axioms

MDPI

Type of the Paper (Article.)

An Improved Clustering Algorithm for Multi-Density Data

Walid Atwa [1]* and Abdulwahab Ali Almazroi[1]

1 University of Jeddah, College of Computing and Information Technology at Khulais, Department of Information Technology, Jeddah, Saudi Arabia; wsatwa@uj.edu.sa, *aalmazroi@uj.edu.sa*

* Correspondence: wsatwa@uj.edu.sa;

Abstract: Clustering is an important method that divides a dataset into groups with similar data using similarity metrics. However, discovering clusters in different densities, shapes and distinct sizes is still a challenging task. DBSCAN is a density-based algorithm that define clusters of different sizes and shapes. However, misapplied to clusters of different densities due to its global attributes (*Eps*. and *MinPts*) that generate single density. Furthermore, most existing algorithms are unsupervised methods where available prior knowledge are useless. In this study, we focus on using the available knowledge in the form of pairwise constraints to increase the clustering performance. Based on the statistical characteristics of the dataset, we partition the dataset into different density levels and expand the clusters using the available pairwise constraints.

Keywords: Clustering; DBSCAN; Pairwise Constraints; Multi-Density Data.

1. Introduction

Clustering is utilized to arrange dataset to a limited set of clusters based on selected similarity metrics [1]. There are different clustering categories like density-based algorithms that can identify clusters of distinct sizes and shape. Hence, this algorithm do not required the specification of clusters number. Rather, clusters are identified in a densely connected region which grows based on the direction of density. Meaning, density based algorithms identify clusters based on regions that has high density that are separated from the regions with low densities [2-5].

DBSCAN is one of the density-based algorithms [6]. The algorithm utilized global criterion such as e-distance and *MinPts* to define density criterion. Multi-density clustering algorithms were also proposed [3], [7], [8]. However, density-based algorithms that can effectively cluster multi-density and arbitrary shapes are limited, with low performance in most of the existing algorithms. In this paper, we propose an active semi-supervised clustering method for clustering multi-density data and arbitrary shapes, SSMD. Our method is separated into two core parts. Firstly, we study the issue of generating a set of pairwise constraint that improve clustering performance. Secondly, we partition the dataset into dissimilar density level sets and then expand the clusters using generated pairwise constraints. We conduct experiments over artificial and real datasets. Comparison results reveal the efficiency and effectiveness of the proposed method.

The main contributions are twofold. Firstly, our proposed algorithm can discover clusters of changing densities, sizes and shapes, even in the existence of outliers and noise. Secondly, we evaluated the performance proposed algorithm on multiple experiments with varied datasets in comparison to other algorithms.

The remaining part of the paper is ordered as follows. In Section 2, the related work is outlined and carefully discussed. The proposed algorithm is discussed in Section 3. Section 4 gives the experimental setup with the detailed explanation of the obtained results. And lastly, the paper is concluded in Section 5.

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Axioms* **2022**, *11*, x. https://doi.org/10.3390/xxxxx

Academic Editor: Firstname Lastname

Received: date
Accepted: date
Published: date

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

2. Related Work

In this section, key related literature on clustering methods are highlighted and discussed. Hence, literature on density-based clustering and pairwise constraints are critically reviewed.

2.1. Density-based Clustering

In a study by Ankerst et al., a new algorithm named OPTICS was proposed for the sole purpose of clustering analysis [9]. The algorithm execute and store dual parameters which are core distance and reachability distance for cluster identification with different densities. The algorithm further create an ordering process for the dataset which will represent its density-based clustering structure. The study revealed the process of extracting not both traditional clustering information and intrinsic clustering structure automatically with good efficiency. The two studies by Jahirabadkar et al. and Liu et al. also utilizes kNN distances of all selected points so as to get the density for a given dataset. The authors further utilize various E-distances for sparse and dense clusters [2], [5]. In another study by Huang et al., the authors proposed a clustering algorithm to successfully determine the e-distance for each density. The authors use the proposed algorithm to identify clusters in a given dataset [10]. The result shows some promise. Based on the issue of identifying clusters in high dimensional data, if we have clusters in distinct sizes, shapes, and density, Ertoz et al. proposed a new technique for clustering to address the issue [11]. Firstly, the proposed algorithm check for the nearest neighbors (NN) of individual data points and further describes the similarity that resides between the points with respect to the number of NN shared by the points. Hence, the algorithm defines and build clusters around these defined points. An experiment on various datasets shows that the proposed algorithm achieves good performance compared to traditional methods.

