x_{i_{n_j} l}, +\infty)$, the function $f_{jl}(\cdot)$ is differentiable and can be expressed as the sum of multiple strictly concave functions. As a result, $f_{jl}(\cdot)$ is strictly concave on these open intervals, with its graph showing a downward jump at two endpoints of the interval. By summing up the derivatives, we obtain one-sided differentials: $f_{jl}(x_{il}^+) = +\infty$ and $f_{jl}(x_{il}^-) = -\infty$, for any $i \in \{i_1, \dots, i_{n_j}\}$. Hence, for any $i \in \{1, \dots, n\}$ with $w_{ij} \neq 0$, $z_{jl} = x_{il}$ serves as a local minimizer of Problem (P2- jl). Note that $f_{jl}(\cdot)$ is not differentiable at its local minimum points. □

To obtain the optimal solution \mathbf{Z}^* of Problem (P2), we address each individual subproblem (P2- jl) for all $j = 1, \dots, k$, $l = 1, \dots, m$, and identify the component value z_{jl} that yield the overall minimum to Problem (P2- jl). Within each (P2- jl), we iterate over the set $\{x_{i_1 l}, \dots, x_{i_{n_j} l}\}$ and select the smallest value that globally minimizes the objective function f_{jl} .

As proved in Selim and Ismail [32], Algorithm 1 converges in a finite number of iterations to a partial optimal solution $(\mathbf{W}^*, \mathbf{Z}^*)$ defined as

$$\begin{aligned} f(\mathbf{W}^*, \mathbf{Z}^*) &\leq f(\mathbf{W}, \mathbf{Z}^*) \quad \text{for all } \mathbf{W} \in \mathcal{C}, \\ f(\mathbf{W}^*, \mathbf{Z}^*) &\leq f(\mathbf{W}^*, \mathbf{Z}) \quad \text{for all } \mathbf{Z} \in \mathbb{R}^{k \times m}. \end{aligned}$$

It is important to note that $(\mathbf{W}^*, \mathbf{Z}^*)$ is a partial optimal solution if and only if \mathbf{W}^* solves Problem (P1) given $\mathbf{Z} = \mathbf{Z}^*$ and \mathbf{Z}^* solves Problem (P2) given $\mathbf{W} = \mathbf{W}^*$. Selim and Ismail demonstrated that partial optimal solutions are Kuhn-Tucker points for Problem (P) provided that $f(\mathbf{W}, \mathbf{Z})$ is differentiable at $(\mathbf{W}^*, \mathbf{Z}^*)$. However, when utilizing sub-one ℓ_p quasi-norm based distance d_p in Problem (P), as discussed in Lemma 1, $f(\mathbf{W}, \mathbf{Z})$ may not be differentiable at $\mathbf{Z} = \mathbf{Z}^*$. In this setup, we establish that partial optimal solutions still function as Kuhn-Tucker points for Problem (P).

Theorem 2 A point $(\mathbf{W}^*, \mathbf{Z}^*)$ is a Kuhn-Tucker point of Problem (P) if and only if it is a partial optimal solution.

Proof For any $\mathbf{Z} = [z_1, \dots, z_k]^T$ and $j = 1, \dots, k$, let $S_l(z_j)$ be an active set of z_j on the l th dimension, defined as

$$S_l(z_j) = \{i \mid z_{jl} \neq x_{il}\}.$$

The Kuhn-Tucker conditions of Problem (P) are given by

- (i) $d_p(\mathbf{x}_i, \mathbf{z}_j) + \lambda_i \geq 0$, $i = 1, \dots, n$, $j = 1, \dots, k$.
- (ii) $w_{ij}(d_p(\mathbf{x}_i, \mathbf{z}_j) + \lambda_i) = 0$, $i = 1, \dots, n$, $j = 1, \dots, k$.
- (iii) $\sum_{j=1}^k w_{ij} = 1$, $i = 1, \dots, n$.
- (iv) $w_{ij} \geq 0$, $i = 1, \dots, n$, $j = 1, \dots, k$.
- (v) $\sum_{i \in S_l(z_j)} p w_{ij} |z_{jl} - x_{il}|^{p-1} \text{sgn}(z_{jl} - x_{il}) = 0$, $l = 1, \dots, m$, $j = 1, \dots, k$.

Assume $(\mathbf{W}^*, \mathbf{Z}^*)$ is a partial optimal solution. By the definition of partial optimal solutions, \mathbf{W}^* solves Problem (P1) given $\mathbf{Z} = \mathbf{Z}^*$, and thus \mathbf{W}^* satisfies the Kuhn-Tucker conditions of

Problem (P1) characterized by (i)–(iv) when $\mathbf{Z} = \mathbf{Z}^*$. Likewise, \mathbf{Z}^* solves Problem (P2) given $\mathbf{W} = \mathbf{W}^*$, and thus \mathbf{Z}^* satisfies the Kuhn-Tucker condition of Problem (P2) characterized by (v) when $\mathbf{W} = \mathbf{W}^*$.

