UDC 004

A COMPARATIVE EVALUATION OF SIMILARITY MEASURES FOR SEMI-SUPERVISED DENSITY PEAKS CLUSTERING

Ahmed Saad Hussein [1,2]

[1] Department of Cybersecurity Engineering Technologies, Technical Engineering College, Al-Farabi University, Baghdad 10001, Iraq
[2] Department of Mobile Communications and Computing Engineering, College of Engineering, University of Information Technology and Communications, Baghdad, Iraq

Ahmed.Hussein@alfarabiuc.edu.iq

Abstract. Semi-supervised Density Peak (SDenPeak) algorithm is known to be efficient and simple in tasks clustering. It improves clustering performance by adding pair-wise constraints, must-link and cannot-link constraints, that drive the grouping process by imposing similarity and dissimilarity between data points. One of the key considerations in clustering accuracy is the selection of similarity measure because various measures reflect diverse structural attributes to data. The problem with the fact that there is no universal best measure of similarity is that it is a tricky task to choose a suitable measure that is dependent on the nature of the data. To explore the effects of the six similarity measures on SDenPeak algorithm performance, the six measures (Euclidean Distance, Cosine Similarity, City Block (Manhattan) Distance, Minkowski Distance, Earth Mover's Distance (EMD), and Rapid Computation of the Maximal Information Coefficient (RapidMIC) Distance) are evaluated systematically in this study in order to understand their influences Real-world datasets are extensively experimented to evaluate the accuracy of clustering and structural consistency in each of the measures. These findings present comparative information on the effectiveness of the various similarity measures and illustrate their applicability to various data distributions providing a useful guide to achieving the best clustering performance in semi-supervised models.

Keywords: clustering, semi-supervised, similarity measurement, density peaks, real-world datasets.

1 INTRODUCTION

Clustering is a simple data mining problem that aims to partition datasets into multiple subsets so that each subset contains objects closer to each other than those in other subsets [1, 33, 39, and 40]. A clustering solution is a collection of such clusters, typically containing all the objects from the dataset [33]. Moreover, clustering methods can also define relationships between clusters, e.g., building a hierarchy where clusters are nested inside one another. However, the quality of clustering results is often influenced by various features of datasets, e.g., noise, overlapping data, shape variation, density, and size variation [39]. Moreover, the performance of clustering algorithms may be impacted by parameter choice and methodologies employed. To address these problems, semi-supervised clustering has been a rising field of research in recent times since it fuses both supervised and unsupervised learning paradigms [3]. Unlike traditional supervised learning, where each point is labeled in advance, semi-supervised clustering applies side information, e.g., pairwise constraints, to direct the clustering procedure. These side constraints are classified into must-link and cannot-link constraints [7]. A must-link constraint asserts that two objects are within the same group, whereas a cannot-link constraint asserts that two are within different groups. By incorporating these constraints, semi-supervised clustering attempts to enhance clustering accuracy, particularly in datasets with high complexity where fully labeled data are not accessible [1, 2, 3, 7, and 40].

Most of the unsupervised clustering algorithms have significantly improved performance by integrating prior knowledge into the clustering process [40]. These kinds of algorithms can work efficiently in a semi-supervised setting. Some widely used clustering algorithms have been proposed to improve their accuracy and robustness. These algorithms are based on must-link and cannot-link constraints [43]. For example, the K-means algorithm has been adapted to include the constraints to define more meaningful cluster topologies and minimize the amount of misclassification experienced. In addition, several other

clustering algorithms have been generalized for semi-supervised learning, such as COP-Kmeans for generalizing K-means with pairwise constraints, constrained spectral clustering for integrating the constraints into the spectral clustering frameworks, graph-based clustering for using graph topology to improve the clustering performance and mean shift clustering for being more adaptive in adjusting to data distribution with the incorporation of the constraints [44]. These advancements demonstrate that adding background knowledge into conventional clustering approaches can greatly improve the quality of the clustering, enhance interpretability, and decrease the common problems that are associated with complete unsupervised clustering methods [4, 5, and 6].

The present paper will identify the most appropriate similarity measurement methodologies for the SDenPeak algorithm. In particular, we initially chose the Semi-supervised Density Peak algorithm recently published in Science [9]. We then discussed six prevalent similarity measurement methodologies. Ultimately, we have conducted exhaustive experiments on eleven real-world datasets to demonstrate the efficacy of various performance metrics by applying various similarity measurement methodologies.

The remaining parts of this paper are structured as follows: Section 2 briefly discusses the SDenPeak algorithm, Section 3 discusses the six similarity techniques, Section 4 reports experimental design demonstrating this algorithm's optimal performance, and Section 5 concludes the paper.

2 RELATED WORK

While these studies have provided valuable insight into semi-supervised clustering and distance measures, there is a lack of systematic comparison of similarity measures in semi-supervised density peaks clustering. Most research has been geared towards unsupervised clustering algorithms or supervised classification methods, with the intersection of semi-supervised density peaks clustering with various similarity measures remaining unexplored. This research seeks to bridge this gap by presenting an extensive performance evaluation of the SDenPeak algorithm on six varied similarity measures: Euclidean, Cosine, Manhattan, Minkowski, Earth Mover's Distance (EMD), and RapidMIC. According to experiments on real-world data, this paper will identify the best similarity measure for semi-supervised density peak clustering, adding to the general area of semi-supervised machine learning. Current advancements in semi-supervised clustering have seen more effective algorithms that use previous knowledge to improve clustering accuracy. The Semi-Supervised Density Peaks Clustering (SDenPeak) algorithm is one of the advancements of the original Density Peaks Clustering (DPC) algorithm, first proposed by Rodriguez and Laio [23]. While DPC correctly identifies cluster centers concerning density and distance, its completely unsupervised nature is inefficient in datasets with overlapping clusters, non-uniform densities, or noise. To overcome these limitations, semi-supervised approaches introduce pairwise constraints in the form of must-link and cannot-link constraints, which limit the clustering algorithm to establishing relationships between some points [24-27]. This integration of supervision is beneficial to improving clustering performance in various applications, including image processing, bioinformatics, and text mining [28]. An essential part of achieving semi-supervised density peak clustering is choosing an appropriate measure of similarity because various distance measures affect the formation of clusters and how well the algorithm can differentiate data points. Multiple studies have analyzed the impact of different distance measures on clustering performance. Aggarwal et al. [29] compared the performance of traditional distance measures, such as Euclidean, Cosine, Manhattan, and Minkowski distances, on high-dimensional clustering tasks. The findings revealed that no distance measure is superior in all situations, as the performance depends on data distribution and feature space properties. Ding et al. [27] subsequently investigated the application of distance measures to semi-supervised clustering. They noted that employing domain-specific similarity measures significantly improved the clustering accuracy of accurate data. One of the most helpful distance metrics for clustering purposes is the Earth Mover's Distance (EMD), which has been widely applied for measuring similarity among probability distributions [31]. EMD has proven particularly useful for image clustering and document retrieval because it can model minimum transformation costs among distributions. In semi-supervised density peaks clustering, Liu et al. [32] demonstrated that EMD improves clustering quality for datasets with complex structures, e.g., those with multi-modal distributions. EMD, however, is computationally expensive and less suitable for large-scale datasets. The RapidMIC (Rapid Computation of the Maximal Information Coefficient) distance metric has been used recently to identify non-linear point relationships [33-35]. RapidMIC is well-suited for clustering high-dimensional data with complex structures since it identifies linear and non-linear correlations, unlike most geometric distance-based metrics. In heterogeneous data-type tasks, Tang et al. [36] proposed a new framework that utilizes AHC algorithms for primary partition generation, taking advantage of the heterogeneity AHC provides to improve the quality of clusters. Our method incorporates a new similarity measure with an adaptive weighting scheme and stringent selection process to minimize computation complexity. The mechanism attempts to capture the stability of the base partition generation process with semi-supervised clustering