Rodriguez and Laio proposed a new approach with the fundamental idea that cluster cores are categorized by greater density in comparison to their neighbors [12]. The result demonstrates the high performance of the proposed approach on various tests. In a study by Comanicin and Meer, a technique was proposed to analyze complex multimodal feature space. [13]. The proposed technique used a mean shift as the plain computational module, which is an old pattern recognition process. An experimental result demonstrates the technique best performance. With the aim of detecting clusters of distinct shapes and sizes, Liu et al. proposed an altered version of DBSCAN algorithm coined Entropy and Probability Distribution (EPDCA) [6]. Testing on benchmark datasets indicates that clustering results based on EPDCA achieve good performance. Kim et al. proposed an optimized combinatorial clustering algorithm specifically design for noisy performance. This algorithm is vital particularly for data that is large with random sampling [14]. The result shows that the proposed clustering algorithm outperformance various traditional approaches that is compared with.

An enhanced clustering algorithm named NG-DBSCAN was proposed by Lulli et al. the proposed algorithm works on arbitrary data [15]. The result of the study based on real and artificial data demonstrates the algorithm scalability and performance. Kim et al. further proposed a density-based clustering algorithm coined approximate adaptive AA-DBSCAN [4]. The proposed algorithm focused on minimizing extra computation needed to determine parameters by utilizing e-distance for each density hen identifying clusters. Based on an experiment conducted, the result shows a significant improvement with regards to clustering performance with decrease in running time. In another study by Zhang et al., proposed ~~coined~~ GCMDDBSCAN because DBSCAN cannot handle databases that are large [3]. With the proposed algorithm, the authors highlighted that the clustering capabilities on datasets that are large has improved together with clustering accuracy.

FlockStream was proposed by Forestiero et al. The algorithm is based on multi-agent system and the results demonstrate that FlockStream has good performance on both synthetic and real dataset [16]. Chen and Tu proposed a framework named D-Stream with

the aim of clustering stream data by utilizing a density-based approach [17]. The result indicates that D-Stream achieves good quality and efficacy.

Recently, semi-supervised clustering algorithms try to discover clusters of random shapes. Nevertheless, they cannot deal with the multi-density data efficiently. C-DBSCAN [18] and SSDBSCAN [19] ~~and is~~ typical semi-supervised clustering algorithms based on DBSCAN. The result revealed that SSDBSCAN is capable of detecting arbitrarily shapes and evolving clusters with current pairwise constraints.

2.2. Pairwise Constraints

Constraint-based methods are proposed in recent years [20]. These methods are an extension of known unsupervised clustering algorithms [21], [22]. The methods utilized constraints in adapted clustering procedure for learning similarity metrics. In recent years, various constraint-based methods were proposed for clustering, specifically in clustering algorithms like spectral clustering and K-means [22] [23].

Basu et al. proposed a clustering framework that handle large and high dimensional datasets. The result shows an improvement in clustering accuracy with little supervision required. Shental et al. proposed a framework for the composition of lateral information in the form of similarity constraints into the model estimation process [24]. The authors further introduce EM procedure and generalized EM procedure that handles both positive constraints (for the former), and negative constraints (for the latter). The algorithm shows significant improvements. In another study by Bilenko et al., a new clustering algorithm was proposed that integrates two methods, which are the constraint-based method and distance-function learning methods for semi-supervised clustering [25]. Based on an experimental study, the result revealed that the proposed algorithm provides better clusters. In another study by Wang et al., a framework for constraint spectral clustering was proposed. The framework encodes the constraints into the constrained optimization problem by encoding the level of belief in constraints [20].

