Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 246 (2024) 830–839

Procedia Computer Science

www.elsevier.com/locate/procedia

28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)

Aggregations of Fuzzy Equivalences in *k*-means Algorithm

Piotr Lasek, Wojciech Rząsa, Anna Król

University of Rzeszów, Pigonia 1, 35-310 Rzeszów, Poland

Abstract

This paper proposes a new approach to the *k*-means clustering algorithm incorporating a variety of fuzzy equivalences and aggregation techniques. Central to this approach is the use of fuzzy equivalences, which serve as an alternative to standard distance metrics, with the goal of improving the clustering process. The modified *k*-means algorithm employs these fuzzy equivalences to better determine the similarity between the data points, particularly in situations where the nearest points may not adequately represent closeness within the data set. To assess the clustering results, we employ a variation of the silhouette coefficient tailored to our method. Additionally, we present theoretical insights into the behavior and benefits of using composition of aggregations and fuzzy equivalences in clustering. Experimental validations carried out on diverse data sets indicate that our method can lead to improved clustering outcomes compared to traditional techniques.

Keywords: *k*-means algorithm, fuzzy equivalence, distance metrics, aggregation methods, silhouette coefficient, cluster quality evaluation.

1. Introduction

Clustering algorithms, with a focus on the *k*-means algorithm, are important in data analysis for grouping data based on similarities. Despite its application, the *k*-means algorithm encounters some difficulties with complex data structures and high-dimensional spaces. This paper proposes a modified *k*-means algorithm that incorporates fuzzy equivalences and a variety of aggregation methods to boost its performance.

The conventional *k*-means algorithm employs Euclidean distance to assess a closeness-distance between data points and cluster centers. Although this is suitable in many cases, it struggles in high-dimensional spaces or with data that does not form neat, spherical clusters. Moreover, the initial choice of centroids can significantly influence the

* Corresponding author. Tel.: +48 17 851 8516
E-mail address: plasek@ur.edu.pl

1877-0509
10.1016/j.procs.2024.09.502

outcome, sometimes resulting in suboptimal clustering. These challenges underscore the necessity for a more flexible and resilient clustering method.

To mitigate some of the Euclidean distance's limitations, this study explores alternative ways of finding similar objects. Combining different aggregations and fuzzy equivalences as measures of closeness offers a strategy for refining the clustering process. Such compositions can better represent the true relationships between data points, allowing for a clustering process customized to the data set's unique characteristics. The significance of such flexibility for achieving precise clustering results is underscored by this aggregation approach.

Fuzzy logic introduces a distinctive method for defining and evaluating similarity, accommodating degrees of similarity beyond a binary decision. This technique is beneficial in scenarios where there is ambiguity in the data or there is no clear distinction between groups. Employing fuzzy equivalences, this document proposes a method for evaluating closeness that can navigate the intrinsic uncertainty in real-world data, potentially leading to more relevant clustering results.

To effectively assess the outcomes of the enhanced k-means algorithm, this research adapts the silhouette coefficient, which, tailored to work with compositions of aggregations and fuzzy equivalences, offers a detailed perspective on cluster quality, focusing on intra-cluster cohesion and inter-cluster separation.

Research conducted on multiple data sets confirms the suggested changes in the k-means algorithm, demonstrating that the incorporation of fuzzy equivalences and aggregations improves clustering, especially in complex data sets, by giving the analyst more flexibility to to adjust the similarity measure when running a clustering algorithm. In addition, the paper investigates the theoretical foundations of these enhancements, analyzing how they lead to a more adaptable and accurate method of clustering.

The document is structured to examine the suggested changes to the k-means algorithm. Following this introduction, Section 2 reviews prior studies, setting the stage for our contributions by evaluating existing methods' strengths and limitations. Section 3 delves into our approach's theoretical basis, detailing the integration of fuzzy logic with the k-means algorithm for more adaptable and nuanced clustering. This includes the mathematical models used and the rationale behind selecting specific fuzzy equivalences and aggregation methods. Section 4 outlines the validation data sets, experimental design, and criteria for assessing the modified algorithm's effectiveness. It describes the testing procedures and the comparison of the enhanced k-means algorithm across different data sets. The methodology for calculating the modified silhouette coefficient and its role in evaluating clustering quality is also elaborated. This section also shares the experimental validation outcomes, emphasizing how fuzzy equivalences and aggregated distance metrics affect clustering. This section contrasts our approach with traditional k-means implementations, discussing the findings within the context of earlier theoretical discussions. Section 5 summarizes our research's main contributions, highlighting the advantages of the proposed k-means algorithm modifications. It considers the potential of fuzzy logic and aggregation methods to overcome traditional clustering techniques' limitations. Future research directions are also proposed, indicating potential areas for further improvements in clustering algorithms and data analysis applications.

2. Related work

Research in machine learning has extensively focused on improving clustering algorithms, with special attention to k-means. This area is crucial for advancing data classification and pattern recognition, highlighting the importance of improving these algorithms. A significant part of this enhancement involves the exploration of different distance metrics and aggregation methods to improve the algorithms' precision and efficiency [1, 14].

k-means, known for its straightforward approach and effectiveness in processing large data sets, traditionally uses Euclidean distance to determine how close data points are to cluster centroids. However, this measurement may not always accurately reflect the data's structure, especially in complex, high-dimensional environments. To address this issue, recent research has investigated various distance metrics and their combinations, aiming for a more accurate definition of clusters [6, 28].

In parallel with the development of distance metrics, the use of fuzzy logic in clustering has gained attention. For example, fuzzy logic addresses the uncertainty and imprecision found in many data sets by defining similarity in degrees of membership rather than in binary terms, logic, and provides a more detailed approach to clustering. The application of fuzzy logic in clustering, particularly through innovative methods like aggregating fuzzy equivalences with the k-nearest neighbors algorithm [2, 11, 23], shows promising improvements for k-means clustering as well.