Assume $(\mathbf{W}^*, \mathbf{Z}^*)$ is a Kuhn-Tucker point of Problem (P). Suppose that $(\mathbf{W}^*, \mathbf{Z}^*)$ is not a partial optimal solution of Problem (P). Either \mathbf{W}^* does not solve Problem (P1) with $\mathbf{Z} = \mathbf{Z}^*$, or \mathbf{Z}^* does not solve Problem (P2) with $\mathbf{W} = \mathbf{W}^*$. If the former is true, then \mathbf{W}^* should not solve Kuhn-Tucker conditions of Problem (P1) with $\mathbf{Z} = \mathbf{Z}^*$, which are given by (i)–(iv). If the latter is true, then \mathbf{Z}^* should not solve Kuhn-Tucker condition of Problem (P2) with $\mathbf{W} = \mathbf{W}^*$, which is given by (v). This is a contradiction and hence $(\mathbf{W}^*, \mathbf{Z}^*)$ is a partial optimal solution. This completes the proof. \square

It is worth noting that when Algorithm 1 terminates at a partial optimal solution $(\mathbf{W}^*, \mathbf{Z}^*)$, if \mathbf{Z}^* is the unique solution of Problem (P2) with $\mathbf{W} = \mathbf{W}^*$, the local optimality of \mathbf{W}^* for Problem (RP) can be guaranteed. This, in turn, makes $(\mathbf{W}^*, \mathbf{Z}^*)$ a local minimizer of Problem (P), according to Selim and Ismail [32].

5 Numerical Experiments

This section provides empirical evidence regarding the effectiveness of our proposed sub-one quasi-norm-based k -means algorithm on both synthetic and real-life datasets, which also serves as a comparative analysis of the proposed ℓ_p quasi-norm-based distance metric against other commonly used distance metrics. Additionally, we further investigate the applicability of our k -means method for handling complex tasks by implementing it within the deep learning based clustering algorithm.

5.1 Experiment Setup

Simulations were executed on all datasets employing the proposed Algorithm 1 and the traditional Lloyd's k -means algorithms with the basic and commonly used distance metrics in most applications. Specifically, the proposed ℓ_p quasi-norm-based distance metric with $p = 0.5$ was tested against conventional distance metrics, including the Manhattan distance (referred to as l_1 distance) and the squared Euclidean distance (referred to as l_2 distance), as detailed in Table 3. Moreover, when applicable, the S -distance, which is chosen among various divergence-based distance functions due to its demonstrated superiority over conventional distance metrics [7], was included in our comparisons.

The various k -means algorithms were executed on the same dataset to facilitate a direct comparison of their clustering results. Given that the Lloyd's k -means clustering process produces a local optimal solution dependent on the initial cluster centers, the implementation of a k -means algorithm was repeated 10 times for each dataset with different initial cluster centers in each iteration. This ensures that any observed performance differences are not due to chance or potential biases stemming from a single initial cluster configuration. In every iteration, a common set of randomly chosen initial cluster centers was employed for for all instances of the k -means algorithm involved in the comparison. Every dataset employed in our experiments was equipped with known class labels for each instance. To safeguard against any impact on the algorithm performance due to the accurate or inaccurate chosen value for k , the input parameter k was set to be the true number of clusters when executing the k -mean algorithms.

Table 3 Different distance metrics under comparison

Distance metric	Definition	Defined space
$l_{0.5}$ quasi-norm-based distance	$\sum_{i=1}^m x_i - y_i ^{\frac{1}{2}}$	\mathbb{R}^m
Manhattan distance (ℓ_1 distance)	$\sum_{i=1}^m x_i - y_i $	\mathbb{R}^m
Squared Euclidean distance (ℓ_2 distance)	$\sum_{i=1}^m (x_i - y_i)^2$	\mathbb{R}^m
S -distance	$\log \prod_{i=1}^m \frac{x_i + y_i}{2} - \frac{\log \prod_{i=1}^m x_i + \log \prod_{i=1}^m y_i}{2}$	\mathbb{R}_{++}^m

We employed the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) to evaluate the similarity between the cluster assignments generated by the k -means algorithms and the true labels. The ARI quantifies the agreement between two class assignments, where a score of 1 indicates a perfect match, and -1 indicates a complete mismatch. Meanwhile, NMI scales the similarity between clustering outcomes and true labels on a range from 0 (indicating no mutual information) to 1 (reflecting perfect correlation). Higher ARI and NMI values signify that the clustering results closely align with the ground-truth.

5.2 Experiments on Synthetic Dataset

To validate the clustering performance of the proposed sub-one quasi-norm-based k -means algorithm, we conducted experiments on synthetic datasets constructed with mixtures of multivariate normal distributions. Initially, we generated multiple clusters following a multivariate Gaussian distribution, each with a mean of 0 and a covariance/correlation matrix denoted as

$$\theta = \begin{bmatrix} 1 & b & \dots & b \\ b & 1 & \dots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \dots & 1 \end{bmatrix},$$

where $0 < b < 1$ is a predefined parameter given by $b = \frac{K^2 - 1}{n - 1 + K^2}$, with $K > 0$ representing the ratio of the maximum eigenvalue to the minimal eigenvalue of θ . This results in multiple normally distributed clusters spreading along the direction of $(1, \dots, 1)^T$, and each variable has identical isotropic covariances. To introduce diversity among clusters, we rotated them to different directions and varied their distances/densities by adding/multiplying them with distinct arbitrary vectors. The final dataset was then formed by overlapping multiple clusters generated in this manner to simulate a multi-cluster scenario.

We conducted experiments on a total of eleven synthetic datasets that encompass a diverse range of scenarios. The first five datasets (c2, c3, c4, c5, c6) are standard multi-cluster datasets, each consisting of 2, 3, 4, 5, 6 clusters, respectively. Each cluster contains 100 records. The remaining six datasets were specifically designed to incorporate various clustering challenges, as summarized below.

- Imbalanced datasets (im3, im4): Dataset im3 comprises 300 records distributed across into three clusters, with the clusters containing 50, 100, and 150 records, respectively. Dataset im4 contains 300 records divided into four clusters, each consisting of 50, 100, 150, 200 records, respectively.

