A Comparative Study of Clustering Algorithms

Manoj Kr. Gupta
RDIAS, Delhi, INDIA.
USIC&T, GGSIPU, New Delhi, INDIA
Email Id: manojkgupta5@gmail.com

Pravin Chandra
USIC&T, GGSIPU, Delhi, INDIA
Email Id: chandra.pravin@gmail.com

***Abstract*—In present era, data analysis plays vital role in various domains. Data clustering is a data analysis technique used for grouping of data objects based on unsupervised learning. Many clustering algorithms have been proposed in the literature. Each algorithm possesses some strengths and weaknesses. Therefore, a set of clustering algorithms are appropriate for one set of application area while another set of clustering algorithm are suitable for another set of application areas. In this paper, popular traditional algorithms are discussed. A comprehensive comparative study of different clustering algorithms is presented in this paper. These clustering algorithms are compared in detail based on various parameters used in these methods.**

Keywords—Clustering Algorithms, Cluster Analysis, Comparative Study, Survey, Data Mining.

I. Introduction

Clustering is a vital exploratory multivariate data analysis method. In clustering, data objects are partitioned into clusters based on distance / dissimilarity among data objects. Data objects which are like or near to each other are placed within the same cluster while unlike or far off data objects are placed in another cluster. Like classification, clustering is also classifying the data objects but unlike the classification, the class labels are unknown because clustering is based on unsupervised learning [16]. The clusters are defined based on the study of the behaviour or characteristics of the data objects by the domain experts.

The clustering algorithms must have following properties:

a) Data objects within the cluster must be like or near to each other as much as possible.
b) Data objects belong to different clusters must be dissimilar or far off to each other as much as possible.
c) The distance / similarity measure must have some practical ability and be clear.

Clustering is also extensively used in many application domains i.e. statistics, image segmentation, image pattern recognition, object recognition, information retrieval, bioinformatics, etc. [21].

Distance and similarity is the foundation for formulation of clusters by the clustering algorithms. Distance measures are preferred for quantitative data whereas similarity measures are preferred for qualitative data [50]. Different distance metrics and similarity measures have been proposed by the many researchers which can be useful for clustering. Though, Euclidean, Manhattan and Mahalanobis metrics are widely used in different clustering algorithms.

The clustering process is presented in section 2. Comparison among various clustering algorithms is systematically analyzed and presented in section 3. Finally, conclusion is drawn in section 4.

II. Clustering Process

A typical data clustering process includes several steps which are presented in *Fig. 1*. In the following subsections, these steps are described [15, 21]:

A. Feature selection

It is a data preprocessing step based on dimensionality reduction. In this step, the appropriate features or characteristics of data are selected. Irrelevant features will leads to the unnecessary complexity in the clustering process.

B. Clustering algorithm

For clustering, either an appropriate existing cluster algorithm is chosen or a clustering algorithm is designed as per the required clustering goal. The chosen algorithms will be executed to find out the clusters within the data set. The appropriate clustering algorithm will generate better results as compared to the other algorithms.

C. Validation of the results

The identified clusters need to be validated as to know whether the clusters are identified appropriately or not. There are three categories of cluster validation techniques: (a) internal (b) external and (c) relative.

D. Interpretation of the results

After validation of the clusters, the clusters needs to analyzed and interpreted to describe the each cluster. The interpretation is to draw the right conclusion.

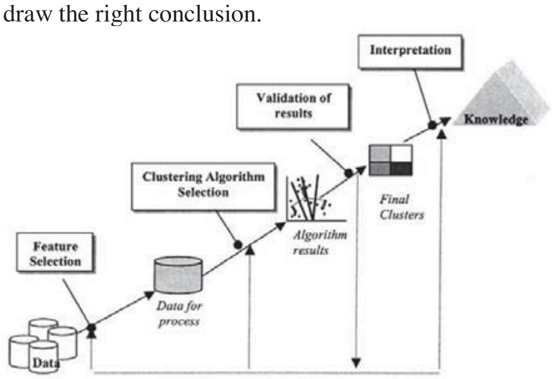

Fig. 1. Steps of a Typical Clustering Process

978-93-80544-34-2$31.00 ©2019 IEEE

III. Comparative Study of Clustering Algorithms

A lot of data clustering algorithms has been proposed by the many researchers so far. Some of these algorithms are based on one method / approach while other algorithms are based on another method / approach. Based on the similarity of method / approach followed by the algorithms, the clustering algorithms can be generally classified into the following categories [21, 16, 11, 45]:

A. Hierarchical Method

A hierarchical decomposition of the sets of data objects is identified and created by hierarchical method. Based on the order of hierarchical decomposition, these methods can be classified as either agglomerative or divisive. SLINK [40], CLINK [18], BIRCH [53], CURE [13], ROCK [14] algorithms follow agglomerative approach whereas DIANA [27] and DISMEA [43] algorithms follow divisive approach. Hierarchical methods can be based on either distance or density or continuity. In these methods, if merge or split is done then it can never be undone. In other words, these cannot correct erroneous decisions. The comparison of various popular clustering algorithms based on hierarchical method is presented in *Table I.* Some more hierarchical algorithms can be found in [20, 41, 46, 50, 58, 59, 70, 71, 72, 73].