The evaluation of cluster quality is another important aspect of this field. The silhouette coefficient, which assesses the compactness and separation of clusters, is a widely used measure. Further adjustments to this measure to align with the method of defining similarity among objects could provide a pathway toward more accurately evaluating cluster quality, particularly in data sets where traditional metrics may be inadequate [15, 27].

Exploring the composition of aggregations and fuzzy equivalences in k-means clustering is a promising direction for future research. Further investigations of these integrations, particularly in the context of various data types, could lead to significant advancements in clustering methods. Additionally, developing more sophisticated tools for evaluating cluster quality, beyond the adjusted silhouette coefficient, would improve our ability to accurately assess and refine the clustering algorithms [3, 11, 29].

The ongoing development of aggregation methods, along with the incorporation of fuzzy equivalences, marks a key advancement in the optimization of clustering algorithms. This work is a step toward creating clustering solutions that are not only more accurate but also adaptable to the varied challenges presented by different data sets. k-means is a widely used clustering algorithm that partitions a non-clustered input data set into k clusters. First, k initial centroids are identified and used to define clusters. A point x is assigned to a cluster represented by a centroid μ with the least distance between x and μ. To find the optimal centroids, k-means alternates between assigning data points to clusters based on the current centroids and then choosing centroids based on the current assignments. In other words, in an unsupervised learning problem, given a training set of feature vectors $x_1, ..., x_m \in \mathbb{R}^n$, the algorithm aims to predict k centroids (the centers of clusters within an input data set) and assign each input data point to its nearest centroid.

3. Fuzzy equivalences in k-means

The notion of fuzzy equivalences can be understood as fuzzy connectives [9] or fuzzy relations [30]. In our research, we consider fuzzy equivalences proposed by Fodor and Roubens.

Definition 1 ([9]). *A fuzzy equivalence is a function* $E : [0, 1]^2 \to [0, 1]$ *that satisfies the following properties*

- $E(0, 1) = 0$,
- $E(x, x) = 1$ *for all* $x \in [0, 1]$,
- $E(x, y) = E(y, x)$ *for all* $x, y \in [0, 1]$,
- $E(x, y) \leq E(u, v)$ *for all* $x, y, u, v \in [0, 1]$ *such that* $x \leq u \leq v \leq y$.

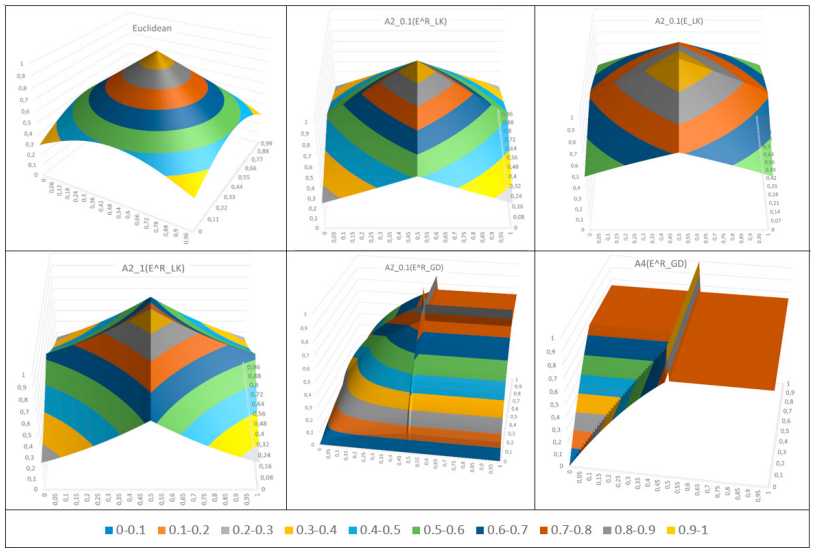

Fig. 1. Compositions $A(E(x - 0.5, y - 0.5))$

Example 1. *Some examples of fuzzy equivalences are generated by the use of fuzzy implications known in the literature under the names Łukasiewicz, Goguen, Gödel and Fodor. For this reason, appropriate fuzzy equivalences are marked with analogous symbols (see [9, 17]).*

- $E_{LK}(x,y) = 1 - |x-y|$,
- $E_{GG}(x,y) = \begin{cases} 1, & \text{if } x = y \\ \frac{\min(x,y)}{\max(x,y)}, & \text{if } x \neq y \end{cases}$,
- $E_{GD}(x,y) = \begin{cases} 1, & \text{if } x = y \\ \min(x,y), & \text{if } x \neq y \end{cases}$,
- $E_{FD}(x,y) = \begin{cases} 1, & \text{if } x = y \\ \max(1-y,x), & \text{if } x < y \\ \max(1-x,y), & \text{if } x > y \end{cases}$.

Example 2. *The next examples come from the above ones by composition with some increasing bijection (see [20]). For the first four functions (named related fuzzy equivalences), an interesting interpretation can be found in [18]. We left the original denoting.*

- $E^R_{LK}(x,y) = (E_{LK}(x,y))^2$,
- $E^R_{GG}(x,y) = E_{GG}(x,y)$,
- $E^R_{GD}(x,y) = E_{GD}(x,y) \cdot (2 - E_{GD}(x,y))$,
- $E^R_{FD}(x,y) = \begin{cases} 1 - 2 \cdot (1 - E_{FD}(x,y))^2, & \text{if } E_{FD}(x,y) \geq 0.5 \\ 2 \cdot (E_{FD}(x,y))^2, & \text{otherwise} \end{cases}$,
- $E_3(x,y) = \frac{2\min(x,y)}{x+y}$ *for* $x + y > 0$, *otherwise* $E_3(x,y) = 1$,
- $E_4(x,y) = \frac{2xy}{x^2+y^2}$ *for* $x + y > 0$, *otherwise* $E_4(x,y) = 1$,
- $E_5(x,y) = \frac{2\min(x^2,y^2)}{x^2+y^2}$ *for* $x + y > 0$, *otherwise* $E_5(x,y) = 1$,
- $E_6(x,y) = \frac{1-|x-y|}{1+|x-y|}$,

where $x, y \in [0,1]$.

