Journal of Education for Pure Science- University of Thi-Qar
Vol.12, No.1 (March, 2022)
Website: jceps.utq.edu.iq Email: jceps@eps.utq.edu.iq

DOI: http://doi.org/10.32792/utq.jceps.12.01.06

Improve Clustering -Based Graph Algorithms Using (MST) Minimal Spanning Trees

Shaima M.Aldkeli College of Science University of Basrah

Kahdhim H.Alibraheemi college of Science University of Basrah

Received 13/1/2022 Accepted 9/3/2022 Published 30/3/2022

Abstract:

Clustering is a significant data mining method that is used in a variety of disciplines. Clustering attempts to divide a big set of data into smaller groups that are more similar within the same group, but distinct from other groups, each subgroup referred to as a "cluster". This paper introduces the concept of clustering, the most important clustering approaches, and the application of graph-based clustering utilizing one of the graph algorithms, the Minimum Spanning Tree (MST) algorithm, which groups related nodes into clusters. The method has been evaluated on five various data sets utilizing cluster validation measures (Adjusted Rand Index, v-measure_ scores), and its performance on all of these data sets has been demonstrated when compared to other algorithms.

Keywords: *Clustering, Minimum Spanning Trees, Graph,* Clustering techniques

1. Introduction

Due to progress in information technology, people are faced daily with a huge amount of information, which is represented as data, and to find or extract useful information from that data, data mining is used, and for the purpose of analyzing and managing data, used of the effective data mining techniques for the purpose of clustering or classification [1].

Cluster analysis is a fundamental concept in data mining, aims to split a large set of data into a smaller set based on similarity and which detect the intrinsic structure of that data. Clustering is an unsupervised learning method in which data is collected without prior knowledge of the category label [2].

Clustering is a technology of major importance and has applications in many fields, including machine learning, data mining, image processing, pattern recognition, biological [3].There are several techniques and methods for clustering In this paper, will use Clustering –based graph algorithms ,where there are several algorithms for the purpose of clustering based on graph ,in our study we using clustering -based a Minimum Spanning Tree (MST), where a (MST) minimum spanning tree distinguished in finding on detecting clusters with irregular boundaries [4].

The paper is structured as follows: Section 2 discusses related research, while Section 3 clustering techniques, in section 4 Evaluation metrics, The clustering algorithm is discussed in Section 5. The experimental results are shown in Section 6. Conclusions are included in Section 7.

2. Related Works:

Spanning tree is, connected undirected non-cyclic sub-graph for a graph G, which involves all the 'vertices' of G. The minim spanning tree (MST) of is the weighing minimal spanning tree of that graph. [5,6]. Clustering based a minimum spanning tree was proposed for the first by C. T. Zahn,1971 [7].
A review of some researches related to clustering techniques based minimum spanning tree.

Table (2.1) Researches related to clustering based on minimum spanning tree

Researchers	*Year*	*Method*	*Reference*
Xu Y ,et al	*2001*	*describe clustering gene expression data multidimensional using a minimal spanning tree .three objective functions and the corresponding cluster algorithm for computing k-partitions*	*[8]*
O. Grygorash et al	*2006*	*propose, two algorithms for clustering using the minim spanning tree. In the first algorithm, a value of K is given to produce the required clusters, and the second algorithm without giving a value of K , The algorithm showed good results compared to the (k-mean ,EM) algorithm through its ability for finding clusters with irregular boundaries..*	*[4]*
Peter. S. John and S. P. Victor	*2010*	*proposed, two approaches to clustering based MST, the first method used the divisional approach, and the second method produced dendrogram (agglomerative).*	*[9]*
A. C. M¨uller et al	*2012*	*proposed a theoretical clustering algorithm to find cluster using optimization to estimate the information mutual between data distribution and cluster assignment. It proposes an efficient algorithm with low runtime complexity. It shows good performance on many datasets*	*[10]*
Wang et al	*2013*	*proposed an efficient clustering method using a minimal spanning tree. Optimizing a MST- based clustering ,by removal density -based outliers, they applied the method to high-dimensional datasets and low-dimensional dataset, and it showed its effectiveness*	*[11]*
Zhong et al	*2015*	*presented an algorithm consisting two phase, the first phase using 'divide and conquer' model and in the second phase refinement. It showed good performance in clustering .*	*[12]*