techniques. Based on constraint knowledge, the proposed method has three main steps: it first creates primary clusters by using various AHC methods; it then builds a new similarity measure to measure the similarity of objects; finally, its re-clusters the primary clusters to create final clusters. Experiments were conducted on real datasets to test the performance of the proposed method. The results indicate appreciable enhancements in clustering precision over comparable approaches. González-Almagro et al. [37] tackled clustering with pairwise and monotonicity requirements resulting from background knowledge. First, the formal framework for clustering under monotonicity restrictions is defined, resulting in a distance measure. After constructing an objective function that fuses the suggested distance measure and a pairwise constraint-based penalty term, pairwise constraints are integrated. This objective function can be optimized using EM. Liu et al. [38] proposed a novel method based on GANs and semi-supervised learning. The structural quality of the network is first enhanced by rewiring it using vertex similarity measures. Then, the new model of generative adversarial networks is built; our model supplies partitions by reconstructing the network and can serve as the base for defining the key communities. The node selection algorithm also uses the local clustering coefficient as the reward signal. Isolated nodes are reallocated, and the final community structure is calculated. The new approach outperforms the previous methods in F1 and Jaccard indices, as experimental results on four large-scale real-world datasets show.

Although the use of semi-supervised algorithms has been promising such as SDenPeak, there has been a lack of research that examines the effect of varying similarity measures on the performance of these algorithms in a systematic manner. A significant part of the existing work is related to the algorithm design or similarity analysis in a fully-supervised or unsupervised environment. Similarity metrics in semi-supervised, constraint-based clustering The role of similarity measures in semi-supervised, constraint-based clustering is not as extensively studied, especially in comparative studies on various datasets. This restricts the pragmatic guidance on the choice of suitable actions to take in a particular type of data. Our study addresses this gap by comparing six similarity metrics of SDenPeak and providing new knowledge, which is not possible in older research.

3 SEMI-SUPERVISED DENSITY PEAKS

The “Fast Search and Find of Density Peaks” (Density Peaks Clustering) algorithm is perhaps the most potent revolution in unsupervised clustering because it adeptly uses density and distance to find cluster centers. The idea behind this algorithm is that cluster centers are more concentrated as opposed to their surrounding data points and are enclosed with less dense objects. They will also be prone to be far between each other thereby making sure that each cluster is mostly segregated. Such an intuitive method has seen the Density Peaks algorithm being a highly desired algorithm in many clustering problems [8]. In 2016, however, a more powerful form of the Density Peaks algorithm was proposed, which instead of being used in an unsupervised manner was now a semi-supervised one, resulting in the Semi-supervised Density Peaks (SDenPeak) algorithm. The new version is much better in performance than the original unsupervised version in terms of clustering since it adds pairwise constraints as an extra information when clustering. These must-link and cannot-link constraints, along with pairwise constraints allow SDenPeak to add external knowledge to the clustering process and produce more accurate, stable and useful clusters. The must-link constraint is used to guarantee that two points should be in the same cluster and the cannot-link constraint is used to guarantee that two points should be given in different clusters [9]. Cleverly exploiting these constraints, SDenPeak addresses common clustering challenges, such as inaccurate cluster boundaries, sensitivity to noise, and misclassification, ultimately escalating to superior results on problematic and real-world data(D) [4], as demonstrated in Algorithm 1:Must-link constraints (similarity, $Con_{=}$): two objects $(d_i, d_{=})$ must be in the equivalent cluster, declared as $(d_i, d_{=}) \in C_K$.

Cannot-link constraints (dissimilarity, $Con_{\neq}$): two objects $(d_i, d_{\neq})$ must not be in the same cluster, declared as $(d_i, d_{\neq}) \in C_K$. Here, the previous information conjoined into the Density Peaks clustering algorithm to improve the performance. The algorithm has four important steps; we represent them as follows [10]:

Step 1: the density calculation p_i of a data object

$$pi = \sum_{j \in D} x(dist(i, j) - Eps),$$

where $x(x) = 1$ if $x < 0$ and $x(x) = 0$ otherwise, and $dist(i, j)$ the similarity among object ***i*** and object ***j***, ***D*** is the actual dataset to be separated. Giving to [10], the most convenient choice is to adjust Eps such that the neighbors’ ordinary number is between 1% and 2% of the total number of objects in the dataset.

Step 2: minimum similarity computation δ_i of a data object

$$\delta_i = \min_{j:j\in D, p_j > p_i} (dist(i,j)).$$

The minimum similarity δ_i of object ***i*** is measured through computing the minimum similarity between the object ***i*** and any other objects with higher density.

Step 3: the following criterion was utilized in the selection of the cluster centers:

$$\gamma_i = \delta_i \times p_i.$$

Cluster centers constantly have comparatively higher p, as well as comparatively higher δ. Thus, the larger the value of γ_i is, the more potential object ***i*** will be taken as a cluster center.