B. Partition-based Method

A partition-based method constructs predetermined k partitions (or clusters) of a data set of n objects, where $k \leq n$. k-Means [28], k-Medoids [30], PAM [26], CLARA [25] and CLARANS [29] algorithms are based on partitioning methods. These methods are effective for data sets up to medium size. But, these can be extended for the large data sets as well. The comparison of various popular clustering algorithms based on partitioned-based method is presented in *Table II*. Some more partition-based algorithms can be found in [20, 26, 33, 38, 41, 47, 50, 55, 56, 57, 70, 71, 72, 73].

C. Density-based Method

In these methods, objects are clustered based on the concept of density instead of distance. Hence, arbitrary-shaped clusters can also be formed by these methods. The idea behind these methods is to keep on increasing a given cluster provided that the density in the neighborhood exceeds some threshold. These may also filter out outliers. DBSCAN [9], OPTICS [2], DENCLUE [17] and RDBC [44] algorithms are the examples of density-based methods. The comparison of various popular clustering algorithms based on density-based method is presented in *Table III*. Some more density-based algorithms can be found in [20, 22, 34, 41, 50, 63, 64, 65, 66, 70, 71, 72, 73].

D. Grid-based Method

These methods quantize the object space into a grid structure. These methods are fast because of independence from number of data objects, yet reliant on grid size in each dimension in the quantized tree. STING [48], CLIQUE [1], OptiGrid [18], GRIDCLUS [36], GDILC [54], WaveCluster [39] are grid-based methods. The comparison of various popular clustering algorithms based on grid-based method is presented in *Table IV*. Some more grid-based algorithms can be found in [20, 41, 42, 50, 67, 68, 69, 70, 71, 72, 73].

E. Fuzzy-based Method

Unlike other crisp methods, a fuzzy-based method assigns the each data object to all the clusters with definite degree of membership. Mostly partitioned-based methods are extended for fuzzy clusters. Fuzzy k-means [3], FCM [4], FCS [7], MM [51], MEC [35] are fuzzy-based clustering algorithms. The comparison of various popular clustering algorithms based on fuzzy-based method is presented in *Table V*. Some more fuzzy-based algorithms can be found in [20, 41, 50, 60, 61, 62, 70, 71, 72, 73].

F. Modern Clustering Methods

Apart from the above-mentioned categories of the clustering algorithms, the modern clustering algorithms are also classified into some more categories as [16, 50, 70, 72, 73] (a) Kernel-based algorithms, (b) Ensemble-based algorithms, (c) nature-inspired algorithms, (d) graph-based algorithms, (e) model-based algorithms, (f) Quantum Theory-based algorithms, etc.

IV. Summary and Conclusion

The paper started with the introduction of clustering followed by typical clustering process. A comprehensive comparative study of traditional clustering algorithms is presented in Table I through Table V. The main objective of this paper is to present the main idea of the common clustering algorithms and compare them based on several parameters. Only some common popular algorithms are discussed in this paper due to difficult to study and present all proposed algorithms because of the availability of a large number of clustering algorithms in a large number of sources. However, a systematic and clear view of the important algorithms is presented in this paper.

References

[1] Agrawal R, Gehrke J, Gunopulos D, Raghavan P (1998) Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings 1998 ACM sigmod international conference on management of data, vol 27, pp 94–105

[2] Ankerst M, Breunig M, Kriegel H, Sander J (1999) OPTICS: ordering points to identify the clustering structure. In: Proceedings on 1999 ACMSIGMOD international conference on management of data, vol 28, pp 49–60

[3] Bezdek, J. (1974). Fuzzy Mathematics in Pattern Classification. PhD thesis, Cornell University, Ithaca, NY.

[4] Bezdek J, Ehrlich R, Full W (1984) FCM: the fuzzy c-means clustering algorithm. Comput Geosci 10:191–203

[5] Bobrowski, L. and Bezdek, J. (1991). c-means clustering with the l1 and l∞ norms. IEEE Transactions on Systems, Man and Cybernetics, 21(3):545–554.