In a study by Wagstaff et al., an algorithm was proposed with background knowledge [21]. The experiment was conducted with artificial constraints on various datasets, the result shows a significant improvement. Rangapuram and Hein proposed a clustering method that is based on tight relaxation of constraint normalized cut [22]. The proposed method guaranteed the satisfy all constrained. The method further allows the optimization of the trade-off between number of violated constraints and normalized cut. The result shows some improvements. Xu et al. proposed a new clustering technique named ACCESS that shows some improvement and has indicated that ACCESS outdoes existing constrained spectral clustering methods.

A method for the selection of pairwise constraints from unlabeled data was proposed by Masud et al. [26]. The proposed method first generate initial set of clusters using the I-nice method. Hence, for each cluster, a set of dense objects are known with local density estimation. One of the merits of the proposed method is the lack of need for label information of data for pairwise constraints selection.

3. SSMD Algorithm

Semi-super vised clustering algorithms utilizes a set of class labels constraints on some examples to help unsupervised clustering. We propose an active semi-super vised clustering for multi-density dataset (called ***SSMD***) that attempts to find out clusters with distinct densities and arbitrary shapes. *SSMD* algorithm has two main parts: (1) Partition the dataset into distinct density level sets. (2) Apply *Semi-DBSCAN* algorithm on each density level set. Where *Semi-DBSCAN* is an extension of *DBSCAN* that make full use of existing pairwise constraints. The pseudocode of our *SSMD* is revealed in algorithm 1.

First, we compute the density function for each data point as shown in Eq (1). Then sorting the density for all the data points in descending order to evaluate the variation

between each two sequence points. Density variation explain how much denser or sparser the points. The density variation of point p_i with respect to p_j is computes in Eq (2).

$$Den(p, MinPts) = \frac{kdist(p,k)}{k} \quad (1)$$

Where $kdist(p, k)$ defines the number of p's neighbors and k is an arbitrary positive integer that explain k-nearest neighbor distance.

$$DenVar(p_i, p_j) = \frac{|Den(p_i,k) - Den(p_j,k)|}{Den(p_i,k)} \quad (2)$$

Finally, partitioning the dataset into a list of density level set (*DLS*). Where each *DLS* contain the set of data points whose densities are approximately the same. We compute density variation threshold (τ) according to the statistical characteristics of the sorted density variation list.

$$\tau = \frac{\sum_{i=0}^{n} DenVar_{(i)}}{n} \quad where \;\; DenVar_{(i)} > 0 \quad (3)$$

Finally, we separate the data points into different densities based on the computed density variation threshold (τ) as follow:

$$p_i, p_j \in DLS_k \quad if \quad DenVar(p_i, p_j) \leq \tau \quad (4)$$

After density levels partitioning, we acquire a list of density level sets (for short *DLSList*). Then we need to find representative *Eps* for each density level set. We compute *Epsi* for DLS_i as follows:

$$Eps_i = \max(DLS_i) \cdot \sqrt{\frac{median(DLS_i)}{mean(DLS_i)}} \quad (5)$$

Where *max*, *median* and *mean* are the maximum, median and mean values of each DLS_i respectively.

Algorithm 1 SSMD

Input:

Data set (*D*), must-link and cannot-link constraints (*ML*, *CL*), number of objects in a neighborhood (*MinPts*)

Output:

Set of clusters

Begin

Compute density value for all data points according to *eq(1)*

Sort density list in descending order

Compute Density variation values using *eq(2)*

Generate Density Level Set (*DLS*) using threshold in *eq(3)*

EpsList = EstimateEps(DLSList);

For each DLS_i in *DLSList*

Semi-DBSCAN(DLS_i, *ML*, *CL*, *MinPts*, Eps_i)

End For

Return all clusters

End

With density levels partitioning and parameters estimation done, we carry on the clustering process: initialize global parameter $MinPts = k$, adopt *Semi-DBSCAN* algorithm for each Eps_i in *EpsList* on corresponding DLS_i. The algorithm s*emi-DBSCAN* can finds a set of clusters and a (probably empty) set of noises. Initially, the constraints is preprocess. Constraints given by human experts may be incomplete. Some constraints are not explicitly given, for example, if we have a $ML(x_i, x_j)$ and $ML(x_i, x_k)$ then we have $ML(x_j, x_k)$. Then, we use constraints to monitor the process of growing clusters in Semi-DB.SCAN as presented in algorithm 2.