Step 4: During this stage, we talked about the unsupervised Density Peaks algorithm as well as the enhancement that turned out to be a Semi-supervised Density Peaks algorithm. The two points that follow explain why this is the case:

1. For unsupervised Density Peaks: each remaining object is allocated to the same cluster as its closest neighbor of higher density.
2. For Semi-Supervised Density Peaks (SDenPeak): it is an effort to allocate each remaining object d_i to its cluster. To be more certain, if the object d_i has some instances of cannot-link in the cluster C_k which d_i is allocated to; it needs to find another cluster as a near similarity δ until it does not violate any constraint. If not, the object d_i is directly allocated to the same cluster as its higher density for the nearest neighbor based on δ, while all the must-link instances linked to object d_i are also allocated to the cluster as similar as the object d_i. Finally, an instances' partition in D should return and satisfies all constraints [9].

Algorithm 1 [26]: Semi-Supervised Density Peaks
Require: Distance matrix (disM), cutoff distance (cd), Dataset original labels (labels), number of clusters (K)
Ensure: the consensus clustering Ck.

1: Get 10% of labels and put it in r
2: **for** (i = 1 : size(r))) **do**
3: **for** (j = i + 1 : size(r)) **do**
4: **if** (di; dj) 2 Ck **then**
5: Constraintsstatus = 1;
6: **else**
7: **if** (di; d≠) 2 Ck then
8: Constraintsstatus = 0;
9: **end if**
10: **end if**
11: disM(i; j) and disM(j; i) = Constraintsstatus
12: **end for**
13: **end for**
14: Density calculation pi = Σj∈D x(dist(i; j) − Eps)
15: Minimum similarity calculation δ i = minj:j∈D; pj>pi(dist(i; j))
16: Find γi to choose the cluster centers γi = δi × pi
17: Allocated each remaining object to the same cluster as its closest neighbor of higher density

4 SIMILARITY MEASUREMENTS IN SDENPEAK

The SDenPeak algorithm involves determining the degree of similarity between each object in the dataset to assign it to the centroid with the least similarity. Within the context of the clustering method, this similarity assessment plays a very important and essential role. According to the research findings, the degree of resemblance between two items can be determined using various methods; hence, we ought to select appropriate methods to get favorable outcomes. When this occurs, it is necessary to consider several crucial aspects, including the data attributes and the dimensions of the dataset. To perform similarity calculations within the SDenPeak algorithm, we utilized the following similarity measurement techniques: “Euclidean” [12], “Cosine” [14], “City block (Manhattan)” [16], “Minkowski” [18], “Earth Mover's” [20], and “Rapid Computation of the Maximal Information Coefficient” [22]. The following are the points that will describe each technique:

4.1 Euclidean distance

The Euclidean distance and the squared Euclidean distance are the most often used metrics for determining the degree of similarity or dissimilarity between data objects in the clustering environment [11, 12]. The similarity between two objects p_i and q_i is defined as follows:

$$d_{Euc} = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}.$$

The Deriving of Euclidean distance of two objects based on calculate the square root of the sum of the squares of the differences among corresponding values. The attitude of an object in a Euclidean n-space is a Euclidean vector [13], so p and q are Euclidean vectors, and their tips point to two objects. Euclidean distance is convenient for data measured at the same scale

4.2 Cosine Distance

Cosine similarity is a metric typically utilized for information retrieval and data mining, specifically in high-dimensional positive spaces [15]. Similarity-based on cosine takes into account the directions. It is measured by taking the cosine of the two vectors' existing angles. Two similar vectors are predicted to have a small angle relating them. The cosine correspondence of two objects S_i and S_j specified by:

$$\cos(\theta) = \frac{\sum_{i=1}^{d}(s_i \times s_i^{'})}{\sqrt{\sum_{i=1}^{d} s_i^2} \times \sqrt{\sum_{i=1}^{d} s_i^{'2}}},$$

where θ is the angle among S_i and S_j. The Cosine similarity based on the smaller the cosine angle is, the larger the resemblance of two objects, If the two objects are completely duplicate, the $\cos(\theta) = 1$; if the two objects are completely different, the $\cos(\theta) = 0$.

4.3 City Block (Manhattan)

The City block similarity [16, 17] involving two objects, a_i and b_i, with k dimensions, is figured as:

$$d_{CB} = \sum_{i=1}^{k} | a_i - b_i |.$$

The similarity between City blocks is always larger than or equal to zero. Whereas the measure would be zero for similar things, it would be high for objects that show a low degree of similarity. This approach is explained in greater detail if you consider two objects located in the plane. The City block similarity is also called the Manhattan similarity [16]. Along the hypotenuse, the Euclidean distance is the shortest similarity between two things. Instead, the city block similarity is measured as the similarity, plus the similarity, which is identical to how you go around a city (like Manhattan), where you should go around the buildings rather than going straight through. This is the method by which the City block similarity is calculated.

4.4 Minkowski Distance

The Euclidean similarity and the Manhattan similarity are both popularized by the Minkowski similarity [18, 19], which is a metric in a normed vector space. It can be thought of as a popularization of both of these similarities. These are the definitions of the Minkowski metric:

$$d_{Mk} = \left(\sum_{i=1}^{n} | x_i - y_i |^p \right)^{1/p},$$

for $p \geq 1$, the Minkowski inequality resulted by Minkowski distance metric. When $P < 1$, the similarity between (0, 0) and (1, 1) is $2^{\wedge}(1/p) > 2$, but the object (0, 1) is within space 1 of each of these objects. Because this contravenes the divergent triangle, for $p < 1$ it does not represent a metric [19].

4.5 Earth Mover's Distance

The earth mover's distance (EMD) [20] measures the degree of similarity between two likelihood distributions in the sphere of statistics. This approach is called the Wasserstein metric [21] in mathematics. It is of the utmost importance to recognize that the practical and functional similarity that can quantify the similarity between two distributions, patterns, or sequences is of supreme significance [20]. Let's say we own a collection of things in the dimension d. As an alternative to assigning a single distribution to the collection of objects, we can group them and then construct the point set in terms of the groups themselves.

"The signature" is the name given to this, and the degree of resemblance that exists between each of the traits is referred to as "ground distance" [20]. This is how the EMD is defined:

$$EMD(P,Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} d_{i,j}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j}}.$$

Let $D = [d_{i,j}]$ be the ground distance between groups p_i and q_i, and want to find a flow $F = [f_{i,j}]$, with $f_{i,j}$ the flow between p_i and q_i, that minimizes the total cost [21].