Lv, Xiao-bo et al	*2018*	*proposes a clustering method based on the minimum spanning tree, by of the group center initialization method , based on the geodesic distance , for find inconsistent edges, where good effectiveness is obtained on the used dataset .*	*[13]*
Li, Jia et al	*2019*	*present two methods, the first is to calculate the' scaled-MST with scaled distance' to detection the lengthy edges in the dataset of different densities. The second method work by merging the MST structure and 'inconsistent edges' find ,in only one step , algorithm apply on image segmentation*	*[14]*

3. Clustering techniques:

This section includes, presenting for the most important clustering techniques widely used in the literature.

3.1 Partitional *methods*:

where the data set is divided into k of clusters through some objective functions, finding the distance or difference between a point and the group model. The figure (2.1) shows the Partitional method. An algorithm k means is defined as the oldest and most prevalent algorithm in the partitional method. The most fundamental step in the (k- means) algorithm is defining the number of clusters, i.e. defining the centers of clusters in prior by the user, one center for each group. Although this algorithm is widely used, it suffers from noise sensitivity as well as its inability to deal with data sets of arbitrary sizes, shapes, and densities [2,15], below are the algorithm steps.

K-Means Algorithm (1)

1-Choose the number of clusters, k.

2-Randomly generate k clusters and determine the cluster centers, or directly generate k random points as cluster centers.

3-Assign each point to the nearest cluster center.

4-Recompute the new cluster centers.

5-Repeat the two previous steps until some convergence criterion is met (usually that the assignment hasn't changed).

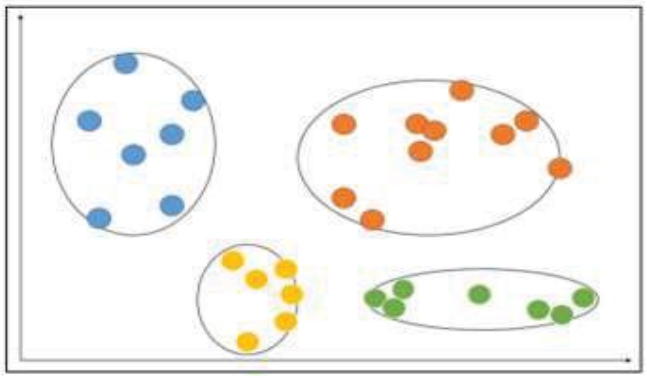

Figure (3.1) Partitional clustering approaches [16]

3. 2 Hierarchical clustering methods:

In contrast to partial methods, hierarchical algorithms find groups are two ways: 'agglomerative' and 'divisive', where the resulting groups are in the form of 'dendrogram '. The algorithms stop merging or splitting when the required number of clusters, the figure shows the hierarchical method, these algorithms suffer from not being able to make corrections when splitting or merging faulty. [15,17].

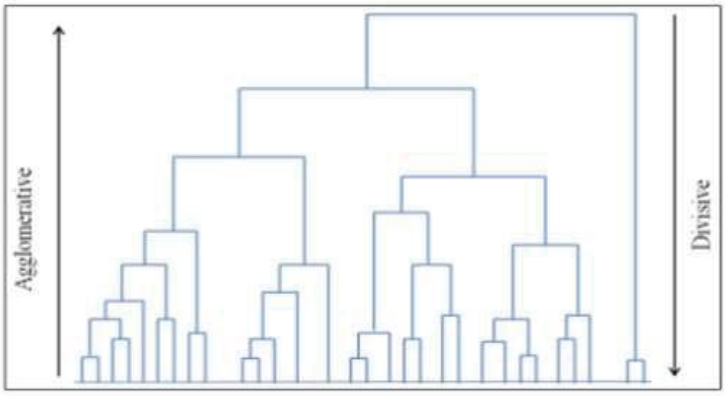

Figure (3.2) Hierarchical clustering methods [17]

3. 3 Clustering Methods Based Density:

In this type of algorithm, contiguous points are grouped into clustering depending on the density conditions. Figure (2.3) show density- based clustering. These algorithms have difficulty specifying input parameters, these algorithms detect randomly shaped clustering [15,6].