First, computing the neighborhood for each unclassified data point and compared with *MinPts* to determine whether it can be added to current cluster or noise set. Then, expanding a cluster using the pairwise constraints as follow:

1. Add all data points have a must-link constraint with d to current cluster.
2. Add all data points in d's neighborhood that does not violated cannot-link constraint with d to current cluster.

Algorithm 2 Semi-DBSCAN (*D, ML, CL, MinPts, Eps***)**

Begin

Cluster Id: = 0;

For each *d* in *D*

If *d* is *UNC LASSIFIED* then

Compute *d's Eps*-neighborhood *neighborhood*;

a. **If** *neighborhood* < *MinPts*

Add d to NOISE set;

Else

Add d to current cluster (*ClusterId*);

b. **For** each data point *o* has a must-link constraints *ML*(*d*, *o*)

Add o to current cluster (*ClusterId*);

c. **For** each data point *p* in neighborhood set that does not violated cannot-link constraints

Add p to current cluster (*ClusterId*);

ClusterId = *ClusterId* +1;

End

4. Experimental Results

In this section, the datasets used is presented, followed by the evaluation metrics utilized, and the results of the study.

4.1. Datasets and Evaluation Metrics

In this section, the datasets utilized are presented. Table 1 outlined the used datasets. For our experiments, real datasets like Magic and Glass were selected from UCI repository. Hence, these datasets are labelled with instances, attributes, and cluster labels as described in Table 1.

Table 1. The Data Sets Used in the Experiments.

Dataset	# Instances	#Attributes	#Clusters
Glass	214	10	6
Ecoli	336	8	8
Segment	2310	19	7
Magic	19020	10	2

We utilized four clustering algorithms, i.e., C-DBSCAN [18], SSDBSCAN [19], AA-DBSCAN [4], and MPCK-mean, along with SSMD to compare their clustering performances. Clust ring performance was m asured in terms of the normalized mutual information (NMI) and Clus ering Accuracy (ACC). NMI and Clustering accuracy is measured using Equation 6 and 7.

$$NMI = \frac{I(X;Y)}{\big(H(X) + H(Y)\big)/2} \tag{6}$$

$$ACC = \frac{\sum_{i=1}^{n} I(\hat{t}_i = t_i)}{n} \tag{7}$$

4.2. Performance Analysis Based on NMI

In this sub-section, the performance evaluation based on NMI is presented. This evaluation result is with respect to the proposed algorithm in comparison to the four algorithms (SSDBSCAN, AA-DBSCAN, MPCKmeans, and C-DBSCAN). The goal is to find out if or not the proposed algorithm performs better with varied constraints on distinct datasets. Hence, this experiment will presents evidence that the proposed algorithm is more effective in comparison to the compared algorithms.