- Datasets with noise (n3, n4): In these datasets, each comprising 3 and 4 clusters respectively, 20% of the data points were augmented with normally distributed noise characterized by a mean of 0 and a standard deviation of 0.1.
- Datasets with outliers (o3, o4): Featuring 3 and 4 clusters respectively, each cluster in these datasets incorporates 20% artificial outliers. These outliers are generated from a uniform distribution on a simplex, where the simplex's vertices were randomly chosen from the intersection of an ℓ_2 norm ball with a radius of 2 and a hyperplane defined by a bias term of 1 and a normal vector forming an angle of 45° with the major direction of the corresponding cluster.

As the S -distance metric is exclusively defined on \mathbb{R}_{++}^m , we compare the performance scores obtained by the k -means algorithms using $\ell_{0.5}$, ℓ_1 , ℓ_2 distances. Table 4 reports the average and standard deviation of ARI and NMI for the clustering results. The table highlights the best ARI and NMI mean values across different algorithms for each experiment instance in bold. Notably, the proposed k -means algorithm utilizing the $\ell_{0.5}$ distance metric generally demonstrates superior performance in comparison to other algorithms on these synthetic datasets. This superiority can be attributed to the non-convex shape of the ℓ_p quasi-norm balls with $p \in (0, 1)$, which enables the ℓ_p quasi-norm-based distance to make better use of similar data-items rather than over-emphasizing dissimilar ones. The enhanced clustering performance of our proposed k -means algorithm on datasets c2, c3, c4, c5, c6 is visually evident from Fig. 3. As the cluster number increases, resulting in clusters with significant variations in every dimension, clustering accuracy naturally decreases. Nevertheless, our proposed algorithm maintains its effectiveness. For datasets with a high cluster number, we observed suboptimal and comparable performances across all k -means algorithms. Additionally, the superior clustering performance of our proposed k -means algorithm on datasets im3, im4, n3, n4, o3, o4 is visually evident from Fig. 4. This highlights the algorithm's efficacy in handling datasets with imbalanced clusters, noise, and outliers. The distinctive features of our proposed algorithm position it as a robust and versatile solution to capture complex patterns and overcome challenges posed by noise and outliers.

5.3 Experiments on Real-Life Dataset

To verify the effectiveness of the proposed k -means algorithm for clustering real-life data, we conducted experiments on several datasets from the UCI Machine Learning Repository [1, 8, 19]. The details of datasets are provided below.

- Wine: This dataset contains 178 instances, each including 13 integer-valued constituents found in wines grown in the same region in Italy but derived from three different cultivars.
- Iris: This dataset consists of three different types of iris plants: Setosa, Versicolor, Virginica. Each type contains 50 instances and each instance has four real-valued attributes.
- Seeds: This dataset contains 210 instances, and each instance includes seven real-valued continuous geometric measurements of wheat kernels belonging to three different varieties of wheat.
- Ceramic: This dataset includes 88 instances that can be classified into body or glaze based on 17 chemical compositions obtained from energy dispersive X-ray fluorescence.
- Algerian forest fires: This dataset comprises 244 instances of forest regions in Algeria, which includes ten attributes and can be classified into fire and not fire classes.
- Facebook live sellers: This dataset contains Facebook pages of Thai fashion and cosmetics retail sellers. The 6622 instances with nine integer-valued attribute can be divided into two classes: video and photo.

Table 4 ARI/NMI levels at the synthetic datasets

Dataset	Distance metric	ARI		NMI	
		Mean	Std Dev	Mean	Std Dev
c2	$\ell_{0,5}$ distance	0.6107	0.2500	0.5426	0.1211
	ℓ_1 distance	0.4969	0.0945	0.5070	0.0625
	ℓ_2 distance	0.2049	0.0069	0.2639	0.0007
c3	$\ell_{0,5}$ distance	0.7787	0.0211	0.7411	0.1366
	ℓ_1 distance	0.6867	0.1490	0.7148	0.1395
	ℓ_2 distance	0.4432	0.0000	0.5373	0.0167
c4	$\ell_{0,5}$ distance	0.6770	0.0406	0.7204	0.0973
	ℓ_1 distance	0.6587	0.0349	0.6666	0.1564
	ℓ_2 distance	0.4018	0.0544	0.4895	0.0352
c5	$\ell_{0,5}$ distance	0.6065	0.0000	0.7155	0.1386
	ℓ_1 distance	0.5376	0.0000	0.6392	0.0694
	ℓ_2 distance	0.4386	0.0000	0.5347	0.0732
c6	$\ell_{0,5}$ distance	0.3827	0.1237	0.6511	0.0862
	ℓ_1 distance	0.3805	0.0910	0.6399	0.0511
	ℓ_2 distance	0.3189	0.0194	0.5910	0.0340
im3	$\ell_{0,5}$ distance	0.7340	0.2195	0.7902	0.1943
	ℓ_1 distance	0.7173	0.2030	0.7164	0.0547
	ℓ_2 distance	0.4917	0.0026	0.6118	0.0128
im4	$\ell_{0,5}$ distance	0.4068	0.0775	0.5856	0.0659
	ℓ_1 distance	0.3799	0.0393	0.5564	0.0121
	ℓ_2 distance	0.3727	0.0307	0.5424	0.0027
n3	$\ell_{0,5}$ distance	0.6253	0.2415	0.6534	0.1336
	ℓ_1 distance	0.5800	0.1834	0.6196	0.0871
	ℓ_2 distance	0.4175	0.0241	0.5105	0.0190
n4	$\ell_{0,5}$ distance	0.4864	0.2170	0.7184	0.0127
	ℓ_1 distance	0.5093	0.1670	0.7039	0.0209
	ℓ_2 distance	0.5060	0.0929	0.6603	0.0064
o3	$\ell_{0,5}$ distance	0.5101	0.2019	0.6219	0.1340
	ℓ_1 distance	0.4500	0.1327	0.5978	0.0822
	ℓ_2 distance	0.4523	0.1384	0.5924	0.0149
o4	$\ell_{0,5}$ distance	0.4831	0.0456	0.5926	0.0728
	ℓ_1 distance	0.4592	0.0440	0.6180	0.0580
	ℓ_2 distance	0.4375	0.0073	0.6110	0.0007