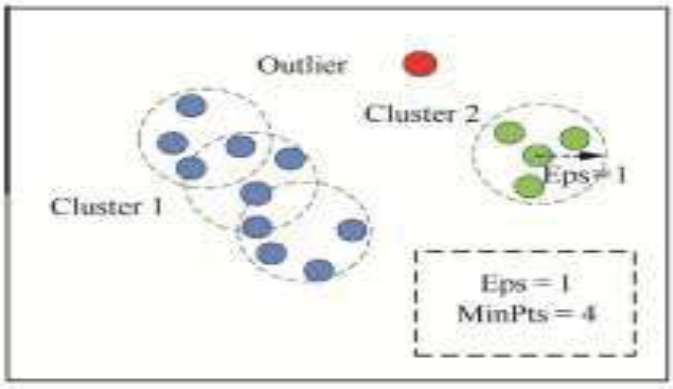

Figure (3.3) Density based clustering [19]

3.4 Clustering –based graph algorithms:

There are several algorithms for the purpose of clustering based on graph, in our study we using clustering based on a minimal spanning tree, where the spanning tree is distinguished in finding clusters of different shapes and sizes. A graph is a structure consists of 'nodes' and 'edges', where these edges connect the nodes. In graph-based grouping, similar vertices are grouped into groups [20,21].

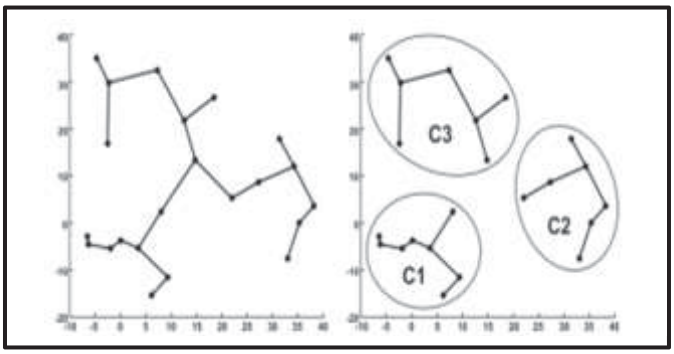

Figure (3.4) Clustering –based graph algorithms [21]

3.4.1 Minimum Spanning Trees algorithm:

For a set of data X where X= $\{x_1, x_2, xn\}$, let a undirection weighted graph be constructed where the complete graph is expressed G= (V, E), where V represents a set of data points from the dataset, either E it represents the edges, which is the the distance between the vertices, let it be the distance between x_1, x_2 is distance calculated using one of the distance measures. The (MST) minimum spanning tree, is a spanning tree is a non-cyclic sub-Graph (G′) of a graph(G) that connects all nodes (V) with the fewest edges(E′) [21,22]. The minimum spanning tree is constructed using either the Kruskal [5] or Prim [6] algorithm.

3.4.1.1 Kruskal's Algorithm

Kruskal's algorithm is an another greedy algorithm to construct the minimal spanning tree. This algorithm starts with a forest (initially each node of the graph represents a tree) and iteratively adds the edge with the lowest cost to the forest connecting trees such a way, that circles in the forest are not enabled. The formal algorithm is given in Algorithm (2)

Kruskal's algorithm (2)

1-Create a forest F in such a way that each vertex (V) of the graph G denotes a separate tree. Let S = E
2-Select an edge e from S with the minimum weight. Set S = S – {e}.
3-If the selected edge e connects two separated trees, then add it to the forest.
4-Step 4 If F is not yet a spanning tree and S ≠ { } go back to Step 2.