[6] Comaniciu D, Meer P (2002) Mean shift: a robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell 24:603–61

[7] Dave R, Bhaswan K (1992) Adaptive fuzzy c-shells clustering and detection of ellipses. IEEE Trans Neural Netw 3:643–662

[8] Defays, D. (1977). An efficient algorithm for a complete link method. The Computer Journal, 20(4):364–366.

[9] Ester M, Kriegel H, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings

of the second ACM SIGKDD international conference on knowledge discovery and data mining, pp 226–231

[10] Faber, V. (1994). Clustering and the continuous k-means algorithm. Los Alamos Science, 22:138–144.

[11] Gan, G., Ma, C., and Wu, J., (2007) 'Data Clustering: Theory, Algorithms, and Applications', American Statistical Association and the Society for Industrial and Applied Mathematics.

[12] Gower,J. and Ross,G. (1969). Minimum spanning trees and single linkage cluster analysis. Applied Statistics, 18(1):54–64.

[13] Guha, S., Rastogi, R., and Shim, K. (1998). CURE: An efficient clustering algorithm for large databases. In Proceedings of the 1998 ACM SIGMOD international conference on management of data, pages 73–84. NewYork: ACM Press.

[14] Guha S, Rastogi R, Shim K (1999) ROCK: a robust clustering algorithm for categorical attributes. In: Proceedings of the 15th international conference on data engineering, pp 512-521

[15] Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001), On Clustering Validation Techniques, Journal of Intelligent Information Systems 17(2-3), DOI: 10.1023/A:1012801612483

[16] Han J, Kamber M, Pei J (2012) 'Data mining concepts and techniques' Elsevier, 3rd Edition

[17] Hinneburg, A. and Keim, D.A. (1998), An efficient approach to clustering in large multimedia databases with noise, in: KDD, Vol. 98, pp. 58-65.

[18] Hinneburg, A. and Keim, D.A. (1999). Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. In Proceedings of the 25th international conference on very large databases (VLDB'99),pages506–517.SanFrancisco: Morgan Kaufmann.

[19] Huang, Z. and Ng, M. (1999). A fuzzy k-modes algorithm for clustering categorical data. IEEE Transactions on Fuzzy Systems, 7(4):446–452.

[20] Jain, A. K. and Dubes, R. C. (1988) 'Algorithms for Clustering Data'. Prentice Hall, Englewood Cliffs, NJ.

[21] Jain, A.K., Murty, M.N. and Flynn, P.J. (1999) 'Data clustering: a review' ACM Comput. Surv. 31, 3, 60 pages.

[22] Jin, H., Wang, S., Zhou, Q., and Li, Y. (2014) 'An improved method for density-based clustering', Int. J. of Data Mining, Modelling and Management, Vol. 6 No. 4

[23] Karypis G, Han E, Kumar V (1999) Chameleon: hierarchical clustering using dynamic modeling. Computer 32:68–75

[24] Kaufman L, Rousseeuw P (1990) Partitioning around medoids (program pam). Finding groups in data: an introduction to cluster analysis. Wiley, Hoboken

[25] Kaufman L, Rousseeuw P (2008) Finding groups in data: an introduction to cluster analysis, vol 344. Wiley, Hoboken. doi:10.1002/9780470316801

[26] Kaufman, L. and Rousseeuw, P.(1990). Finding Groups in Data—An Introduction to Cluster Analysis. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons, Inc.

[27] Macnaughton-Smith, P., Williams, W. T. , Dale, M. B., Mockett L. G. (1964), Dissimilarity Analysis: a new Technique of Hierarchical Sub-division, Nature, volume 202, pages 1034–1035.

[28] MacQueen J (1967) Some methods for classification and analysis of multivariate observations. Proc Fifth Berkeley Symp Math Stat Probab 1:281–297

[29] Ng R, Han J (2002) CLARANS: a method for clustering objects for spatial data mining. IEEE Trans Knowl Data Eng 14:1003–1016

[30] Park H, Jun C (2009) A simple and fast algorithm for K-medoids clustering. Expert Syst Appl 36:3336–3341

[31] Pelleg, D. and Moore, A. (2000). x-means: Extending k-means with efficient estimation of the number of clusters. In Proceedings of the seventeenth international conference on machine learning, pages 727–734. San Francisco: Morgan Kaufmann.

[32] Phillips,S.(2002).Acceleration of k-means and related clustering algorithms. In Mount, D. and Stein, C., editors, ALENEX: International workshop on algorithm engineering and experimentation,LNCS,volume 2409,pages 166–177.SanFrancicsco: Springer-Verlag.