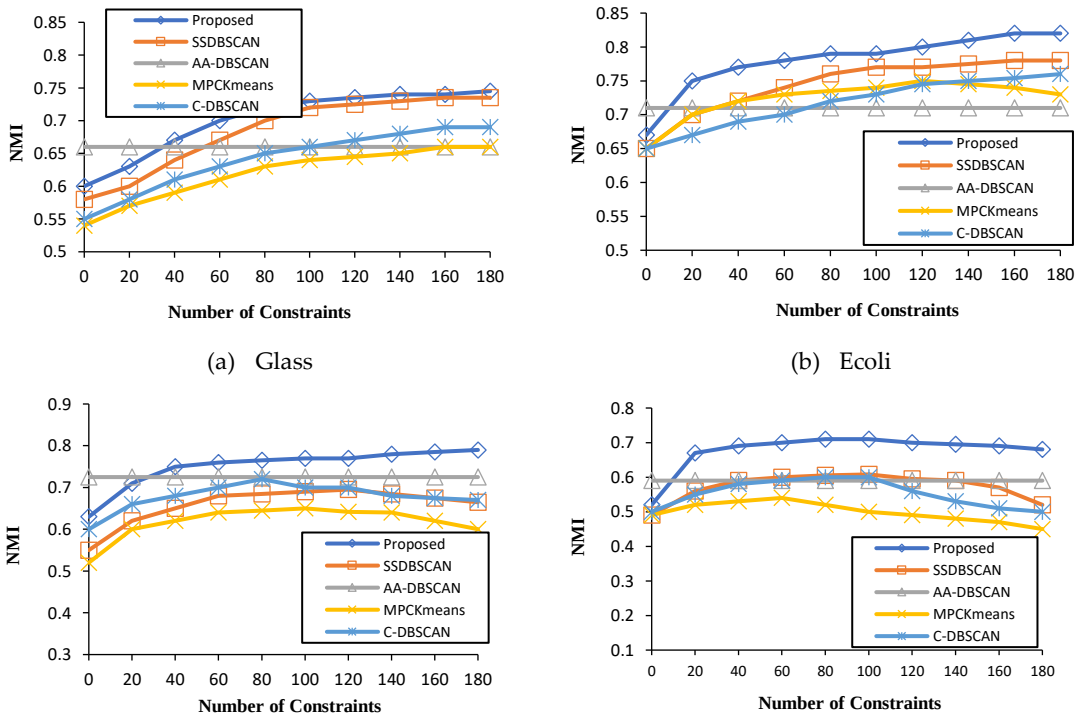


Figure 1. Comparison of NMI.

Figure 1 presents the performance results of the four different datasets (Glass, Ecoli, Segment, and Magic) as presented in Table 1. For each dataset, a number of constraints were utilized with respect to the NMI. The result demonstrates that the proposed algorithm performance better than the compared algorithms in general. For instance, in Figure 1.A, with 20 constraints the proposed algorithm performance superior with greater than 0.6 NMI, and with 80 constraints, the proposed algorithm achieve NMI of more than 0.7, respectively. Looking very carefully, one can see that SSDBSCAN is the second most

effective, followed by C-DBSCAN. Even though AA-DBSCAN has a higher performance in the early stages, ~~yet~~ the performance eventually drops. The reader can spot that the effectiveness of the proposed algorithm progressively decreases in the figure (Figure 1.A). However, ~~despite this,~~ the performance of the proposed algorithm surpass the four algorithms compared with on Glass dataset.

With respect to the Ecoli dataset, the proposed algorithm also has shown to be more effective, with 0.8 NMI with 80 constraints. This is same with the other two datasets, which are Segment and Magic. Considering the overall result in Figure 1, we can confidently conclude that with respect to NMI performance evaluation, the proposed algorithm outperformance all the four compared algorithms.

4.3. Performance Analysis Based on ACC

This section presents result based on the clustering performances of the proposed algorithm in comparison to compared algorithms on the datasets outlined in Table 1. Hence, the results of each algorithm are displayed in varied colors as shown in Figure 2. From Figure 2, one can see that in all datasets, the proposed algorithm outperform the compared algorithms ~~in~~ great margin. For instance, looking at Figure 2.A (experiment on Glass), with 20 constraints and above, the algorithm performance better. This performance is persistent to the other experiments on distinct datasets.

Considering the general clustering performance on all datasets in Figure 2, SSDBSCAN is the second most effective algorithm with respect to clustering accuracy on Glass, Ecoli, and Segment datasets, followed by AA-DBSCAN and C-DBSCAN. The main advantage of the proposed algorithm is that it can attain good performance with fewer constraints and progressively maintain the performance. However, we observed that other algorithms compared with required more constraints to get a good performance. Some algorithms performance also subside considerable.