Let us notice that fuzzy equivalences, as binary operations, can measure a closeness of two points. In order to use them for objects of multidimensional spaces, we need n-ary operations that will aggregate results obtained for each pair of points. Aggregations serve as a means to combine multiple metrics or equivalences into a single measure, whereas fuzzy equivalences offer a nuanced approach to assess similarity or closeness between data points in a fuzzy set context. Aggregation functions are crucial for merging individual distance metrics or similarity measures into a comprehensive score. Now we present useful information about aggregation functions.

Definition 2 (cf. [4]). *Let* $n \in \mathbb{N}$. *A function* $A : [0,1]^n \to [0,1]$ *which is increasing, i.e., for* $x_i, y_i \in [0,1]$, $x_i \leq y_i$, $i = 1, \ldots, n$, $A(x_1, \ldots, x_n) \leq A(y_1, \ldots, y_n)$, *is called an aggregation function if* $A(0, \ldots, 0) = 0$, $A(1, \ldots, 1) = 1$. *Moreover, we call an aggregation function A a mean if it is idempotent, i.e.,* $A(x, \ldots, x) = x$, $x \in [0,1]$.

Example 3 (cf. [4]). *Some examples of aggregation functions are given by standard means such as lattice operations min, max, projections, geometric mean, and*

- *weighted arithmetic means* $A_w(x_1, \ldots, x_n) = \sum_{k=1}^n w_k x_k$, *for* $w_k > 0$, $\sum_{k=1}^n w_k = 1$,
- *quasi-arithmetic means* $M_\phi(x_1, \ldots, x_n) = \phi^{-1}\left(\frac{1}{n}\sum_{k=1}^n \phi(x_k)\right)$, *where* $\sum_{k=1}^n w_k = 1$, $x_1, \ldots, x_n \in [0,1]$ *and* $\phi : [0,1] \to \mathbb{R}$ *is a continuous, strictly increasing function.*

Example 4. *The following aggregation functions will be used in a sequel*

- *Arithmetic Mean:* $A1(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$,
- *Power Mean:* $A2_p(x_1, \ldots, x_n) = \left(\frac{1}{n} \sum_{i=1}^{n} (x_i)^p\right)^{\frac{1}{p}}$, *where* $p > 0$,
- *Minimum:* $A3(x_1, \ldots, x_n) = \min\{x_1, \ldots, x_n\}$,
- *Maximum:* $A4(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$,

where $A2_1 = A1$.

The capability of using compositions of aggregation functions and fuzzy equivalences as a closeness measure validates their application to some data mining solutions that have previously used distance-based algorithms.

Definition 3. *Let $n \in \mathbb{N}$ and A be an aggregation function. For a fuzzy equivalence E, we consider a binary operation for all $\mathbf{x}, \mathbf{y} \in [0, 1]^n$*

$$A(E)(\mathbf{x}, \mathbf{y}) = A(E(x_1, y_1), \ldots, E(x_n, y_n)). \tag{1}$$

In the sequel, we consider such defined measure of closeness of objects of multidimensional spaces in k-means algorithm, which is a continuation of the research presented in [11] and [18], where the usefulness of such closeness measures by comparing them with metrics in *kNN* algorithm was proved. The distant measure is obtained by the formula $1 - A(E)$.

Figure 1 illustrates the shapes of spheres centered at the point (0.5, 0.5) for the Euclidean metric and compositions of several selected aggregations A and fuzzy equivalences E. Different colors correspond to different closeness levels to the point (0.5, 0.5). For more graphics of compositions of aggregations from Example 4 and fuzzy equivalences from Examples 1 and 2 see [11].

ID	Name	Cnt.	Dim.	Cat.
1	Fertility	100	9	2
2	Iris	150	4	3
3	Zoo	101	16	7
4	Ecoli	336	7	8
5	Glass Identification	214	9	6
6	Leaf	340	14	30
7	LED Display Domain	500	7	10
8	Parkinson Dataset	195	22	2
9	Speaker accent	329	12	6
10	Statlog (German Credit Data)	1000	20	2
11	Banknote Authentication	1372	4	2
12	Wine Quality	6497	11	7

Table 1. The UCI Machine Learning Repository data sets used for experiments. Cnt. - Number of objects, Dim. - dimensionality, Cat. - Number of categories (k).

ID	Name	Dim.	Cat.
1	gen_1000_2_4	2	4
2	gen_1000_4_4	4	4
3	gen_1000_6_4	6	4
4	gen_1000_8_4	8	4
5	gen_1000_2_6	2	6
6	gen_1000_4_6	4	6
7	gen_1000_6_6	6	6
8	gen_1000_8_6	8	6
9	gen_1000_2_8	2	8
10	gen_1000_4_8	4	8
11	gen_1000_6_8	6	8
12	gen_1000_8_8	8	8

Table 2. Artificiall data sets. Dim. - dimensionality, Cat. - Number of categories (k), Cnt. = 1000.

4. Experiments

4.1. Data sets

Real data sets The UCI Machine Learning Repository is a vital resource for researchers and practitioners in the fields of data mining, machine learning, and statistics, offering a wide range of data sets for various tasks [16].