4.5 Rapid computation of the maximal information coefficient (RapidMic [22])

The maximal information coefficient, often known as the MIC, is a distance measurement of two variables developed specifically to facilitate the rapid exploration of datasets that contain numerous dimensions [22]. MIC is part of a large class of maximum information-based nonparametric discovery (MINE) statistics. They can be employed to characterize the significant connections in data sets and detect such relationships. It is necessary to compute the dependence measurement for each pair of items to discover a high-dimensional dataset. RapidMic is a cross-platform tool that uses parallel computing methods to achieve maximal information coefficients. It has played an important role in facilitating quick calculation. It has demonstrated that parallel processing can impact and reduce the time required for computation when working with large biological data. The system accommodates four methods of parallel analysis: one-pair analysis, all-pair analysis, two-set analysis, and master analysis.

5 EXPERIMENTAL DESIGN

Several experiments were run to determine which distance measurement with the maximum efficacy is possible from the semi-supervised Density Peaks algorithm for clustering. The experiments carried out utilizing KEEL tool, MATLAB, and RStudio. We then carry out the Friedman test [42] to obtain statistical justification in comparing the results.

Datasets

Eleven datasets, characterized by a mix of continuous and categorical variables and drawn from diverse application domains, were selected for this study to evaluate the semi-supervised density peak clustering algorithm with various distance measures. This number and the inherent diversity of these datasets align with the evaluation protocols established in prior seminal works in the field of clustering [45, 46, 47, 48], which have similarly employed a focused yet varied set of datasets to provide a robust initial assessment of algorithmic performance and facilitate comparative analysis. Our selection adheres to these established standards, allowing for a rigorous evaluation of the proposed approach under conditions commonly considered representative in the literature. Below is an overview of the key characteristics of the datasets listed in Table 1.

Table 1. Dataset Description

No.	Dataset	Short name	Objects	Features	Classes
1	Contextual Deterding vowel recognition data	vowel_context	990	13	11
2	Cleveland heart disease database	heartdisseaseC	303	13	5
3	QSAR biodegradation	Biodeg	1055	41	2
4	Flower bouquet	bouquet	880	892	3
5	Vowel	vowe	990	13	11
6	no2t	no2t	500	7	6
7	Glass Identification	Glas	214	9	6
8	Hungary heart disease database	Heartdisseaseh	294	13	5
9	BUPA Medical Research -Liver Disorders data	bupa	345	6	2
10	Blood Transfusion data tran	Tran	748	4	2
11	Plrx	Plrx	182	12	2

Clustering Evaluation Metrics

We must have clustering performance metrics to quantify the performance of clustering algorithms. The most widely used performance metrics are Root Mean Square Error (RMSE), Accuracy, and Micro-Precision (Micro-P). The root mean square error (RMSE) is a widespread statistical metric that states the degree to which the predicted cluster labels are different from the actual cluster labels as the square root of

the mean of the squared variations. The smaller the RMSE, the more successfully the clustering algorithm has grouped similar points and minimized errors in the assignment process. In quantifying clustering robustness, the root mean square error (RMSE) is a valuable statistic since it is especially useful when the impact of high errors is to be minimized. Accuracy, on the other hand, is the proportion of instances that were clustered correctly. This is accomplished by matching the predicted cluster labels with the ground truth labels, if there were any. High accuracy is one of the most essential indicators for evaluating the validity of clustering algorithms, particularly when dealing with semi-supervised or supervised clustering tasks. This is because high accuracy implies that the clustering model represents the underlying structure of the data in a good way. When there is an imbalance in the data, using merely accuracy can still result in an inaccurate representation of the precision and recall trade-offs. Since it calculates accuracy for all classes by merging true positives and false positives before generating the precision score, Micro-accuracy (Micro-P) is a more accurate measure, particularly in the case of multi-class clustering. This is because it calculates precision for all of the classes.

In contrast to Micro-P, which assigns high weights to heavy classes, Macro-precision treats all classes similarly. Because of this, this measure would be advantageous in situations where the class distributions of the dataset are not balanced. This statistic guarantees that the evaluation is conducted fairly based on the overall distribution of cases within the data set. A thorough explanation of clustering performance is presented by RMSE, Accuracy, and Micro-P when taken together. This description enables researchers and practitioners to select the most appropriate clustering method by considering the features of the data set and the particular objectives of the clustering task. Experiments are being carried out. For measurements, the outcome of clustering can be written as:

$$Micro - p = \frac{1}{m}\sum_{t=1}^{c} a_t.$$

The better the clustering performance should be the larger value of micro-p.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - y_i\right)^2}, \qquad Accuracy = \frac{Correctly\ Clustered\ Points}{Total\ Points} \times 100.$$

6 RESULT ANALYSIS

The experimental findings in this paper are presented using different datasets and performance metrics to allow a fair comparison of the literature. We selected eleven different datasets, varying in objects, features, and classes, for our experiments to comprehensively evaluate the Semi-supervised Density Peaks (SDenPeak) clustering algorithm. We wanted to examine the influence of employing different distance measures on the performance of clustering. For this purpose, we employed different distance measures in the SDenPeak algorithm and obtained clustering results for each dataset. When we received the results, we employed micro-precision as a performance measure to compare what distance measure provided the best clustering results. Coherently comparing the results allowed us to determine the most appropriate distance measure for SDenPeak. The experimental results shown in Table 2 provide us with the differences in performance on datasets and distance measures and allow us to select the most suitable distance measure for further enhancing SDenPeak's performance.

We observed from the experimental results that the RapidMIC distance measure worked better in ten out of eleven datasets with systematic improvement, confirming its effectiveness in guiding the SDenPeak algorithm to generate more accurate cluster structures. However, for the Plrx dataset, Earth Mover's Distance (EMD) was the best, with the highest performance value of 0.6264, outperforming all the other distances. Despite that, the RapidMIC distance also performed well on the same dataset with a score of 0.5989, the second best after EMD. These findings indicate that while Earth Mover's Distance may be suitable for some datasets, RapidMIC distance is the strongest and most reliable distance metric overall to use with the Semi-supervised Density Peaks Clustering Algorithm. RapidMIC distance can be considered the ideal similarity measurement to enhance SDenPeak's clustering performance and be a good choice for future clustering applications as it performs well consistently on a range of datasets (see Table 2).

For Micro-p, the results confirm that RapidMIC distance remains the most consistent and reliable metric when paired with the SDenPeak algorithm, even though Earth Mover's Distance is more appropriate to some datasets such as Plrx. The RapidMIC distance is the most optimal similarity measure for achieving best clustering performance due to its exceptional performance on every dataset. It is a reliable option for future clustering as it is immune to the inherent problem of other datasets with minimal performance variation. In semi-supervised learning scenarios where the prior knowledge would be utilized to guide the clustering, RapidMIC distance comes through as a reliable method that can powerfully improve the

accuracy and performance of cluster algorithms, particularly with the emergence of the clustering field. On that basis, it provides an excellent choice for general application of cluster algorithms in the real world, particularly where high-quality and reliable results are needed (see Figure 1).