Table 5 provides the average and standard deviation of ARI and NMI for clustering results across the six real-life datasets, which underscores the superior performance of the proposed distance metric in capturing the underlying structures in specific real-life datasets. The table highlights the best ARI and NMI mean values across different algorithms for each experiment instance in bold. On the Wine, Ceramic, Forest fire, and Facebook live seller datasets, the k -means algorithm using the $\ell_{0,5}$ quasi-norm-based distance demonstrates superior clustering accuracy compared to the other algorithms.

However, on the Seed dataset, while the proposed k -means algorithm maintains comparable performance, it does not exhibit a discernible advantage over the other algorithms. In contrast, on dataset Iris, the utilization of the $\ell_{0.5}$ quasi-norm-based distance metric yields improved clustering results compared to both Manhattan and Euclidean distances, although it falls short of surpassing the performance achieved by the S -distance.

To assess the robustness of the proposed sub-one quasi-norm based k -means clustering algorithm, we conducted experiments on intentionally corrupted datasets with added noise. Specifically, we introduced a pre-specified amount of noise into two real-life datasets, namely Iris and Seeds, where the ℓ_p quasi-norm-based distance metric does not yield the best results in Table 5. We generated normally distributed noise with a mean of 0 and applied to approximately 10%, 20%, 30% of the data points in each dataset. To account for the different scales of variations across variables, the standard deviations of the added noise were set equal to the mean value within each variable. The clustering experiments involved 10 trials for each combination of distance metric and noise percentage. The presence of noise introduces challenges in recovering the underlying cluster structures of the dataset.

Table 6 presents the average and standard deviation of ARI and NMI for k -means clustering results on the noisy data, categorized by datasets, noise percentage, and distance metrics. The table highlights the best ARI and NMI mean values across different algorithms for each experiment instance in bold. A comparison between Tables 5 and 6 reveals a deterioration in the quality of clustering results when noise is present in the data. However, both datasets consistently indicate the proposed $\ell_{0.5}$ quasi-norm-based distance produces the best clustering results when compared to other distance metrics. This underscores the effectiveness of the proposed k -means method in clustering data while mitigating the adverse effects of noise. This suggests that the k -means clustering method with ℓ_p quasi-norm-based distance exhibits resistance to noise and outliers in the data, highlighting the efficacy and robustness of the proposed algorithm.

5.4 Experiments with Deep-Learning Models

To investigate the practical applicability of the sub-one quasi-norm-based metric, we have integrated the proposed k -means algorithm into advanced deep learning models. Evaluating the clustering algorithm within a deep learning context provides valuable insights into its applicability for more complex tasks and real-world scenarios, especially when dealing with high-dimensional data.

This evaluation was conducted on the cutting-edge generalized deep learning clustering (GDLC) algorithm [36]. The GDLC algorithm focused on dimensionality reduction by finding the product of two low-dimensional matrices that approximates the original high-dimensional matrix. This deep learning approach enables the representation of high-dimensional data in a lower-dimensional space and facilitates the completion of clustering tasks in this reduced space. In the GDLC framework, a nonlinear constrained non-negative matrix factorization based on stochastic gradient descent is initially employed as an element update method. The resulting generalized weights, generalized biases, and activation function are then combined to form a generalized deep learning network for updating elements in the low-dimensional matrix.

In the original GDLC algorithm, an ℓ_2 -norm-based k -means algorithm is employed for clustering. Notably, we replaced a segment of the existing GDLC's clustering process with