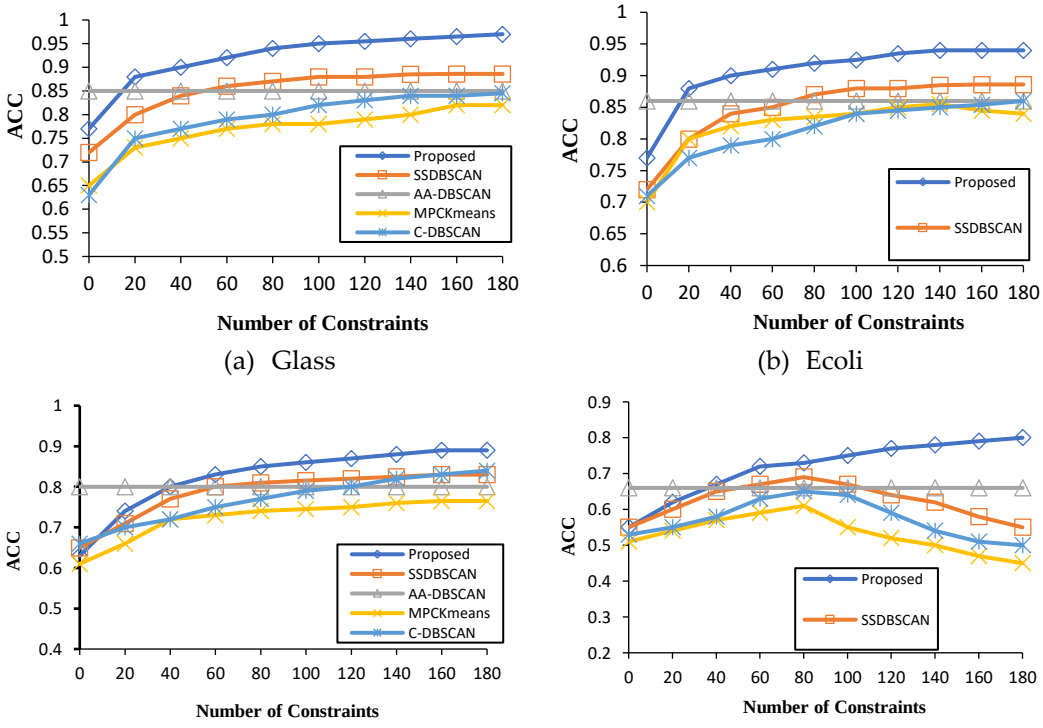
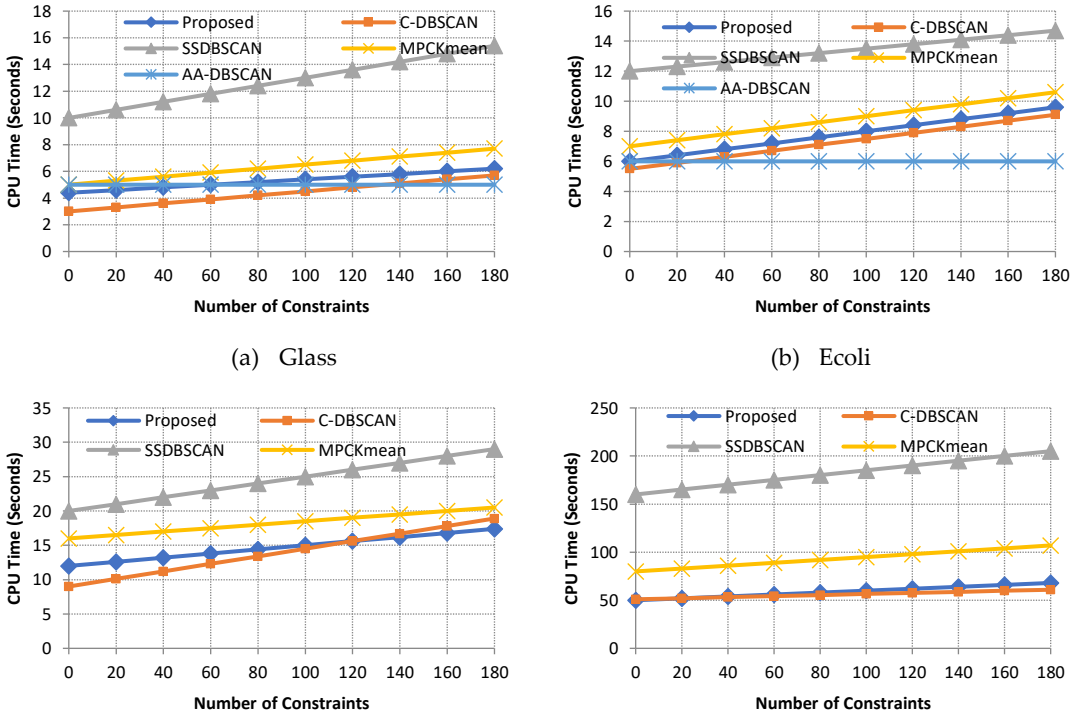