Table 2. Average micro-precisions (The highest micro-precision among different distance on each dataset is bolded)

Dataset	RapidMic	EMD	Euclidean	Cosine	Manhattan	Minkowski
1	**0.2263**	0.1030	0.1091	0.1758	0.1030	0.1091
2	**0.5314**	0.3036	0.2904	0.2739	0.2475	0.2904
3	**0.5330**	0.5182	0.5083	0.5000	0.5215	0.5083
4	**0.6375**	0.4057	0.5364	0.6034	0.4773	0.5364
5	**0.2263**	0.1030	0.1091	0.1758	0.1030	0.1091
6	**0.4000**	0.2460	0.2460	0.2920	0.2420	0.2460
7	**0.5000**	0.4065	0.4019	0.4907	0.4533	0.4019
8	**0.5102**	0.2721	0.3673	0.3401	0.3946	0.3673
9	**0.6986**	0.5188	0.5478	0.5420	0.5623	0.5478
10	**0.7620**	0.5053	0.6043	0.6417	0.5735	0.6043
11	0.5989	**0.6264**	0.5385	0.5275	0.5604	0.5385
AVG	**0.5113**	0.3644	0.3872	0.4148	0.3853	0.3872

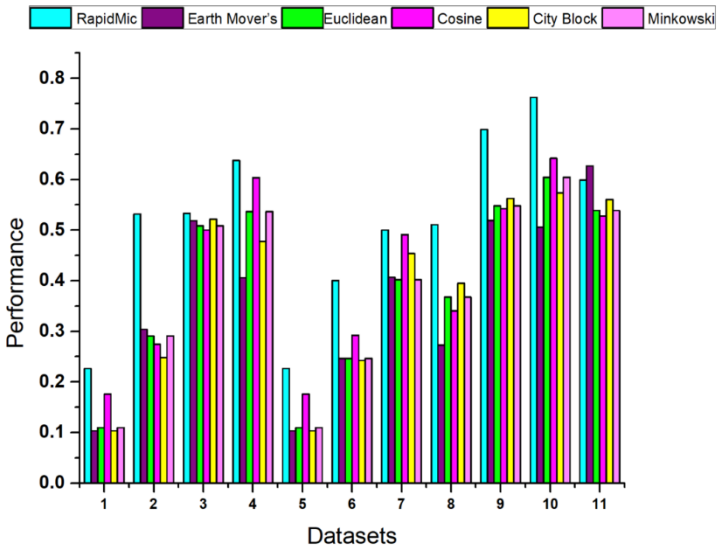

Fig. 1. Micro-precisions for all methods

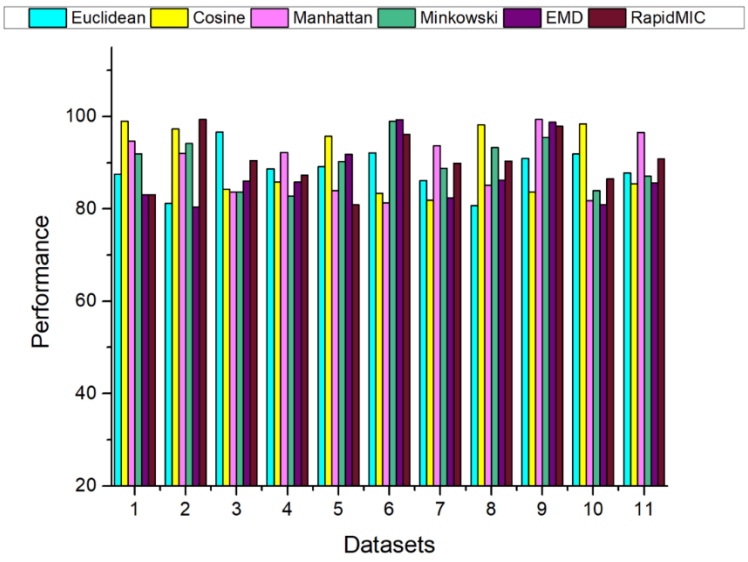

Fig. 2. Accuracy for all methods

For Accuracy and RMSE, the findings that were gathered from our study offer a substantial amount of information concerning their operational efficiency and performance. In terms of accuracy, Euclidean Distance provides consistent performance throughout most datasets and frequently obtains the best accuracy. On the other hand, Cosine Similarity and Manhattan Distance provide good clustering performance, although there are some minor changes among datasets related to their performance. The Euclidean Distance is significantly superior to the Minkowski Distance (p=3), which produces reliable results but does not perform significantly better. On the other hand, Earth Mover's Distance (EMD) and RapidMIC are less precise, particularly when the datasets are high-dimensional, such as "bouquet." There are occasions when Manhattan Distance performs better than Euclidean in huge datasets regarding features. This is likely because it is better able to handle high-dimensional data.

Additionally, the RMSE study lends credence to the findings above by demonstrating that the Euclidean Distance has the lowest RMSE for most datasets, substantiating its effectiveness. On the other hand, the Minkowski and Manhattan Distances have larger RMSE values, but they are still acceptable. On the other hand, EMD and RapidMIC have the highest RMSE, which indicates that they are responsible for a bigger portion of the clustering errors.

Table 3. Average of Accuracy

Dataset	Euclidean	Cosine	Manhattan	Minkowski	EMD	RapidMIC
vowel_context	87.49	99.01	94.63	91.97	83.12	83.11
heartdisseaseC	81.16	97.32	92.02	94.16	80.41	99.39
biodeg	96.64	84.24	83.63	83.66	86.08	90.49
bouquet	88.63	85.82	92.23	82.78	85.84	87.32
vowe	89.12	95.70	83.99	90.28	91.84	80.92
no2t	92.15	83.41	81.30	98.97	99.31	96.16
glas	86.09	81.95	93.68	88.80	82.44	89.90
heartdisseaseh	80.68	98.18	85.17	93.25	86.23	90.40
bupa	90.93	83.69	99.39	95.50	98.78	97.89
tran	91.95	98.43	81.76	83.91	80.90	86.50
Plrx	87.77	85.42	96.57	87.13	85.61	90.85
AVG	88.41	90.28	89.48	90.03	87.32	90.26