Among these, the Fertility Dataset explores semen characteristics' impact on fertility with 100 instances and 10 attributes [10]. The Iris Dataset, known for its utility in pattern recognition, includes measurements of 150 iris flowers across three species with four features [8]. Other notable data sets include the Zoo Dataset for animal classification, the Ecoli Dataset for protein localization in E.coli cells [24], and the Glass Identification Dataset for classifying types of glass based on their chemical composition [7]. The Leaf Dataset provides data on leaf images from different plant

Input: D the input not clustered data set of m data points x_i
    k the number of clusters to find ($k \leq m$)
    A, E the aggregation and the fuzzy equivalence used as $A(E)$ to compute closeness of points x_i to centroids
Output: The clustered data points from set D into k clusters, with their centers $\mu_1, \mu_2, \ldots, \mu_k$
Initialize cluster centroids $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly
repeat
    for each i (data point $x_i \in \mathbb{R}^n$) **do**
        $c_i \leftarrow \arg\min_j \{1 - A(E)(x_i, \mu_j)\}$
    end for
    for each j (centroid μ_j) **do**
        $\mu_j \leftarrow mean\ of\ \{x_i : c_i = \mu_j\}$
    end for
until convergence or max number of iterations was reached

Fig. 2. The general pseudocode of the modified k-means algorithm used in experiments

species, while the LED Display Domain Dataset focuses on digit recognition from LED displays. The Parkinson Dataset aids in distinguishing individuals with Parkinson's disease through voice measurement analysis [21], and the collection also includes data sets for financial risk modeling with the Statlog (German Credit Data) [13], banknote authentication [22], and wine quality assessment based on physicochemical tests [5]. The summary of the UCM ML data sets is provided in Table 1.

Artificial data sets Data sets were generated using a custom function that leverages the `make_blobs` function from the Python library. Each data set is characterized by a set of parameters: the number of samples (n), the number of centers or clusters (k), the dimensionality of the data (d), and a random state seed (r) to ensure reproducibility. The `make_blobs` function creates n samples distributed across k clusters, with each cluster being roughly spherical and having a dimensionality of d. The clusters are generated within a specified bounding box, here chosen to be between -6 and 6 along each dimension, allowing for a wide dispersion of data points. The generated data sets are saved as CSV files, where each file is named according to the pattern `gen_{n}_{d}_{k}.csv`, making it easy to identify the characteristics of the data set from the filename. This systematic approach facilitates the creation of a diverse array of data sets to test and benchmark machine learning algorithms, particularly those designed for clustering analysis. The summary of the artificial data sets is given in Table 2.

4.2. Experimental setup

In each experiment, the optimal set of clusters was sought using the k-means algorithm. The values of the parameter k of the algorithm were equal to the number of values of the decision attribute of the input data sets. The results of experiments obtained using the k-means algorithm implemented in Python were used as a reference point. The quality of the resulting division of objects into clusters was expressed using the Silhouette coefficient. The experiments were then repeated, but instead of the Euclidean metric a distinct combination of an aggregation method and fuzzy equivalence was used to calculate the closeness (cl) of input objects to centroids (Figure 2). To evaluate the performance of clustering using silhouette coefficients, the closeness measure was replaced with a distance measure ($dist$) (greater closeness = less distance) calculated according to the formula dist = $1 - cl$. Their values were collected in Table 3. The same distance measures used in the modified k-means algorithm to determine the distances of points from the centroids were applied to calculate the values of the Silhouette coefficient. The experiments were conducted on real and artificial data sets named with the convention `gen_1000_X_Y`, where `X` indicates the dimensionality of the data set and `Y` the number of categories or clusters. Different combinations of aggregation methods ($A2_{0.1}$, $A2_1$, $A2_4$, $A3$, $A4$) and fuzzy equivalences (e.g., E_{LK}, E^R_{LK}, etc.) were used to compute the silhouette coefficients. The pseudocode of the modified *k-means* algorithm is presented as Figure 2 while the source code is available on the Github repository [19].

The experiments to evaluate the performance of the k-means clustering algorithm with various distance measures computed as $1 - A(E)$ were structured and executed methodically, leveraging Python and its powerful libraries such