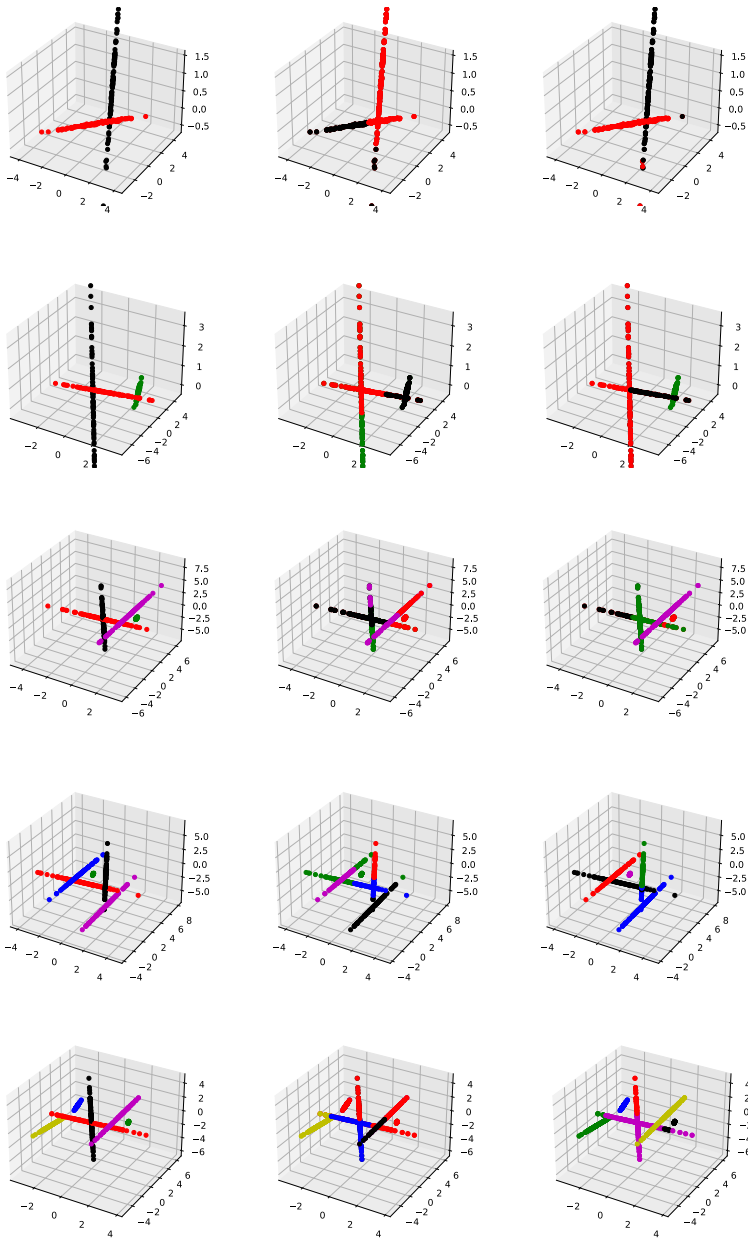


Fig. 3 The figure shows the results obtained on standard multi-cluster datasets with varying cluster numbers. Each row corresponds to a different dataset (c2, c3, c4, c5, c6). Column 1 illustrates the original cluster structure. Columns 2 and 3 indicate the class assignments generated by the Lloyd's k -means algorithm with the squared Euclidean distance and the proposed k -means algorithm with the $\ell_{0.5}$ quasi-norm-based distance metric, respectively

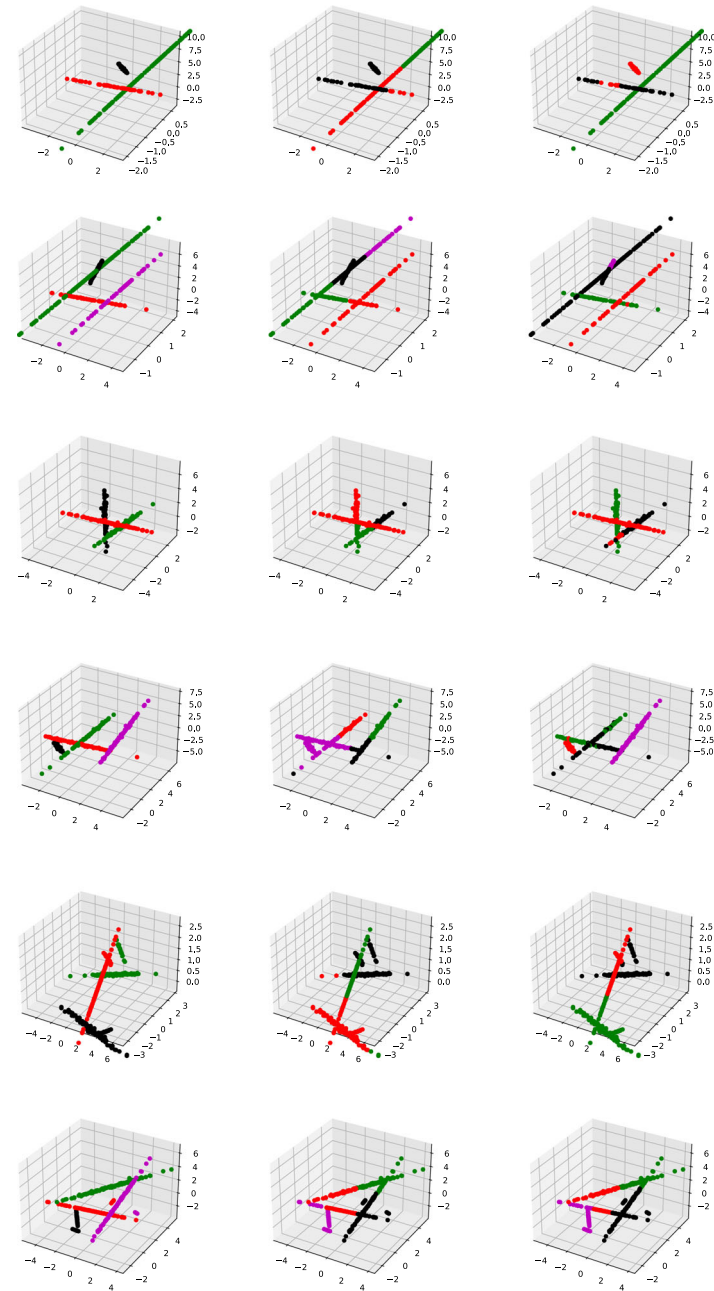


Fig. 4 The figure shows the results obtained on the complex synthetic datasets. Each row corresponds to a different dataset (im3, im4, n3, n4, o3, o4). Column 1 illustrates the original cluster structure. Columns 2 and 3 indicate the class assignments generated by the Lloyd's k -means algorithm with the squared Euclidean distance and the proposed k -means algorithm with the $\ell_{0.5}$ quasi-norm-based distance metric, respectively