Table 4. Average of RMSE

Dataset	Euclidean	Cosine	Manhattan	Minkowski	EMD	RapidMIC
1	0.0711	0.1703	0.0612	0.1980	0.1658	0.0798
2	0.0508	0.1723	0.1560	0.1594	0.1657	0.0611
3	0.1038	0.0674	0.1795	0.1435	0.0996	0.0595
4	0.0966	0.0988	0.1594	0.1456	0.1831	0.1208
5	0.0679	0.1570	0.1641	0.1342	0.1656	0.1241
6	0.1284	0.1141	0.0538	0.0662	0.0547	0.1455
7	0.0972	0.1263	0.1861	0.0874	0.1116	0.1633
8	0.0843	0.0615	0.0935	0.0742	0.1895	0.1712
9	0.1450	0.1807	0.1706	0.0780	0.1839	0.1309
10	0.1711	0.1844	0.0977	0.0665	0.0842	0.1141
11	0.1727	0.1791	0.0510	0.1266	0.1126	0.0833
AVG	0.1081	0.1375	0.1248	0.1163	0.1378	0.1140

Table 5 describes the Friedman test, lower average rankings reflect better overall performance. RapidMIC's average ranking of 3.09 shows it was best performing on average, and EMD's higher average ranking of 3.91 indicates it was worst performing across the datasets. The remaining similarity measures, Manhattan, Cosine, Minkowski, and Euclidean, are intermediate, with Manhattan and Cosine generally performing better than Minkowski and Euclidean. These rankings help identify which similarity measures are better or more accurate to use in clustering jobs across different types of datasets. The Friedman test results would suggest substantial differences in the performance of these similarity measures and that one needs to select the best measure for specific datasets with caution.

In table 6, none of the RapidMIC comparisons with the similarity measures are statistically significant at Holm's adjusted α levels. All p-values are larger than the corresponding thresholds in Holm's procedure, and hence all null hypotheses are not rejected. This implies that although RapidMIC had the best average rank, the difference in performance is not statistically significant compared to other methods based on Holm's criteria. This discovery attests to the postulation that while RapidMIC performs consistently well,

similarity metrics such as Manhattan, Cosine, and Minkowski are statistically competitive options for using the SDenPeak clustering algorithm.

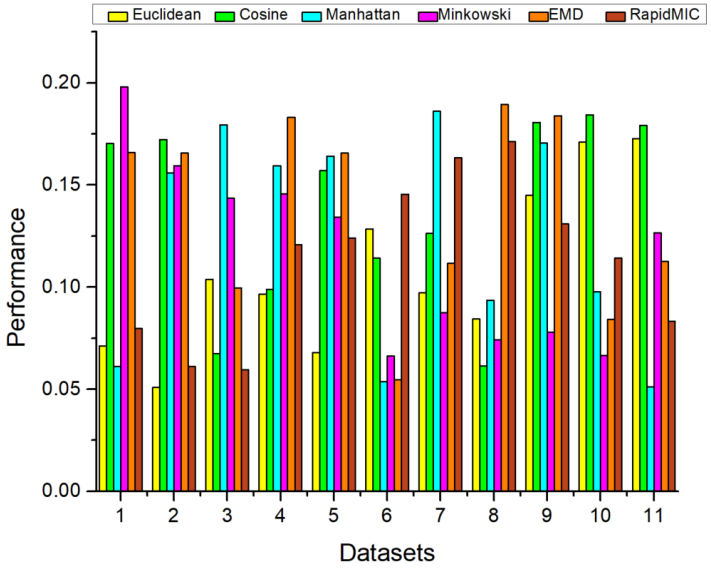

Fig. 3. RMSE for all methods

Table 5. Rankings obtained through Friedman's test

Similarity Measure (Methods)	Ranking
RapidMIC	3.09
Manhattan	3.36
Cosine	3.45
Minkowski	3.55
Euclidean	3.64
EMD	3.91
RapidMIC	3.09
Manhattan	3.36
Cosine	3.45
Minkowski	3.55
Euclidean	3.64

Table 6. Holm's Post-Hoc Test (α = 0.05), RapidMIC as the control

No.	Algorithm	$z = (R_0 - R_i) / SE$	p value	Holm/Hochberg/Hommel (α/i)	Hypothesis
1	EMD	-1.03	0.3023	0.010	Not Rejected
2	Euclidean	-0.69	0.4884	0.0125	Not Rejected
3	Minkowski	-0.58	0.5608	0.0167	Not Rejected
4	Cosine	-0.45	0.6517	0.025	Not Rejected
5	Manhattan	-0.34	0.7357	0.050	Not Rejected

7 CONCLUSION AND FUTURE WORKS

Based on the experimental results that we have obtained, we found that the application of the RapidMIC distance measure in combination with the Semi-supervised Density Peaks (SDenPeak) Clustering algorithm performed well in micro-p on the majority of datasets. Our experiments were conducted on eleven different datasets with varying characteristics to comprehensively assess and verify the algorithm's effectiveness and applicability with varying data conditions. Our findings stress the importance of selecting an appropriate distance measure based on the specific traits of a dataset, as an inappropriate one can lead to increased computational complexity and lower clustering quality. As this is a significant factor, we intend to extend our work by further experimenting with SDenPeak using different distance measures to examine their effects on clustering quality. Moreover, While this paper provides a comprehensive evaluation of the SDenPeak algorithm with the selected 11 datasets, we recognize the need for a more extensive analysis using a larger set of datasets; we plan to generalize our study to support other

semi-supervised clustering algorithms, including semi-supervised fuzzy c-means, semi-supervised k-means, semi-supervised affinity propagation, and multi-view clustering algorithms. By extending our research to these algorithms, we hope to gain more insight into how different measures of distance impact semi-supervised clustering and encourage the general usefulness of clustering techniques on a broader range of data-based applications.