#	Agg (A_i)	Dist. (E)	Real datasets												Artificial datasets											
			Fer.	Iris	Zoo	Eco.	Gla	Lea	LED	Par.	Spe.	Cre.	Ban.	Win.	1	2	3	4	5	6	7	8	9	10	11	12
1	Euclidean		0.13	0.51	0.28	0.16	0.35	0.26	0.32	0.24	0.14	0.08	0.33	0.13	0.64	0.43	0.65	0.73	0.36	0.43	0.52	0.25	0.43	0.32	0.35	0.43
2	2_0.1	E_LK	0.24	0.54	0.24	0.07	0.34	0.25	0.31	0.33	0.15	0.12	0.36	0.15	0.65	0.45	0.66	0.45	0.42	0.49	0.51	0.50	0.41	0.40	0.41	0.51
3	2_0.1	E_LK^R	0.19	0.50	0.24	0.13	0.32	0.24	0.30	0.27	0.14	0.09	0.33	0.15	0.43	0.66	0.52	0.72	0.36	0.33	0.29	0.24	0.36	0.52	0.39	0.38
4	2_0.1	E_GG	0.10	0.31	0.23	0.07	0.43	0.14	0.32	0.24	0.02	0.06	0.28	0.13	0.43	0.32	0.32	0.57	0.34	0.41	0.20	0.50	0.39	0.26	0.33	0.46
5	2_0.1	E_GD	0.06	0.10	0.26	0.00	0.14	-0.02	0.30	0.00	-0.01	0.05	0.06	0.01	0.01	0.17	0.11	0.04	0.01	-0.02	0.00	0.09	0.06	-0.05	0.02	0.02
6	2_0.1	E_GD^R	0.11	0.45	0.23	0.08	0.50	0.20	0.31	0.35	0.08	0.07	0.40	0.12	0.38	0.71	0.70	0.73	0.53	0.43	0.57	0.66	0.51	0.53	0.55	0.52
7	2_0.1	E_FD	0.22	0.21	0.24	0.04	0.10	0.04	0.32	0.22	0.06	0.10	0.17	0.01	0.31	0.35	0.32	0.22	0.04	0.16	0.20	0.21	0.03	0.18	0.21	0.18
8	2_0.1	E_FD^R	0.22	0.31	0.24	0.10	0.21	0.08	0.30	0.26	0.09	0.10	0.26	0.07	0.42	0.53	0.49	0.34	0.27	0.33	0.30	0.19	0.08	0.26	0.36	0.38
9	2_0.1	E3	0.01	0.35	0.00	0.00	0.00	0.16	0.03	0.29	0.00	0.00	0.31	0.10	0.59	0.48	0.59	0.40	0.42	0.50	0.45	0.38	0.32	0.36	0.33	0.35
10	2_0.1	E4	0.02	0.46	0.00	0.00	0.01	0.21	0.03	0.37	0.08	0.00	0.43	0.15	0.41	0.73	0.73	0.75	0.54	0.62	0.36	0.68	0.53	0.57	0.51	0.58
11	2_0.1	E5	0.01	0.32	0.00	0.00	0.00	0.17	0.03	0.23	0.03	0.00	0.28	0.11	0.37	0.57	0.56	0.59	0.41	0.36	0.24	0.39	0.39	0.44	0.35	0.32
12	2_0.1	E6	0.11	0.49	0.24	0.11	0.30	0.23	0.32	0.27	0.03	0.10	0.32	0.11	0.60	0.64	0.51	0.69	0.41	0.41	0.34	0.21	0.41	0.42	0.26	0.37
13	2_1	E_LK	0.17	0.52	0.37	0.15	0.32	0.24	0.34	0.33	0.14	0.08	0.36	0.12	0.64	0.68	0.55	0.43	0.43	0.43	0.34	0.39	0.43	0.46	0.41	0.54
14	2_1	E_LK^R	0.15	0.49	0.37	0.16	0.28	0.22	0.37	0.32	0.03	0.06	0.32	0.09	0.60	0.39	0.50	0.68	0.42	0.56	0.33	0.44	0.37	0.50	0.48	0.44
15	2_1	E_GG	0.16	0.37	0.36	0.14	0.27	0.15	0.37	0.25	0.05	0.07	0.23	0.11	0.29	0.40	0.52	0.57	0.38	0.25	0.29	0.30	0.33	0.30	0.24	0.02
16	2_1	E_GD	0.16	0.14	0.36	-0.01	0.08	-0.03	0.32	0.01	-0.01	0.05	0.06	0.00	0.04	-0.03	0.11	0.11	0.05	0.04	0.01	0.11	-0.02	0.00	-0.02	0.04
17	2_1	E_GD^R	0.20	0.51	0.36	0.20	0.45	0.24	0.36	0.38	0.05	0.09	0.37	0.16	0.71	0.59	0.46	0.34	0.58	0.54	0.58	0.49	0.50	0.43	0.58	0.39
18	2_1	E_FD	0.16	0.23	0.37	0.05	0.13	0.05	0.37	0.18	0.05	0.07	0.16	0.01	0.25	0.33	0.28	0.18	0.14	0.20	0.12	0.21	0.04	0.08	0.10	0.19
19	2_1	E_FD^R	0.16	0.31	0.37	0.10	0.14	0.05	0.38	0.33	0.07	0.08	0.25	0.04	0.19	0.51	0.47	0.26	0.24	0.33	0.33	0.36	-0.03	0.19	0.23	0.31
20	2_1	E3	0.06	0.40	0.02	0.05	0.03	0.17	0.05	0.32	0.04	0.03	0.29	0.10	0.59	0.38	0.39	0.63	0.36	0.45	0.36	0.36	0.35	0.36	0.37	0.38
21	2_1	E4	0.06	0.54	0.03	0.04	0.06	0.25	-0.02	0.40	0.05	0.05	0.40	0.21	0.73	0.57	0.73	0.50	0.45	0.46	0.60	0.33	0.50	0.57	0.60	0.40
22	2_1	E5	0.08	0.37	0.02	0.07	0.06	0.14	-0.12	0.24	0.04	0.03	0.23	0.08	0.56	0.57	0.53	0.59	0.40	0.39	0.24	0.26	0.39	0.30	0.18	0.33
23	2_1	E6	0.15	0.48	0.37	0.16	0.28	0.21	0.42	0.30	0.03	0.06	0.31	0.15	0.58	0.39	0.49	0.66	0.42	0.37	0.43	0.52	0.36	0.40	0.47	0.36
24	2_4	E_LK	0.16	0.51	0.38	0.15	0.23	0.21	0.39	0.35	0.04	0.07	0.29	0.11	0.42	0.64	0.59	0.40	0.43	0.38	0.18	0.25	0.41	0.33	0.46	0.33
25	2_4	E_LK^R	0.18	0.47	0.37	0.12	0.17	0.15	0.36	0.34	0.03	0.08	0.25	0.10	0.39	0.31	0.51	0.57	0.33	0.43	0.29	0.32	0.38	0.23	0.25	0.38
26	2_4	E_GG	0.16	0.34	0.37	0.11	0.18	0.07	0.33	0.20	0.02	0.06	0.17	0.07	0.46	0.28	0.45	0.51	0.29	0.37	0.34	0.29	0.26	0.18	0.19	0.26
27	2_4	E_GD	0.18	0.24	0.37	0.01	0.06	-0.03	0.34	0.09	0.03	0.08	0.13	0.02	0.04	0.09	0.10	0.16	0.12	-0.08	0.10	0.08	-0.06	-0.07	0.03	-0.02