4. Evaluation metrics:

Cluster Evaluation metrics are classified into external and internal indexes: external metrics evaluate clusters according to the knowledge of the correct class labels, but if class labels are not available, internal indexes are used as they depend in estimating the quality on the data alone **[23,24].** Examples of these metrics are *(Adjusted Rand index, V-measure).*

4.1 Adjusted Rand index:

This metrics is considered one of the external measures and is used to verify the validity of the clusters because it is a measure to show the extent to which the correct group matches what is expected in the clustering. The value The adjusted Rand index, its value, is 0.0 when the labeling are random, for a number of groups and samples, and 1.0, when the grouping is the identical, and it is expressed in the mathematical formula below:

$$ARI = \frac{RI - E[RI]}{max(RI) - E[RI]} \quad \ldots\ldots (4.5)$$

Where (RI) is Rand index

Homogeneity, completeness and V-measure:

Homogeneity: Each group contains objects of one class Just, completeness: All objects belonging to a certain class are assigned to the same group, The V-measure harmonic mean to completenes and homogeneity. These measures range in value between 0 and 1, If the value is increased to 1, it is better.

5. Proposed methodology:

The steps for a methodology (CBMST) is explained on the following diagram.

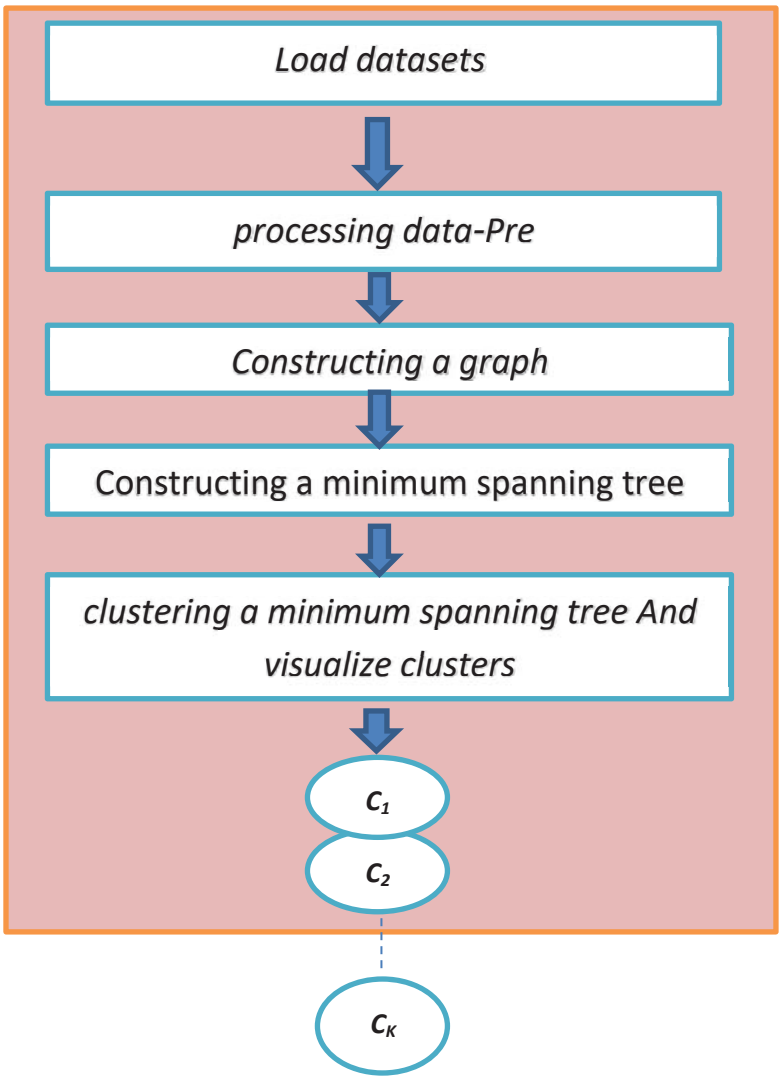

Figure (5.1) Flowchart Method

1- Initially, the data set input, let it be, "D= {d_1, d_2, dn}", where "n ", is number of objects in a dataset with p- dimensions.

2- Pre-processing step is performed for each the dataset, which is considered an important stage before apply any algorithm. Pre-processing is used for the purpose of easier the process of data clustering, data quality has a significant effect on the processing process. Pre-processing consists of several steps that are used based to the type of data, including (outlier detection, transformation, normalization, dimensional reduction and feature selection).