REFERENCES

[1] *Zhong G, Pun CM.* Self-taught multi-view spectral clustering. Pattern Recognition. 2023 Jun 1;138:109349.
[2] *Zhang C, Ni M, Zhong Y, Wei H, Qiu K.* Density-ratio peak based semi-supervised algorithm for access network user behavior analysis. IEEE Access. 2019 May 6;7:62904-10.
[3] *Jain, A., Jin, R., & Chitta, R.* (2014). Semi-supervised clustering. Handbook of Cluster Analysis, 1-35.
[4] *Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S.* (2001, June). Constrained k-means clustering with background knowledge. In ICML (Vol. 1, pp. 577-584).
[5] *Sinha A, Jana PK.* Improved affinity propagation clustering algorithms: A PSO-based approach. Knowledge and Information Systems. 2025 Feb;67(2):1681-711.
[6] *Shabani N, Wu J, Beheshti A, Sheng QZ, Foo J, Haghighi V, Hanif A, Shahabikargar M.* A comprehensive survey on graph summarization with graph neural networks. IEEE Transactions on Artificial Intelligence. 2024 Jan 8;5(8):3780-800.
[7] *González-Almagro G, Peralta D, De Poorter E, Cano JR, García S.* Semi-supervised constrained clustering: An in-depth overview, ranked taxonomy and future research directions. Artificial Intelligence Review. 2025 Mar 7;58(5):157.
[8] *Basu, S., Banerjee, A., & Mooney, R.* (2002). Semi-supervised clustering by seeding. In In Proceedings of 19th International Conference on Machine Learning (ICML-2002.
[9] *Fan, W. Q., Wang, C. D., & Lai, J. H.* (2016, March). SDenPeak: Semi-supervised Nonlinear Clustering Based on Density and Distance. In Big Data Computing Service and Applications (BigDataService), 2016 IEEE Second International Conference on (pp. 269-275). IEEE.
[10] *Rodriguez, A., & Laio, A*. (2014). Clustering by fast search and find of density peaks. Science, 344(6191), 1492-1496.
[11] *Liberti, L., Lavor, C., Maculan, N., & Mucherino, A.* (2014). Euclidean distance geometry and applications. Siam Review, 56(1), 3-69.
[12] *Alsuhibany, S. A., Almushyti, M., Alghasham, N., & Alkhudier, F.* (2016, November). Analysis of free-text keystroke dynamics for Arabic language using Euclidean distance. In Innovations in Information Technology (IIT), 2016 12th International Conference on (pp. 1-6). IEEE.
[13] *Nourbakhsh A, Jadidi M, Shahriari K.* Clustering bike sharing stations using Quantum Machine Learning: A case study of Toronto, Canada. Transportation Research Interdisciplinary Perspectives. 2024 Sep 1;27:101201.
[14] *Hao, L., & Gang, N.* (2016, October). A novel diagnosis method for intelligent IETM platform based on cosine similarity and fuzzy semantic inference. In Prognostics and System Health Management Conference (PHM-Chengdu), 2016 (pp. 1-6). IEEE.
[15] *Hernandez, A. F. R., & Garcia, N. Y. G.* (2016). Distributed processing using cosine similarity for mapping Big Data in Hadoop. IEEE Latin America Transactions, 14(6), 2857-2861.
[16] *Wang W, Chen T, Liu H, Zhang J, Wang Q, Jiang Q.* Depth perception optimization of mixed reality simulation systems based on multiple-cue fusion. Journal of the Society for Information Display. 2024 Aug;32(8):568-79.
[17] *De Carvalho, F. D. A., Barbosa, G. B., & Pimentel, J. T.* (2013, October). Partitioning fuzzy c-means clustering algorithms for interval-valued data based on city-block distances. In Intelligent Systems (BRACIS), 2013 Brazilian Conference on (pp. 113-118). IEEE.
[18] *Rodrigues ÉO.* Combining Minkowski and Chebyshev: New distance proposal and survey of distance metrics using k-nearest neighbours classifier. Pattern Recognition Letters. 2018 Jul 15;110:66-71.
[19] *Howard, S. D., & Sirianunpiboon, S.* (2012, August). Fast tests for the common causality of time-of-arrival events from their mutual Minkowski distances. In Statistical Signal Processing Workshop (SSP), 2012 IEEE (pp. 101-104). IEEE.
[20] *Xu, J., Lei, B., Gu, Y., Winslett, M., Yu, G., & Zhang, Z.* (2015). Efficient similarity join based on earth mover's distance using MapReduce. IEEE Transactions on Knowledge and Data Engineering, 27(8), 2148-2162.
[21] *Beecks, C., Uysal, M. S., & Seidl, T.* (2015, December). Earth Mover's Distance vs. Quadratic form Distance: An Analytical and Empirical Comparison. In Multimedia (ISM), 2015 IEEE International Symposium on (pp. 233-236). IEEE.