[33] Reddy, D., and Jana, P.K. (2014) 'A new clustering algorithm based on Voronoi diagram', Int. J. of Data Mining, Modelling and Management, Vol. 6 No. 1

[34] Rehioui, H., Idrissi, A., Abourezq, M. Zegrari, F. (2016), DENCLUE-IM: A New Approach for Big Data Clustering, Procedia Computer Science, Volume 83, 2016, Pages 560-567

[35] Rose, K., Gurewitz, E., and Fox, G. (1990). Statistical mechanics and phase transitions in clustering. Physical Review Letters, 65(8):945–948.

[36] Schikuta, E. (1996). Grid-clustering: An efficient hierarchical clustering method for very large data sets. In Proceedings of the 13th International Conference on Pattern Recognition, volume 2, pages 101–105. Vienna, Austria: IEEE.

[37] Schikuta, E. and Erhart, M.(1997). The BANG- clustering system: Grid-based data analysis. In Liu, X., Cohen, P., and Berthold, M., editors, Lecture Notes in Computer Science, volume 1280, pages 513–524. Berlin, Heidelberg: Springer-Verlag.

[38] Shah-Hosseini, H. (2013) 'Improving K-means clustering algorithm with the intelligent water drops (IWD) algorithm', Int. J. of Data Mining, Modelling and Management, Vol. 5 No. 4

[39] Sheikholeslami, G., Chatterjee, S., and Zhang, A. (2000). WaveCluster: a wavelet-based clustering approach for spatial data in very large databases. The VLDB Journal, 8(3– 4):289–304.

[40] Sibson, R. (1973). SLINK: An optimally efficient algorithm for the single link cluster method. The Computer Journal, 16(1):30–34.

[41] Silva, J. A., Faria, E. R., Barros, R. C., Hruschka, E. R., de Carvalho, A. C. P. L. F., and Gama, J. (2013) 'Data stream clustering: a survey', ACM Comput. Surv. 46, 1, Article 13, 31 pages.

[42] Sim, K., Gopalkrishnan, V., Zimek, A. and Cong, G. (2012) 'A survey on enhanced subspace clustering', Data Mining & Knowledge Discovery, Vol. 26 No. 2"

[43] Späth, H. (1980). Cluster Analysis Algorithms. West Sussex, UK: Ellis Horwood Limited.

[44] Su, Z., Yang, Q., Zhang, H., Xu, X., Hu, Y. (2001), Correlation-based Document Clustering using Web Logs, Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34)-Volume 5, p.5022.

[45] Tan KC, Teoh EJ, Yua Q, Goh KC. (2009) 'A hybrid evolutionary algorithm for attribute selection in data mining', Expert Systems with Applications;36(No.4): 8616e30.

[46] Wang, B., Rahal, I., and Dong, A. (2011) 'Parallel hierarchical clustering using weighted confidence affinity', Int. J. of Data Mining, Modelling and Management, Vol. 3 No. 2

[47] Wang, L., & Dong, M., (2015) 'Exemplar-based low-rank matrix decomposition for data clustering', Data Mining & Knowledge Discovery, Springer, No 29, pp 324-357

[48] Wang, W.,Yang, J., and Muntz, R. (1997). STING: A statistical information grid approach to spatial data mining. In Twenty-third international conference on very large databases, pages 186–195.Athens, Greece: Morgan Kaufmann.

[49] Wang, W., Yang, J., and Muntz, R. (1999). STING+: An approach to active spatial data mining. In Fifteenth international conference on data engineering, pages 116–125. Sydney, Australia: IEEE Computer Society.

[50] Xu D. and Tian Y. (2015) 'A Comprehensive survey of Clustering Algorithms', Annals of Data Science, Springer-Verleg Berlin Heildelberg, DOI: 10.1007/s40745-015-0040-1

[51] Yager R, Filev D (1994) Approximate clustering via the mountain method. IEEE Trans Syst Man Cybern 24:1279–1284

[52] Zhang, B., Hsu, M., and Dayal, U. (2000). k-harmonic means: A spatial clustering algorithm with boosting. In International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining, TSDM 2000, volume 2007 of Lecture Notes in Artificial Intelli- gence, Lyon, France: Springer.

[53] Zhang, T., Ramakrishnan, R., and Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. In Proceedings

of the 1996 ACMSIGMOD international conference on management of data, pages 103–114. NewYork: ACM Press.

[54] Zhao, Y. and Song, J.(2001). GDILC: A grid-based density-isoline clustering algorithm. In Proceedings of the international conferences on info-tech and info-net, 2001, volume 3, pages 140–145. Beijing: IEEE.