3- The algorithm starts by finding the nearest neighbors between the data points based on the distances between the points and this is done according to the (nearest neighbors) algorithm that finds the nearest neighbors.The nearest neighbor is find by implementing algorithm (KNN Graph),value (k) represents the number of neighbors where (k<n), where the outputs of the previous step are the inputs to the algorithm, which finds the distance between the data points and the specified number of neighbors, the Euclidean distance is applied, where the algorithm works to find the distances according to the specified search algorithm ,not done specify the type of search algorithm as it works to find the best distances between data points, after finding the nearest neighbors for all the input data, undirection weighted graph is produced, where the vertices of the graph are the data points, and the edges between the vertices represent the weighted (Euclidean distance). weighted graph is built, where the vertices represent the number of points and the edges represent the distance values obtained from the previous step.

4- In this step, the minimum spanning tree is created, the minimum spanning tree algorithm is created using the (kruskul) algorithm, where this algorithm works to find (spanning tree) with the least weight, this algorithm starts with a forest (initially each node of the graph represents a tree) and iteratively adds the edge with the lowest cost to the forest connecting trees such a way, that circles in the forest are not enabled.

5- After (minimum spanning tree) is built, we search for the inconsistent edge among the tree edges, finding the inconsistent edge and then removing it. The edge deletion is inconsistent will divide the tree into a group of connected components, as each component It is a cluster, the algorithm continues to find the inconsistent edges until the desired number of clusters is found, where the output of the algorithm is the label for each cluster.

6. Experiment Results:

The algorithm is tested on a set of 2-dimensional data, where these selected data are containing of various shapes, sizes, densities. Table (6.1) datasets details shows. The three data sets (DS1, DS2, DS3) are from [25], and the datasets (DS4, DS5) are from [26].

Table (6.1) Details of the 2-dimensional datasets

Name of the data set	*number of data points(n)*	*Number of clusters (k)*	*Description*
DS1	*200*	*4*	*Blob(Gaussian clusters)*
DS2	*800*	*2*	*half-moon*
DS3	*500*	*2*	*nested circles*
DS4	*788*	*7*	*aggregation*
DS5	*312*	*3*	*spiral*

Where the data set on which the algorithm was examined, each data set describes a various type for clustering problems. The algorithm was implemented on the five datasets shown in the figure (6.1), where we note that it is able to separate clusters well with each of the datasets, and as shown in Figure (6.2)
Cluster evaluation are performed using two measures **Adjusted Rand Index** (ARI) and (**V-measure**), which was mentioned in the section 4, each of the five datasets was evaluated with metrics, and shows good performance.
The figure(6.2) shows the able of the minimum spanning tree algorithm in separating clusters well with the five datasets compared to the algorithm (K means) with the same data set and under the same conditions, as figure (6.3).