Table 5 ARI/NMI levels on the real-life datasets

Dataset	Distance metric	ARI		NMI	
		Mean	Std Dev	Mean	Std Dev
Wine	$\ell_{0.5}$ distance	0.7389	0.0622	0.6774	0.0380
	ℓ_1 distance	0.1379	0.0042	0.1516	0.0030
	ℓ_2 distance	0.1106	0.0027	0.1127	0.0000
	S-distance	0.6484	0.0005	0.6485	0.0000
Iris	$\ell_{0.5}$ distance	0.6840	0.1008	0.7118	0.0082
	ℓ_1 distance	0.6663	0.1295	0.7507	0.0000
	ℓ_2 distance	0.6617	0.1539	0.7582	0.0000
	S-distance	0.7852	0.1795	0.8642	0.0000
Seed	$\ell_{0.5}$ distance	0.6981	0.0000	0.7011	0.0070
	ℓ_1 distance	0.6994	0.0000	0.7029	0.0000
	ℓ_2 distance	0.7103	0.0000	0.7101	0.0000
	S-distance	0.5059	0.1482	0.5041	0.0000
Ceramic	$\ell_{0.5}$ distance	0.8244	0.0000	0.7766	0.0000
	ℓ_1 distance	0.7610	0.0000	0.7387	0.0000
	ℓ_2 distance	0.5925	0.0000	0.5868	0.0000
	S-distance	0.5028	0.0000	0.5006	0.0000
Forest fire	$\ell_{0.5}$ distance	0.5274	0.0109	0.4374	0.0000
	ℓ_1 distance	0.2032	0.0131	0.2842	0.0000
	ℓ_2 distance	0.0628	0.0000	0.1477	0.0000
	S-distance	0.3116	0.0029	0.2516	0.0000
Facebook live seller	$\ell_{0.5}$ distance	0.1972	0.1178	0.1999	0.0733
	ℓ_1 distance	0.1604	0.0978	0.1519	0.0343
	ℓ_2 distance	0.0603	0.0000	0.0267	0.0000
	S-distance	0.0787	0.0008	0.1130	0.0000

our proposed k -means algorithm. This modification allows us to compare the performance of two GDLC versions utilizing different metrics. Our experiments involve eight public datasets that pose a significant data challenge due to their high dimensional feature spaces. These datasets encompass six biographical datasets, one image dataset, and one text dataset [3, 6, 27].

- ALLAML: This dataset contains a total of 72 instances classified into two categories, ALL and AML. Each instance includes 7129 gene expression values.
- LUNG_DISCRETE: This dataset features 73 instances distributed among seven classes. Each instance consists of 325 gene expression values.
- ARCENE: This dataset contains 200 instances with 10000 features derived from mass-spectrometric data, categorized in two classes: cancer pattern and normal pattern.
- leukemia: This dataset consists of 72 bone marrow instances based on probes from human genes, with two classes: acute lymphoblastic and acute myeloid.
- colon: This dataset contains 62 instances with 2000 genes, divided into two classes: tumor tissues and normal tissues.
- nci9: This biological dataset encompasses 60 instances, each characterized by 9712 features and classified into nine distinct classes.

Table 6 ARI/NMI levels on the Seeds and Iris datasets with noise

Dataset	Noise percentage	Distance metric	ARI		NMI	
			Mean	Std Dev	Mean	Std Dev
Seeds	10%	$\ell_{0.5}$ distance	0.6981	0.0048	0.6898	0.0038
		ℓ_1 distance	0.6895	0.0000	0.6792	0.0000
		ℓ_2 distance	0.6329	0.0000	0.5748	0.0000
		S-distance	0.4998	0.0000	0.4719	0.0000
	20%	$\ell_{0.5}$ distance	0.6681	0.0062	0.6443	0.0000
		ℓ_1 distance	0.6607	0.0000	0.6285	0.0000
		ℓ_2 distance	0.5754	0.0000	0.5668	0.0000
		S-distance	0.4295	0.0000	0.4069	0.0000
	30%	$\ell_{0.5}$ distance	0.6394	0.0000	0.5899	0.0137
		ℓ_1 distance	0.6170	0.0056	0.5368	0.0000
		ℓ_2 distance	0.5269	0.0000	0.4515	0.0000
		S-distance	0.3972	0.0000	0.4138	0.0000
Iris	10%	$\ell_{0.5}$ distance	0.6609	0.1039	0.7281	0.0745
		ℓ_1 distance	0.6031	0.0935	0.6617	0.0647
		ℓ_2 distance	0.4789	0.0168	0.5415	0.0101
		S-distance	0.4846	0.0137	0.6025	0.0135
	20%	$\ell_{0.5}$ distance	0.5192	0.1383	0.6301	0.0572
		ℓ_1 distance	0.4570	0.1191	0.5415	0.0156
		ℓ_2 distance	0.3293	0.0209	0.4298	0.0315
		S-distance	0.4033	0.0033	0.4676	0.0153
	30%	$\ell_{0.5}$ distance	0.5288	0.0790	0.5666	0.0333
		ℓ_1 distance	0.4228	0.0553	0.4713	0.0296
		ℓ_2 distance	0.367	0.0300	0.3840	0.0404
		S-distance	0.3508	0.0012	0.4317	0.0023

- **ORL**: This dataset comprises 400 instances representing ten distinct images of 40 different subjects. Each instance is characterized by 1024 features and falls into one of eleven classes.
- **RELATHE**: This dataset comprises 1427 instances, each defined by 4322 features, and is organized into two distinct newsgroups.