Figure 2. Comparison of the clustering accuracy.

4.4. Efficiency Analysis Based on Times

In this section, the clustering times for the proposed algorithm and the four compared algorithms is presented. Figure 3 give the results for the execution times suing distinct number of pairwise constraints on the four datasets. Achieving low executing time indicates that an algorithm has a better performance, and vice versa with a high execution time. From Figure 3.A, SSDBSCAN has the highest execution time with all pairwise constraints taking into consideration. We observed that the proposed algorithm on Glass dataset achieve relatively high execution time with low constraints, however, overall, the execution time drops and better performance was recorded for the proposed algorithm. Looking at all the experiments conducted on the other datasets, the reader can see that consistently the proposed algorithm performs better by achieving lower execution times on all the datasets in comparison with all the compared algorithms.



(a) Glass (b) Ecoli

(c) Segment (d) Magic

Figure 3. Comparison of the execution time.

5. Conclusions

In this paper, SSMD was proposed which is an active semi-supervised clustering method for clustering multi-density data and arbitrary shapes. The proposed algorithm partition dataset into distinct density level sets by examining the statistical characteristics of its variation with respect to density, and then expand the clusters using selected active pairwise constraints. We conducted experiments to assess the clustering performance and execution time on both artificial and real datasets of distinct dimensions and sizes. The experimental result revealed that SSMD attain better clustering performance in comaprision to the existing state-of-the-art. Furthermore, SSMD has significantly reduced the execution time in comparison to the compared algorithms in most scenarios.

Funding: "This research was funded by the University of Jeddah, Saudi Arabia, under grant No. (UJ-02-069-DR).

Acknowledgments: This work was funded by the University of Jeddah, Saudi Arabia, under grant No. (UJ-02-069-DR). The author, therefore, acknowledges with thanks the university technical and financial support.