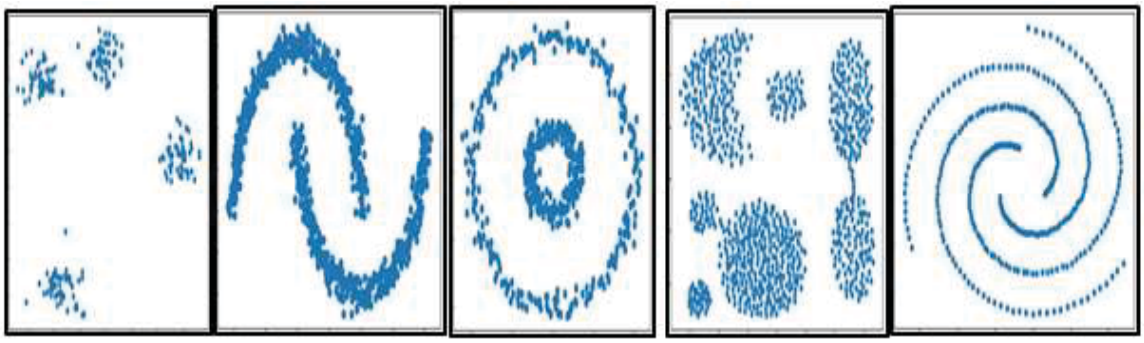

Figure (6.1) Five synthetic datasets

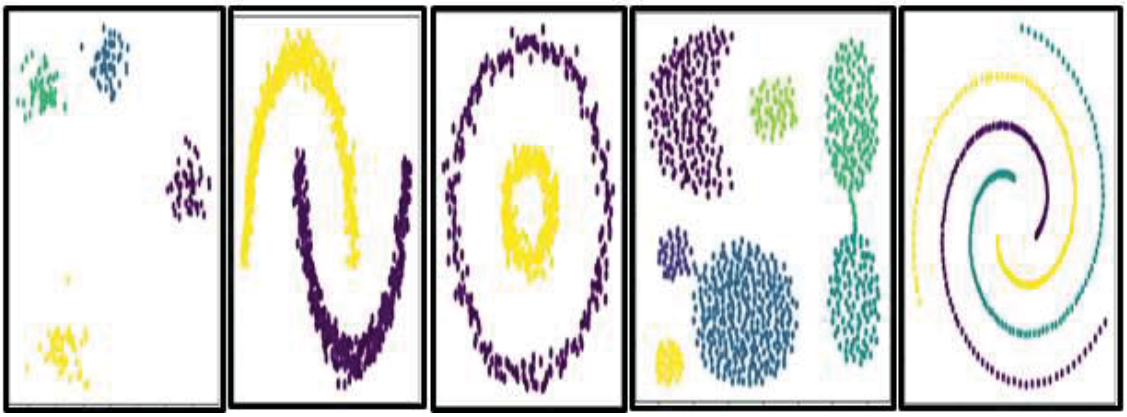

Figure (6.2) Five datasets with CBMST

Figure (6.3) Datasets Five with k-means algorithm

The algorithm was compared with the (kmeans) algorithm and the (agglomerative) algorithm for the same data set using evaluation metrics, as shown in Figures (6.4) and (6.5).