28	2_4	E_GD^R	0.19	0.51	0.37	0.24	0.24	0.16	0.33	0.34	0.05	0.08	0.33	0.13	0.54	0.52	0.66	0.72	0.52	0.50	0.39	0.44	0.38	0.41	0.47	0.53
29	2_4	E_FD	0.17	0.26	0.37	0.02	0.06	0.02	0.32	0.47	0.05	0.08	0.22	0.01	0.26	0.26	0.23	0.20	0.03	0.16	0.13	0.16	0.07	0.10	0.07	0.11
30	2_4	E_FD^R	0.17	0.34	0.37	0.05	0.14	0.03	0.36	0.42	0.08	0.08	0.32	0.01	0.42	0.41	0.29	0.45	0.25	0.27	0.28	0.23	0.14	0.16	0.11	0.15
31	2_4	E3	0.08	0.41	0.03	0.11	0.05	0.11	-0.05	0.23	0.03	0.05	0.24	0.10	0.54	0.56	0.43	0.59	0.41	0.27	0.30	0.30	0.19	0.25	0.40	0.29
32	2_4	E4	0.14	0.53	0.03	0.09	0.04	0.20	0.01	0.37	0.06	0.05	0.35	0.14	0.70	0.74	0.70	0.75	0.58	0.61	0.49	0.47	0.49	0.46	0.51	0.44
33	2_4	E5	0.09	0.35	0.03	0.10	0.06	0.07	-0.06	0.07	0.02	0.07	0.18	0.09	0.19	0.47	0.34	0.11	0.28	0.24	0.27	0.28	0.20	0.23	0.16	0.35
34	2_4	E6	0.18	0.47	0.37	0.12	0.14	0.15	0.33	0.34	0.03	0.08	0.24	0.10	0.39	0.54	0.42	0.30	0.30	0.34	0.13	0.32	0.29	0.28	0.32	0.33
35	3	E_LK	0.10	0.46	0.19	0.14	0.31	0.20	0.22	0.16	0.07	-0.01	0.27	0.12	0.64	0.33	0.49	0.71	0.39	0.58	0.48	0.48	0.35	0.30	0.41	0.46
36	3	E_LK^R	0.05	0.40	0.18	0.14	0.30	0.19	0.21	0.12	0.07	0.00	0.23	0.10	0.36	0.59	0.58	0.62	0.34	0.43	0.45	0.32	0.43	0.41	0.36	0.28
37	3	E_GG	0.04	0.23	0.18	0.13	0.39	0.10	0.25	0.07	0.03	0.00	0.26	0.11	0.35	0.43	0.40	0.38	0.25	0.24	0.26	0.26	0.37	0.27	0.26	0.21
38	3	E_GD	0.02	0.01	0.17	0.00	0.06	0.00	0.21	0.00	-0.03	0.00	0.12	-0.01	0.07	0.01	0.00	-0.04	-0.05	0.00	0.01	-0.03	-0.02	-0.02	0.02	-0.02
39	3	E_GD^R	0.07	0.25	0.19	0.20	0.52	0.15	0.24	0.13	0.06	0.00	0.40	0.13	0.44	0.56	0.45	0.32	0.57	0.17	0.44	0.40	0.51	0.45	0.43	0.37
40	3	E_FD	0.09	0.17	0.17	0.05	0.10	0.04	0.22	0.08	0.02	0.06	0.13	0.04	0.35	0.17	0.25	0.22	0.18	0.14	0.17	0.16	0.12	0.18	0.17	0.19
41	3	E_FD^R	0.08	0.25	0.18	0.09	0.17	0.07	0.21	0.09	0.03	0.04	0.19	0.10	0.46	0.48	0.44	0.25	0.28	0.28	0.22	0.22	0.13	0.30	0.28	0.19
42	3	E3	0.00	0.28	0.00	0.00	0.02	0.14	0.07	0.11	0.03	0.00	0.33	0.16	0.56	0.50	0.49	0.39	0.40	0.41	0.31	0.36	0.32	0.19	0.34	0.32
43	3	E4	0.00	0.26	0.00	0.00	0.02	0.16	0.07	0.14	0.06	0.00	0.44	0.12	0.65	0.57	0.28	0.60	0.58	0.53	0.30	0.54	0.49	0.42	0.38	0.37
44	3	E5	0.00	0.23	0.00	0.00	0.01	0.10	0.07	0.05	0.04	0.00	0.24	0.05	0.53	0.38	0.39	0.22	0.27	0.24	0.27	0.23	0.38	0.28	0.23	0.18
45	3	E6	0.07	0.39	0.18	0.13	0.29	0.18	0.21	0.12	0.06	0.00	0.22	0.12	0.33	0.36	0.56	0.60	0.40	0.36	0.43	0.41	0.34	0.23	0.36	0.45
46	4	E_LK	-0.32	0.38	-	0.90	0.46	0.29	-	0.39	-0.11	-	0.05	-0.03	0.19	0.13	0.16	0.05	0.07	0.01	0.04	-0.04	0.17	-0.03	-0.06	-0.10
47	4	E_LK^R	-0.32	0.38	-	0.90	0.47	0.31	-	0.38	-0.10	-	0.05	-0.04	0.08	0.15	0.16	0.01	0.05	0.00	-0.02	-0.06	0.03	-0.01	-0.08	-0.08
48	4	E_GG	-0.32	0.31	-	-	0.47	0.14	-	0.25	-0.10	-	0.15	-0.01	0.36	0.26	0.14	0.08	0.12	0.04	0.01	-0.04	0.04	-0.05	0.00	-0.05
49	4	E_GD	-0.32	0.34	-	0.90	0.42	-0.09	-	0.26	0.03	-	-0.12	0.03	0.01	-0.02	0.12	-0.06	0.21	0.06	0.11	0.07	0.01	-0.10	-0.07	0.10
50	4	E_GD^R	-0.32	0.30	-	0.90	0.28	-0.01	-	0.15	-0.24	-	0.14	-0.07	0.41	0.44	-0.12	0.18	0.20	0.01	-0.07	-0.03	-0.08	-0.07	-0.12	-0.17
51	4	E_FD	-0.32	0.22	-	-	0.53	0.10	-	0.51	0.05	-	0.14	-0.04	0.33	0.21	0.13	0.18	0.05	-0.07	0.17	0.15	-0.08	-0.05	-0.03	0.15
52	4	E_FD^R	-0.32	0.34	-	0.90	0.15	-0.13	-	0.73	0.06	-	0.23	-0.16	0.07	0.05	0.26	0.29	-0.21	-0.05	0.20	0.19	-0.04	0.14	0.12	0.26
53	4	E3	-0.59	0.31	-	-0.06	-0.06	0.18	-0.46	0.25	-0.10	-0.35	0.15	-0.02	0.44	0.07	0.24	-0.10	0.16	0.03	-0.04	-0.05	0.09	-0.05	-0.04	-0.07
54	4	E4	-0.59	0.29	-	-0.22	-0.21	-0.01	-0.46	0.15	-0.26	-0.36	0.13	-0.07	0.11	0.47	0.24	-0.09	0.11	-0.03	0.00	-0.10	0.12	-0.17	-0.14	-0.18
55	4	E5	-0.58	0.31	-	-0.06	-0.05	0.18	-0.46	0.25	-0.10	-0.35	0.15	-0.01	0.14	0.07	0.10	0.01	0.22	0.04	-0.05	-0.07	0.10	-0.11	-0.05	-0.09
56	4	E6	-0.32	0.38	-	0.90	0.47	0.37	-	0.38	-0.10	-	0.06	-0.05	0.24	0.13	0.09	0.00	0.14	0.07	-0.02	-0.02	0.07	-0.02	-0.03	-0.08