References

1. M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: ordering points to identify the clustering structure," ACM Sigmod Rec., vol. 28, no. 2, pp. 49–60, 1999.
2. P. Liu, D. Zhou, and N. Wu, "VDBSCAN: varied density based spatial clustering of applications with noise," in 2007 International conference on service systems and service management, 2007, pp. 1–4.
3. S. Jahirabadkar and P. Kulkarni, "Algorithm to determine ε-distance parameter in density based clustering," Expert Syst. Appl., vol. 41, no. 6, pp. 2939–2946, 2014.
4. T. Huang, Y. Yu, K. Li, and W. Zeng, "Reckon the parameter of DBSCAN for multi-density data sets with constraints," in 2009 International Conference on Artificial Intelligence and Computational Intelligence, 2009, vol. 4, pp. 375–379.
5. L. Ertöz, M. Steinbach, and V. Kumar, "Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data," in Proceedings of the 2003 SIAM international conference on data mining, 2003, pp. 47–58.
6. Lulli, M. Dell'Amico, P. Michiardi, and L. Ricci, "NG-DBSCAN: scalable density-based clustering for arbitrary data," Proc. VLDB Endow., vol. 10, no. 3, pp. 157–168, 2016.
7. W. Atwa, and K. Li. "Constraint-based clustering algorithm for multi-density data and arbitrary shapes." In Industrial Conference on Data Mining, pp. 78-92. Springer, Cham, 2017.
8. M. Ester, H.P. Kriegel, J. Sander, and X. Xu, 'A density based algorithm for discovering clusters in large spatial databases with noise'. In Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining, pp. 226–231, (1996).
9. X. Liu, Q. Yang, and L. He, "A novel DBSCAN with entropy and probability for mixed data," Cluster Comput., vol. 20, no. 2, pp. 1313–1323, 2017.
10. J.-H. Kim, J.-H. Choi, K.-H. Yoo, and A. Nasridinov, "AA-DBSCAN: an approximate adaptive DBSCAN for finding clusters with varying densities," J. Supercomput., vol. 75, no. 1, pp. 142–169, 2019.
11. L. Zhang, Z. Xu, and F. Si, "GCMDDBSCAN: multi-density DBSCAN based on grid and contribution," in 2013 IEEE 11th International Conference on Dependable, Autonomic and Secure Computing, 2013, pp. 502–507.
12. M. A. Masud, J. Z. Huang, M. Zhong, and X. Fu, "Generate pairwise constraints from unlabeled data for semi-supervised clustering," Data Knowl. Eng., vol. 123, p. 101715, 2019.
13. W. Atwa, and K. Li. "Active query selection for constraint-based clustering algorithms." In International Conference on Database and Expert Systems Applications, pp. 438-445. Springer, Cham, 2014.
14. W. Atwa, and M. Emam. "Improving Semi-Supervised Clustering Algorithms with Active Query Selection." Advances in Systems Science and Applications 19, no. 4 (2019): 25-44.
15. W. Atwa, "A Supervised Feature Selection Method with Active Pairwise Constraints." In Proceedings of the 11th International Conference on Informatics & Systems (INFOS 2018). 2018.
16. J. Kim, W. Lee, J. J. Song, and S.-B. Lee, "Optimized combinatorial clustering for stochastic processes," Cluster Comput., vol. 20, no. 2, pp. 1135–1148, 2017.
17. C. Ruiz, M. Spiliopoulou and E. Menasalvas. C-DBSCAN: Density-Based Clustering with Constraints. In Proceedings of the International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, pages 216- 223, 2007.
18. L. Lelis and J. Sander. Semi-Supervised Density-Based Clustering. In Proceeding of the Ninth IEEE International Conference on Data Mining, 842-847, 2009.
19. W. Atwa, and K. Li. "Semi-supervised Clustering Method for Multi-density Data." In International Conference on Database Systems for Advanced Applications, pp. 313-319. Springer, Cham, 2015.
20. S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Batch mode active learning and its application to medical image classification," in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 417–424.
21. S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Semisupervised SVM batch mode active learning with applications to image retrieval," ACM Trans. Inf. Syst., vol. 27, no. 3, pp. 1–29, 2009.
22. S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," in Advances in neural information processing systems, 2010, pp. 892–900.
23. P. K. Mallapragada, R. Jin, and A. K. Jain, "Active query selection for semi-supervised clustering," in 2008 19th International Conference on Pattern Recognition, 2008, pp. 1–4.
24. S. Xiong, J. Azimi, and X. Z. Fern, "Active learning of constraints for semi-supervised clustering," IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 43–54, 2013.
25. H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," in Proceedings of the twenty-first international conference on Machine learning, 2004, p. 79.
26. V.-V. Vu, N. Labroche, and B. Bouchon-Meunier, "Active learning for semi-supervised k-means clustering," in 2010 22nd IEEE International Conference on Tools with Artificial Intelligence, 2010, vol. 1, pp. 12–15.
27. W. Zhao, Q. He, H. Ma, and Z. Shi, "Effective semi-supervised document clustering via active learning with instance-level constraints," Knowl. Inf. Syst., vol. 30, no. 3, pp. 569–587, 2012.

28. X. Wang and I. Davidson, "Active spectral clustering," in 2010 IEEE International Conference on Data Mining, 2010, pp. 561–568.
29. R. Huang, W. Lam, and Z. Zhang, "Active learning of constraints for semi-supervised text clustering," in Proceedings of the 2007 SIAM International Conference on Data Mining, 2007, pp. 113–124.
30. M. C. Fernandes, T. F. Covões, and A. L. V. Pereira, "Improving evolutionary constrained clustering using Active Learning," Knowledge-Based Syst., vol. 209, p. 106452, 2020.