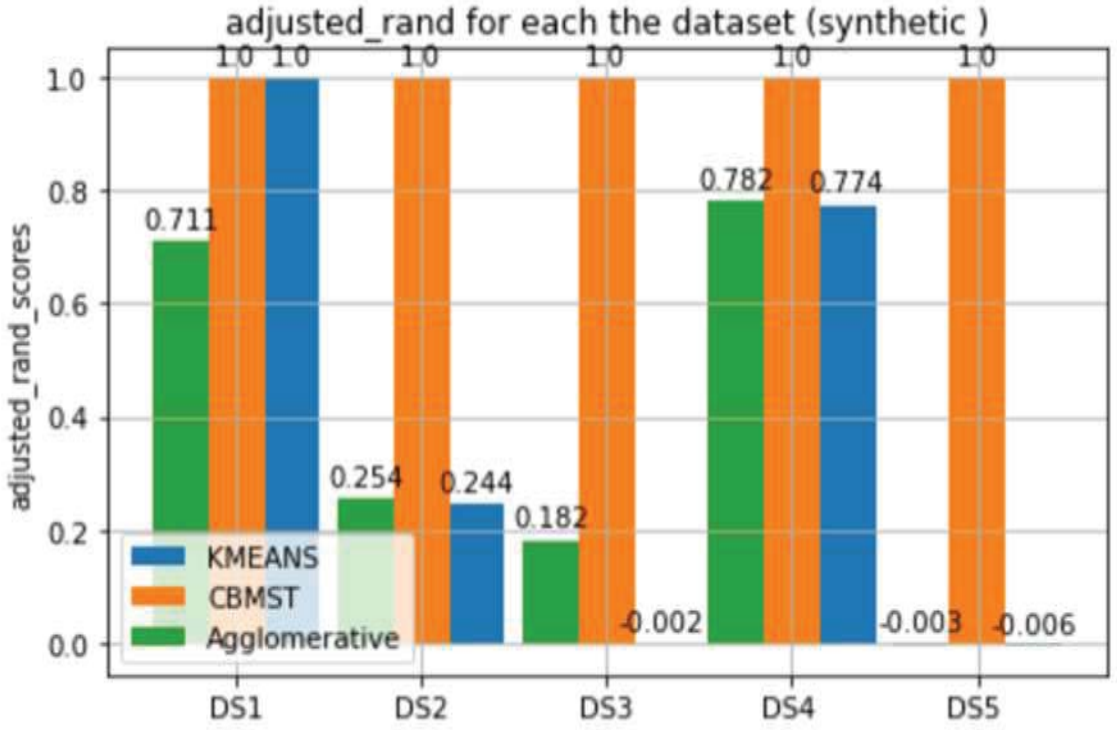

Figure (6.4) Adjusted Rand Index by all synthetic datasets

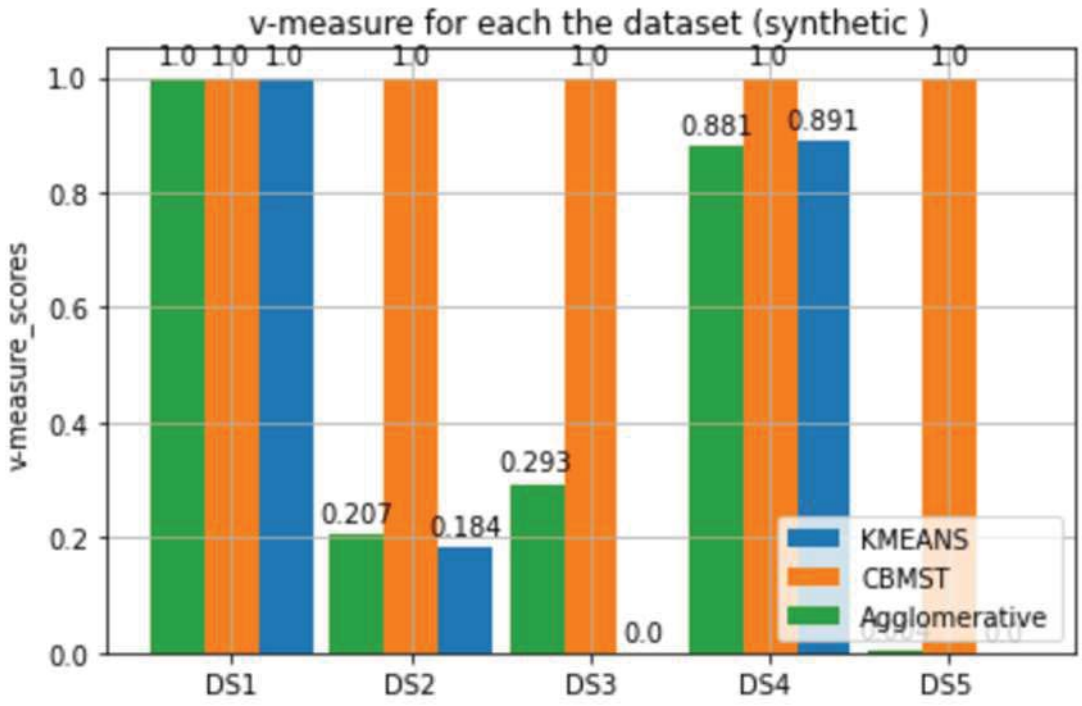

Figure (6.5) v-measure_ scores by all synthetic datasets

7. Conclusions:

In our study, the clustering method was discussed using one of the graph algorithms, the minimal spanning tree (MST) algorithm, which showed its ability to find clusters in the data set of different shapes

and sizes, and the results also showed that the algorithm works well in the two-dimensional dataset. In addition to being more efficient compared to the (Kmeans, Agglomerative) algorithms and based on the results of cluster validation metrics in Figures (6.4) and (6.5), our method can be extended to work on a high-dimensional data set or can be combined with one of the other clustering algorithms to find clusters in future work.