[22] *Tang, D., Wang, M., Zheng, W., & Wang, H.* (2014). RapidMic: Rapid Computation of the Maximal Information Coefficient. Evolutionary bioinformatics, 10, 11.
[23] *Rodriguez, A., & Laio, A.* (2014). Clustering by fast search and find of density peaks. Science, 344(6191), 1492–1496.
[24] *Wang, H., Zhang, R., Li, Y., & Sun, M.* (2020). Improving density peaks clustering with semi-supervised pairwise constraints. Knowledge-Based Systems, 192, 105369.
[25] *Yang, C. H., Lee, B., Lee, Y. I., Chung, Y. F., & Lin, Y. D.* (2025). An autoencoder-based arithmetic optimization clustering algorithm to enhance principal component analysis to study the relations between industrial market stock indices in real estate. Expert Systems with Applications, 266, 126165.
[26] *Mustafa K, Wang H, Zhou Y, Song J.* Semi-supervised cluster ensemble based on density peaks. In Data Science and Knowledge Engineering for Sensing Decision Support: Proceedings of the 13th International FLINS Conference (FLINS 2018) 2018 (pp. 645-651).
[27] *Wang, M.* (2025). Hybrid data clustering algorithm and interactive experience in E-learning electronic course simulation of legal education. Entertainment Computing, 52, 100760.
[28] *Yang, J., Hu, K., Wang, F., Zhang, J., Bao, J., & Liu, W.* (2025). A Partial Discharge Diagnosis Method for GIS Based on a Semi-supervised Classification Framework and Density Peak Clustering Algorithm. IEEE Transactions on Instrumentation and Measurement.
[29] *Zhang, X., Li, J., & Zhao, Y.* (2021). Enhancing clustering performance using semi-supervised density peaks with adaptive similarity measures. Journal of Machine Learning Research, 22(1), 1–20.
[30] *Aggarwal, C. C., Hinneburg, A., & Keim, D. A.* (2001). On the surprising behavior of distance metrics in high-dimensional space. International Conference on Database Theory (ICDT), 420–434.
[31] *Ding, C., He, X., Zha, H., Gu, M., & Simon, H. D.* (2008). A min-max cut algorithm for graph partitioning and data clustering. Proceedings of the IEEE International Conference on Data Mining (ICDM), 107–114.
[32] *Rubner, Y., Tomasi, C., & Guibas, L. J.* (2000). The Earth Mover's Distance as a metric for image retrieval. International Journal of Computer Vision, 40(2), 99–121.
[33] *Liu, Y., Wang, J., Zhang, X., & Li, H.* (2021). A comparative study of distance metrics for semi-supervised clustering algorithms. Pattern Recognition Letters, 144, 12–19.
[34] *Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., ... & Mitzenmacher, M.* (2011). Detecting novel associations in large datasets. Science, 334(6062), 1518–1524.
[35] *Tang, D., Wang, M., Zheng, W., & Wang, H.* (2014). RapidMic: rapid computation of the maximal information coefficient. Evolutionary bioinformatics, 10, EBO-S13121.
[36] *Malmberg, C., Torpner, J., Fernberg, J., Öhrn, H., Ångström, J., Johansson, C., ... & Kreuger, J.* (2022). Evaluation of the speed, accuracy and precision of the QuickMIC rapid antibiotic susceptibility testing assay with Gram-negative bacteria in a clinical setting. Frontiers in Cellular and Infection Microbiology, 12, 758262.
[37] *Tang, J., Xu, D., Cai, Q., Li, S., & Rezaeipanah, A.* (2024). Towards a semi-supervised ensemble clustering framework with flexible weighting mechanism and constraints information. Engineering Applications of Artificial Intelligence, 136, 108976.
[38] *Kadhim MR, Tian W, Khan T.* Rapid clustering with semi-supervised ensemble density centers. In2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing 2019 Dec 14 (pp. 230-235). IEEE.
[39] *González-Almagro, G., Sánchez-Bermejo, P., Suarez, J. L., Cano, J. R., & García, S.* (2024). Semi-supervised clustering with two types of background knowledge: Fusing pairwise constraints and monotonicity constraints. Information Fusion, 102, 102064.
[40] *Liu, X., Zhang, M., Liu, Y., Liu, C., Li, C., Wang, W., ... & Bouyer, A.* (2024). Semi-supervised community detection method based on generative adversarial networks. Journal of King Saud University-Computer and Information Sciences, 36(3), 102008.
[41] *Hussein, A. S., Li, T., Yohannese, C. W., & Bashir, K.* (2019). A Novel Hybrid Approach Based on Rough Set for Classification: An Empirical Comparative Study. Journal of Multiple-Valued Logic & Soft Computing, 33.
[42] *Diallo, B., Hu, J., Li, T., Khan, G. A., & Hussein, A. S.* (2022). Multi-view document clustering based on geometrical similarity measurement. International Journal of Machine Learning and Cybernetics, 1-13.
[43] *Cleophas, T. J., Zwinderman, A. H., Cleophas, T. J., & Zwinderman, A. H.* (2016). Non-parametric tests for three or more samples (Friedman and Kruskal-Wallis). Clinical data analysis on a pocket calculator: understanding the scientific methods of statistical reasoning and hypothesis testing, 193-197.

[44] *Kadhim MR, Zhou G, Tian W.* A novel self-directed learning framework for cluster ensemble. Journal of King Saud University-Computer and Information Sciences. 2022 Nov 1;34(10):7841-55.
[45] *Hasan N, Alam MG, Ripon SH, Pham PH, Hassan MM.* An autoencoder-based confederated clustering leveraging a robust model fusion strategy for federated unsupervised learning. Information Fusion. 2025 Mar 1;115:102751.
[46] *Sun H, Pan J.* Heart disease prediction using machine learning algorithms with self-measurable physical condition indicators. Journal of data analysis and information processing. 2023 Jan 18;11(1):1-0.
[47] *Zhang J, Wu M, Sun Z, Zhou C.* Learning from Crowds Using Graph Neural Networks with Attention Mechanism. IEEE Transactions on Big Data. 2024 Mar 19.
[48] *de Menezes JA, Gomes JC, de Carvalho Hazin V, Dantas JC, Rodrigues MC, Nogueira PL, dos Santos WP.* Classification of Motor Imagery EEG Signals Based on Sparse Representations of Empirical Mode Decomposition Features. In Advanced Electroencephalography Analytical Methods (pp. 208-241). CRC Press.

Received: December 15,2025

Citation: *Ahmed Saad Hussein.* 2026. A Comparative Evaluation of Similarity Measures for Semi-Supervised Density Peaks Clustering. *International journal of theoretical and applied issues of digital technologies*. Volume 9, Issue 2, pp. 7-19. https://doi.org/10.62132/ijdt.v9i2.371.

СРАВНИТЕЛЬНАЯ ОЦЕНКА МЕР СХОДСТВА ДЛЯ КЛАСТЕРИЗАЦИИ ПИКОВ ПЛОТНОСТИ С ПОЛУКОНТРОЛИРУЕМЫМ ОБУЧЕНИЕМ

Ahmed Saad Hussein [1,2]

[1] Department of Cybersecurity Engineering Technologies, Technical Engineering College, Al-Farabi University, Baghdad 10001, Iraq
[2] Department of Mobile Communications and Computing Engineering, College of Engineering, University of Information Technology and Communications, Baghdad, Iraq

Ahmed.Hussein@alfarabiuc.edu.iq

Аннотация. Алгоритм полуконтролируемой кластеризации Density Peak (SDenPeak) известен своей эффективностью и простотой в задачах кластеризации. Он улучшает производительность кластеризации за счет добавления попарных ограничений, ограничений на обязательное и невозможное связывание, которые управляют процессом группировки, устанавливая сходство и несходство между точками данных. Одним из ключевых факторов точности кластеризации является выбор меры сходства, поскольку различные меры отражают разнообразные структурные характеристики данных. Проблема заключается в том, что не существует универсальной лучшей меры сходства, поскольку выбор подходящей меры является сложной задачей, зависящей от характера данных. Для изучения влияния шести мер сходства на производительность алгоритма SDenPeak в данном исследовании систематически оцениваются шесть мер евклидово расстояние, косинусное сходство, расстояние между городскими кварталами (Манхэттенское расстояние), расстояние Минковского, расстояние перемещения земли и расстояние быстрого вычисления максимального информационного коэффициента с целью понимания их влияния. Для оценки точности кластеризации и структурной согласованности каждой из мер были проведены обширные эксперименты на реальных наборах данных. Полученные результаты представляют сравнительную информацию об эффективности различных мер сходства и иллюстрируют их применимость к различным распределениям данных, предоставляя полезное руководство по достижению наилучшей производительности кластеризации в полуконтролируемых моделях.

Ключевые слова: кластеризация, полуконтролируемое обучение, измерение сходства, пики плотности, реальные наборы данных.