Table 7 presents the average and standard deviation of ARI and NMI for the clustering results resulting from 10 random implementations of the GDLC algorithm. The table highlights the best ARI and NMI mean values across different algorithms for each experiment instance in bold. On the ALLAML, LUNG_DISCRETE, ARCENE, leukemia, colon, and nci9 datasets, the incorporation of our proposed k -means algorithm into GDLC outperforms the original GDLC algorithm in terms of ARI or NMI. This underscores the effectiveness of our proposed metric in capturing distinctive patterns exhibited by certain data types, such as biological data involving gene expression patterns or molecular profiles. However, the updated GDLS algorithm demonstrates relatively lower performance on the ORL and RELATHE datasets, indicating challenges in handling the inherent complexities associated with text and image datasets. These datasets typically possess different feature spaces and structures that require clustering algorithms

Table 7 ARI/NMI levels for the GDLC algorithm using different distances

Dataset	Distance metric	ARI		NMI	
		Mean	Std Dev	Mean	Std Dev
ALLAML	$\ell_{0.5}$ distance	0.4946	0.0008	0.1379	0.0476
	ℓ_2 distance	0.5290	0.0129	0.0851	0.0310
LUNG_DISCRETE	$\ell_{0.5}$ distance	0.7449	0.0143	0.2244	0.0224
	ℓ_2 distance	0.7342	0.0037	0.2035	0.0385
ARCENE	$\ell_{0.5}$ distance	0.5068	0.0018	0.0114	0.0040
	ℓ_2 distance	0.4979	0.0001	0.0003	0.0001
leukemia	$\ell_{0.5}$ distance	0.4965	0.0000	0.1686	0.0191
	ℓ_2 distance	0.5729	0.0175	0.0847	0.0070
colon	$\ell_{0.5}$ distance	0.5270	0.0163	0.0186	0.0057
	ℓ_2 distance	0.4921	0.0007	0.0002	0.0004
nci9	$\ell_{0.5}$ distance	0.8288	0.0157	0.4210	0.0309
	ℓ_2 distance	0.8178	0.0077	0.3883	0.0366
ORL	$\ell_{0.5}$ distance	0.8636	0.1654	0.4893	0.0636
	ℓ_2 distance	0.9760	0.0110	0.7827	0.1536
RELATHE	$\ell_{0.5}$ distance	0.5028	0.0011	0.0118	0.0157
	ℓ_2 distance	0.9397	0.0043	0.8276	0.0148

to adapt to diverse modalities. This discrepancy in performance across data types indicates the sensitivity of our proposed k -means algorithm to the intrinsic characteristics and structures present in different types of data, emphasizing its context-dependent applicability.

6 Summary

The k -means algorithm is a widely-used clustering algorithm that partitions a given dataset into k clusters based on their similarity. In this study, we propose an enhancement to the k -means algorithm by employing the ℓ_p quasi-norm-based distance metric, specifically with $p \in (0, 1)$. Compared to commonly used distance metrics, our proposed sub-one quasi-norm-based distance metric excels at capturing similarities among data items while mitigating the overemphasis on dissimilarities. We substantiate the effectiveness of our proposed k -means algorithm through theoretical analysis, providing a proof of convergence to the Kuhn-Tucker point. Extensive experimental results on synthetic and real-life datasets further highlight the superiority of our algorithm over other distance metrics across diverse scenarios. Notably, for datasets with added noise, the sub-one ℓ_p quasi-norm-based distance metric successfully addresses a known drawback of k -means, demonstrating its robustness against noise. We also integrated our proposed k -means method into the generalized deep learning clustering algorithm that indicates its sensitivity to the inherent characteristics and structures of data and its context-dependent applicability. As a prospective avenue, extending the implementation of

the sub-one ℓ_p quasi-norm-based distance to Fuzzy c -means (FCM) type algorithms presents a promising direction for future research.

Author Contributions Q.A. developed the theory and performed the numerical experiments. S.J. derived the models and analyzed the data. Q.A. and S.J. prepared and reviewed the manuscript.

Funding This work was supported by the National Natural Science Foundation of China (NSFC Grant No. 12201528), the Research Funds for Young Scholars in the Open University of China (No. Q22F0029), and the Fundamental Research Funds for the Central Universities of Xiamen University (No. 2072021127).

Declarations

Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Abid F, Izeboudjen N (2019) Predicting forest fire in Algeria using data mining techniques: Case study of the decision tree algorithm. *Int Conf Adv Intell Syst Sustain Dev* 1105:363–370. https://doi.org/10.1007/978-3-030-36674-2_37
2. Aloise D, Deshpande A, Hansen P et al (2009) NP-hardness of Euclidean sum-of-squares clustering. *Mach learn* 75:245–248. <https://doi.org/10.1007/s10994-009-5103-0>
3. Alon U, Barkai N, Notterman DA et al (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci* 96(12):6745–6750. <https://doi.org/10.1073/pnas.96.12.6745>
4. Banerjee A, Merugu S, Dhillon IS et al (2005) Clustering with Bregman divergences. *J Mach Learn Res* 6(10):1705–1749. <https://doi.org/10.5555/1046920.1194902>
5. Bobrowski L, Bezdek JC (1991) c -means clustering with the l_1 and l_∞ norms. *IEEE Transact Syst Man Cybern* 21(3):545–554. <https://doi.org/10.1109/21.97475>
6. Cai D, He X, Hu Y et al (2007) Learning a spatially smooth subspace for face recognition. 2007 IEEE Conf Comput Vis Pattern Recognit pp 1–7. <https://doi.org/10.1109/CVPR.2007.383054>
7. Chakraborty S, Das S (2017) k -means clustering with a new divergence-based distance metric: Convergence and performance analysis. *Pattern Recognit Lett* 100:67–73. <https://doi.org/10.1016/j.patrec.2017.09.025>
8. Charytanowicz M, Niewczas J, Kulczycki P et al (2010) Complete gradient clustering algorithm for features analysis of X-ray images. *Inf Technol Biomed* 69:15–24. https://doi.org/10.1007/978-3-642-13105-9_2
9. Chen WJ, Tian YJ (2010) l_p -norm proximal support vector machine and its applications. *Proced Comput Sci* 1(1):2417–2423. <https://doi.org/10.1016/j.procs.2010.04.272>
10. Chiu SM, Chen YC, Chang TY et al (2016) (2016) A fast way for finding similar friends in social networks by using neuro-fuzzy networks. *Int Conf Mach Learn Cybern* 2:541–545. <https://doi.org/10.1109/ICMLC.2016.7872945>
11. de Amorim RC, Mirkin B (2012) Minkowski metric, feature weighting and anomalous cluster initializing in k -means clustering. *Pattern Recognit* 45(3):1061–1075. <https://doi.org/10.1016/j.patcog.2011.08.012>
12. Deng P, Li T, Wang H et al (2023) Graph regularized sparse non-negative matrix factorization for clustering. *IEEE Trans Comput Soc Syst* 10(3):910–921. <https://doi.org/10.1109/TCSS.2022.3154030>
13. Dhillon IS, Mallela S, Kumar R (2003) A divisive information theoretic feature clustering algorithm for text classification. *The J Mach learn Res* 3:1265–1287. <https://doi.org/10.5555/944919.944973>
14. Duda RO, Hart PE, Stork DG (2000) *Pattern Classif*, 2nd edn. Wiley-Interscience, USA