References:

[1] Rui Xu and D. Wunsch, "Survey of clustering algorithms," IEEE Transactions on Neural Networks, vol. 16, no. 3, pp. 645-678, May 2005.
[2] K. P. Sinaga and M. Yang, "Unsupervised K-Means Clustering Algorithm," IEEE Access, vol. 8, pp. 80716-80727, 2020.
[3] Sathiyakumari, K. et al. "A Survey on Various Approaches in Document Clustering." (2011).
[4] O. Grygorash, Y. Zhou and Z. Jorgensen, "Minimum Spanning Tree Based Clustering Algorithms," 2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06), 2006, pp. 73-81.
[5] Kruskal, Joseph B. "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem." Proceedings of the American Mathematical Society, vol. 7, no. 1, American Mathematical Society, 1956, pp. 48–50.
[6] R. C. Prim, "Shortest connection networks and some generalizations," in The Bell System Technical Journal, vol. 36, no. 6, pp. 1389-1401, Nov. 1957.
[7] C. T. Zahn, "Graph-theoretical methods for detecting and describing gestalt clusters," IEEE Trans. Computers, Vol. 20, No. 1, pp. 68–86, January 1971.
[8] Xu Y, Olman V, Xu D. Minimum spanning trees for gene expression data clustering. Genome Inform. 2001; 12:24-33. PMID: 11791221.
[9] Peter, S. John, and S. P. Victor. "A Novel Algorithm for Informative Meta Similarity Clusters Using Minimum Spanning Tree." arXiv preprint arXiv:1005.4585 (2010).
[10] Müller A.C., Nowozin S., Lampert C.H. (2012)," Information Theoretic Clustering Using Minimum Spanning Trees". In: Pinz A., Pock T., Bischof H., Leberl F. (eds) Pattern Recognition. DAGM/OAGM 2012. Lecture Notes in Computer Science, vol 7476. Springer, Berlin, Heidelberg.
[11] Wang, Xiaochun et al. "Enhancing minimum spanning tree-based clustering by removing density-based outliers." Digit. Signal Process. 23 (2013): 1523-1538.
[12] Zhong, C., Malinen, M., Miao, D., Franti, P. (2015). ¨ A fast minimum spanning tree algorithm based on k-means. Information Sciences.
[13] Lv, Xiao-bo et al. "CciMST: A Clustering Algorithm Based on Minimum Spanning Tree and Cluster Centers." Mathematical Problems in Engineering (2018): n. pag.
[14] Li, Jia et al. "A scaled-MST-based clustering algorithm and application on image segmentation." Journal of Intelligent Information Systems 54 (2019): 501-525.
[15] Sisodia, Deepti et al. "Clustering Techniques: A Brief Survey of Different Clustering Algorithms." (2012).
[16] Faizan, Muhammad et al. "Applications of Clustering Techniques in Data Mining: A Comparative Study." International Journal of Advanced Computer Science and Applications 11 (2020): n. pag
[17] Saxena, Amit, et al. "A review of clustering techniques and developments." Neurocomputing 267 (2017): 664-681
[18] Madhulatha, T. Soni. "An overview on clustering methods." arXiv preprint arXiv:1205.1117 (2012).

[19] Wang, Rixin, et al. "Fault detection of flywheel system based on clustering and principal component analysis." Chinese Journal of Aeronautics 28.6 (2015): 1676-1688.
[20] P. Praveen, B. Rama and T. Sampath Kumar, "An efficient clustering algorithm of minimum Spanning Tree," Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2017, pp. 131-135.
[21] J. G. M. Esgario and R. A. Krohling, "Clustering with Minimum Spanning Tree using TOPSIS with Multi-Criteria Information," IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2018, pp. 1-7.
[22] Vathy-Fogarassy, Ágnes, and Janos Abonyi. Graph-based Clustering and Data Visualization Algorithms. Springer, 2013.
[23] Handl, Julia, Joshua Knowles, and Douglas B. Kell. "Computational cluster validation in post-genomic data analysis." Bioinformatics 21.15 (2005): 3201-3212.
[24] Al-Mhairat, Muthana et al. (2019). Performance Evaluation of clustering Algorthims.
[25] Pedregosa, Fabian et al. “Scikit-learn: Machine Learning in Python.” J. Mach. Learn. Res. 12 (2011): 2825-2830.
[26] http://cs.joensuu.fi/sipu/datasets/