Table 3. The silhouette coefficient values calculated for different data sets, aggregations, fuzzy equivalences. Outputs for Euclidean distance based *k*-means algorithm are showed as a reference at the top.The columns contain the IDs of experiments (column 1), names of aggregations (column 2), fuzzy equivalences (column 3) and data set names (columns 3 - 27). The color gradient represents the silhouette values, ranging from the lowest (red) to the highest (green).

as NumPy [12], pandas [25], and scikit-learn [26]. Essential Python libraries were imported for data manipulation (NumPy, pandas), generating synthetic data sets (scikit-learn), calculating distances (scipy), and measuring clustering quality (scikit-learn's silhouette scores). The `max_iters` variable was defined to limit the number of iterations for the *k*-means algorithm. For each data set, the distance matrix was calculated using the selected distance metric, modified by the current aggregation function. The partial function from the functools library enabled the dynamic modification of the distance function with the aggregation parameter. The modified distance matrix served as input to *k*-means, and the clustering algorithm was executed with the specified number of clusters and iterations. After clustering, the silhouette score for the resulting labels was calculated against the distance matrix, providing a measure of cluster quality. Special cases where clustering resulted in a single cluster or failed were handled with predefined silhouette scores (-1

and -2, respectively). The silhouette scores for each combination of distance metric and aggregation function were compiled into a results dictionary, keyed by data set name, with an array of scores corresponding to each experimental run.

4.3. Results and Observations

The results of the experiments are shown in Table 3, and this subsection is dedicated to discussing these results and observations.

For each of the UCI ML repository data set, we selected 3 best results (except for the Ecoli and Parkinson data sets, where the best result was significantly better from the second and third one) and examined the relevant aggregations and fuzzy equivalences. The most valuable aggregations, fuzzy equivalences, compositions of aggregations and fuzzy equivalences are:

- $A2_1$ - 19 times; $A2_{0.1}$ - 18 times; $A2_4$ - 11 times; $A4$ - 10 times; $A3$ - 3 times
- E_{LK} - 12 times; E^R_{LK} - 7 times; E_{FD} - 6 times; E^R_{FD} - 6 times; E_GD - 3 times; E^R_{GD} - 7 times; E_{GG} - 1 time; E_3 - 1 time; E_4 - 7 times; E_6 - 7 times.
- $A2_{0.1}(E_{LK})$ - 5 times and $A2_1(E^R_{GD})$ - 4 times

This summary shows that the Euclidean distance and variants of E_{LK} frequently resulted in higher silhouette coefficients, indicating that these measures are more effective for clustering in the tested configurations. Our evaluation of clustering performance across various data sets revealed that the distance measure E_{LK} consistently yielded superior average silhouette scores, particularly when combined with the aggregation method $A2_{0.1}$. This combination's inherent robustness and adaptability suggest a strong compatibility with the data structures present in the evaluated data sets.

Modifications in the aggregation methods led to notable variations in silhouette scores, highlighting the profound influence of the method used to aggregate distances or determine cluster centroids on the overall effectiveness of the clustering. For instance, transitioning from $A2_{0.1}$ to $A2_1$ or $A2_4$ not only altered the silhouette scores quantitatively but also qualitatively reflected changes in clustering performance.

Yet another observation is that slight changes of the way the distance $1 - A(E)$ is computed may lead to significant differences of output clusters (cf. adjacent records of Table 3 for E_{GD} and E^R_{GD} and artifficial datasets) since fuzzy eguivalences E and E^R return similar values for the same inputs.

Data sets characterized by increased dimensionality and a higher number of categories typically exhibited a wider range of silhouette scores. This demonstrates the growing complexity in clustering high-dimensional data or data sets with numerous categories, thus emphasizing the need for more sophisticated or adaptable clustering approaches.

Further analysis revealed that the *Fertility* and *Iris* data sets achieved the highest silhouette scores using the aggregation method $A2_{0.1}$ and the distance measure E_{LK}. In contrast, the *Zoo* data set exhibited optimal clustering performance with the aggregation method $A2_4$ and the same distance measure. This highlights that specific data sets may favor certain methodological pairings for optimal clustering.