15. Filippone M, Camastra F, Masulli F et al (2008) A survey of kernel and spectral methods for clustering. *Pattern Recognit* 41(1):176–190. <https://doi.org/10.1016/j.patcog.2007.05.018>
16. Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Ann Eugen* 7(2):179–188. <https://doi.org/10.1111/j.1469-1809.1936>
17. Ge D, Jiang X, Ye Y (2011) A note on the complexity of l_p minimization. *Math Program* 129:285–299. <https://doi.org/10.5555/3119419.3119628>
18. Hathaway RJ, Bezdek JC, Hu Y (2000) Generalized fuzzy c -means clustering strategies using l_p norm distances. *IEEE Trans Fuzzy Syst* 8(5):576–582. <https://doi.org/10.1109/91.873580>
19. He Z, Zhang M, Zhang H (2016) Data-driven research on chemical features of Jingdezhen and Longquan celadon by energy dispersive X-ray fluorescence. *Ceram Int* 42(4):5123–5129. <https://doi.org/10.1016/j.ceramint.2015.12.030>
20. Jain AK (2010) Data clustering: 50 years beyond k -means. *Pattern Recognit Lett* 31(8):651–666. <https://doi.org/10.1016/j.patrec.2009.09.011>
21. Jain AK, Dubes RC (1988) *Algorithms for Clustering Data*. Prentice-Hall, USA. <https://doi.org/10.5555/46712>
22. Jain P, Kar P (2017) Non-convex optimization for machine learning. *Found Trends Mach Learn* 10(3–4):142–363. <https://doi.org/10.1561/22000000058>
23. Jiang S, Fang SC, An Q et al (2019) A sub-one quasi-norm-based similarity measure for collaborative filtering in recommender systems. *Inf Sci* 487:142–155. <https://doi.org/10.1016/j.ins.2019.03.011>
24. Kabashima Y, Wadayama T, Tanaka T (2009) A typical reconstruction limit for compressed sensing based on l_p -norm minimization. *J Stat Mech: Theory Exp* 2009:L09003. <https://doi.org/10.1088/1742-5468/2009/09/L09003>
25. Kersten PR (1997) Implementation issues in the fuzzy c -medians clustering algorithm. *Proc 6th Int Fuzzy Syst Conf* 2:957–962. <https://doi.org/10.1109/FUZZY.1997.622838>
26. MacQueen JB (1967) Some methods for classification and analysis of multivariate observations. In: *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* pp 281–297. <https://api.semanticscholar.org/CorpusID:6278891>
27. Nardone D (2019) Biological datasets for SMBA. Zenodo. <https://doi.org/10.5281/zenodo.2709491>
28. Nie F, Wang H, Cai X et al (2012) Robust matrix completion via joint Schatten p -norm and l_p -norm minimization. *2012 IEEE 12th Int Conf Data Mining* pp 566–574. <https://doi.org/10.1109/ICDM.2012.160>
29. Nielsen F, Nock R, Si Amari (2014) On clustering histograms with k -means by using mixed α -divergences. *Entropy* 16(6):3273–3301. <https://doi.org/10.3390/e16063273>
30. Saha A, Das S (2016) Geometric divergence based fuzzy clustering with strong resilience to noise features. *Pattern Recognit Lett* 79:60–67. <https://doi.org/10.1016/j.patrec.2016.04.013>
31. Seal A, Karlekar A, Krejcar O et al (2020) Fuzzy c -means clustering using Jeffreys-divergence based similarity measure. *Appl Soft Comput* 88:106016. <https://doi.org/10.1016/j.asoc.2019.106016>
32. Selim SZ, Ismail MA (1984) k -means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Trans Pattern Anal Mach Intell PAMI-6* 1:81–87. <https://doi.org/10.1109/tpami.1984.4767478>
33. Singh A, Yadav A, Rana A (2013) k -means with three different distance metrics. *Int J Comput Appl* 67(10):13–17. <https://doi.org/10.5120/11430-6785>
34. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J Royal Stat Soc: Ser B (Methodol)* 58(1):267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
35. Wang D, Li T, Deng P et al (2023) A generalized deep learning algorithm based on nmf for multi-view clustering. *IEEE Trans on Big Data* 9:328–340. <https://api.semanticscholar.org/CorpusID:247874882>
36. Wang D, Li T, Deng P et al (2023) A generalized deep learning clustering algorithm based on non-negative matrix factorization. *ACM Trans Knowl Discov from Data* 17(7):1–20. <https://doi.org/10.1145/3584862>
37. Wang D, Li T, Huang W et al (2023) A multi-view clustering algorithm based on deep semi-nmf. *Inf Fusion* 99(C). <https://doi.org/10.1016/j.inffus.2023.101884>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.