A significant observation was the identification of negative silhouette values, indicative of suboptimal clustering. This phenomenon was particularly evident with the aggregation method $A4$, where negative values were prevalent across the majority of distance measures, indicating its unsuitability for the data sets examined. These findings are crucial for guiding future applications and theoretical developments in data clustering, offering insight into the intricate dynamics between distance measures, aggregation methods, and data set characteristics.

Observations conducted on artificially generated data demonstrated that while the clusters were approximately spherical (hence, reference results should be drawn from line 1 of Fig. 4), certain alternative combinations of aggregation and equivalence yielded comparable or superior outcomes.

5. Conclusions

The analysis of silhouette coefficients across various clustering configurations reveals significant insights into the performance of different distance measures and aggregation methods. It highlights the importance of carefully se-

lecting the clustering algorithm's parameters based on the data set's specific characteristics. Some distance measures exhibit robust performance across multiple data sets, indicating their general suitability for the type of clustering problems encountered.

However, the aggregation method plays a critical role, with different methods leading to vastly different outcomes. As the complexity of the data sets increases, particularly with regard to the dimensionality and number of categories, the selection of appropriate clustering strategies becomes even more critical. Future work should focus on exploring adaptive clustering methods that can automatically adjust to the complexities inherent in high-dimensional and multi-category data sets.

The compositions $A2_{0.1}(E_{LK})$ and $A2_1(E^R_{GD})$ were most often among the top 3 clustering results. Therefore, these compositions, along with the Euclidean distance, are worth using for data clustering. However, neither of these two compositions was among the top 3 results for 6, i.e., half of tested data sets among those from UCI ML. Therefore, it may be worthwhile to explore the utilization of alternative aggregations and fuzzy equivalences in clustering tasks, particularly for data sets where the quality of clustering achieved with the Euclidean metric or the $A2_1(E_{LK})$ and $A2_1(E^R_{GD})$ measures is deemed unsatisfactory.

References

[1] Arthur, David and Vassilvitskii, Sergei. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[2] Bezdek, James C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[3] Bezdek, James C and Keller, James and Krishnapuram, Raghu and Pal, Nikhil R. *Fuzzy models and algorithms for pattern recognition and image processing*. Kluwer Academic Publishers, 1984.

[4] Calvo, Tomasa and Kolesárová, Anna and Komorniková, Magda and Mesiar, Radko. Aggregation operators: Properties, classes and construction methods. pages 3–104, 2002.

[5] Cortez, Paulo and Cerdeira, António and Almeida, Fernando and Matos, Telmo and Reis, José. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.

[6] Dhillon, Inderjit S and Guan, Yuqiang and Kulis, Brian. Weighted graph cuts without eigenvectors: A multilevel approach. *IEEE transactions on pattern analysis and machine intelligence*, 29(11):1944–1957, 2007.

[7] Evett, Ian W and Spiehler, Ernst J. Rule induction in forensic science. Technical report, Kansas State Univ Manhattan Dept of Statistics, 1987.

[8] Fisher, Ronald A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[9] Fodor, János C and Roubens, Marc. *Fuzzy Preference Modelling and Multicriteria Decision Support*. Kluwer Academic Publishers, 1994.

[10] Gil, David and Girela, Jose and De Juan, Joaquín and Gomez-Torres, María José and Johnsson, Magnus. Fertility Diagnosis by Using Semen Characteristics. *Journal of Systems Science and Complexity*, 25(1):2–3, 2012.

[11] Grochowalski, Piotr and Król, Anna and Rząsa, Wojciech. Radius kNN Classifier Using Aggregation of Fuzzy Equivalences. In *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, page 1–6. IEEE Press, 2021.

[12] Harris, Charles R. et al. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[13] Hofmann,Hans. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

[14] Jain, Anil K. Data clustering: 50 years beyond K-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[15] Kaufman, Leonard and Rousseeuw, Peter J. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[16] Kelly, Markelle and Longjohn, Rachel and Nottingham, Kolby. The UCI machine learning repository. *URL https://archive. ics. uci. edu*, 2023.

[17] Król, Anna. Fuzzy (C,I)-equivalences. In *Proceedings of the 8th International Summer School on Aggregation Operators*, pages 157–161, 2015.

[18] Król, Anna and Rząsa, Wojciech and Grochowalski, Piotr. Aggregation of Fuzzy Equivalences in Data Exploration by kNN Classifier. In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, page 1–6. IEEE Press, 2020.

[19] Piotr Lasek. Clustering using Fuzzy Equivalences. https://github.com/piotrlasek/fuzzy-equivalences.

[20] Li, Yingfang and Qin, Keyun and He, Xingxing. Some new approaches to constructing similarity measures. *Fuzzy Sets Syst.*, 234:46–60, 2014.

[21] Little, Max A and McSharry, Patrick E and Roberts, Stephen J and Costello, Declan AE and Moroz, Irene M. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering OnLine*, 6(1):23, 2007.

[22] Lohr, Wolfgang. Statistical Mechanics of Money, Income, and Wealth: A Short Survey. *Econophysics of Income and Wealth Distributions*, pages 51–60, 2011.

[23] Miyamoto, Sadaaki and Ichihashi, Hidetomo and Honda, Katsuhiro. *Algorithms for Fuzzy Clustering: Methods in c-Means Clustering with Applications*. Springer, 2008.

[24] Nakai, Kenta and Kanehisa, Minoru. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, 14(4):897–911, 1992.

[25] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

[26] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. volume 12, pages 2825–2830, 2011.

[27] Rousseeuw, Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
[28] Xu, Dongkuan and Tian, Yingjie. Comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015.
[29] Zadeh, Lotfi A. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.
[30] Zadeh, Lotfi A. Similarity relations and fuzzy orderings. *Information Sciences*, 3:177–200, 1971.