[55] Boley D, Gini M, Gross R, Han E, Hastings K, Karypis G, Kumar V, Mobasher B, Moore J (1999), Partitioning-based clustering for web document categorization. Decis Support Syst 27:329–341

[56] Jain A (2010) Data clustering: 50 years beyond K-means. Pattern Recognit Lett 31:651–666

[57] Velmurugan T, Santhanam T (2011) A survey of partition based clustering algorithms in data mining: an experimental approach. Inf Technol J 10:478–484

[58] Murtagh F (1983)Asurvey of recent advances in hierarchical clustering algorithms.Comput J 26:354–359

[59] Carlsson G, Mémoli F (2010) Characterization, stability and convergence of hierarchical clustering methods. J Mach Learn Res 11:1425–1470

[60] Yang M (1993) A survey of fuzzy clustering. Math Comput Model 18:1–16

[61] Baraldi A, Blonda P (1999) A survey of fuzzy clustering algorithms for pattern recognition. I. IEEE Trans Syst Man Cybern Part B 29:778–785

[62] Höppner F (1999) Fuzzy cluster analysis: methods for classification, data analysis and image recognition. Wiley, Hoboken

[63] Januzaj E, Kriegel H, PfeifleM(2004) Scalable density-based distributed clustering. In: Proceedings of the 8th european conference on principles and practice of knowledge discovery in databases, pp 231–244

[64] KriegelH, PfeifleM(2005) Density-based clustering of uncertain data. In: Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining, pp 672–677

[65] Chen Y, Tu L (2007) Density-based clustering for real-time stream data. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining, pp 133–142

[66] Duan L, Xu L, Guo F, Lee J, Yan B (2007) A local-density based spatial clustering algorithm with noise. Inf Syst 32:978–986

[67] Ma E, Chow T (2004) A new shifting grid clustering algorithm. Pattern Recognit 37:503–514

[68] Park N, LeeW(2004) Statistical grid-based clustering over data streams. ACMSIGMOD Rec 33:32–37

[69] Pilevar A, Sukumar M (2005) GCHL: a grid-clustering algorithm for high-dimensional very large spatial data bases. Pattern Recognit Lett 26:999–1010

[70] IndiraPriya, P. and Ghosh, D.K. (2013), A Survey on Different Clustering Algorithms in Data Mining Technique, International Journal of Modern Engineering Research, Vol.3, Issue.1, pp-267-274

[71] Pujari, A.K., Rajesh, K., Reddy, D.S. (2001) Clustering Techniques in Data Mining—A Survey, IETE Journal of Research, 47:1-2, 19-28, DOI: 10.1080/03772063.2001.11416199

[72] Greenlaw, R. and Kantabutra, S. (2013), Survey of Clustering: Algorithms and Applications, International Journal of Information Retrieval Research, 3(2), 1-293

[73] Berkhin P. (2006) A Survey of Clustering Data Mining Techniques. In: Kogan J., Nicholas C., Teboulle M. (eds) Grouping Multidimensional Data. Springer, Berlin, Heidelberg

TABLE I. COMPARISON AMONG CLUSTERING ALGORITHMS BASED ON HIERARCHICAL METHOD

Algorithm Basis	**SLINK**	**CLINK**	**BIRCH**	**CURE**	**ROCK**	**Chameleon**	**DIANA**	**DISMEA**
Reference	[40]	[18]	[53]	[13]	[14]	[23]	[27]	[43]
Algorithm Type	Agglomerative	Agglomerative	Agglomerative	Agglomerative	Agglomerative	Agglomerative	Divisive	Divisive
Method	Graph	Graph	Geometric	Geometric	Geometric	Graph	Gini Index	k-means
Representation	Pointer Representation of Dendrogram	Pointer Representation of Dendrogram	Clustering feature vector	Fixed number of points	links	k-nearest neighbor graph	Series of Successive splits in Dendrogram or Banners	--
Carried out using	Arbitrary dissimilarity coefficient	Dissimilarity measure	CF tree	Combination of random sampling and partitioning	Computation of links	Construction of Sparse Graph	Largest dissimilarity	k-means algorithm to subdivide a cluster
Type of Data	Numerical / Categorical	Numerical / Categorical	Numerical	Numerical	Boolean and Categorical	Numerical	Numerical / Categorical	Categorical
Shape of Clusters	Arbitrary	Convex	Convex with uniform size	Arbitrary with wide variance in size	Arbitrary with wide variance in size	Arbitrary shape	Convex with uniform size	Convex
Sensitive to Outliers / Noise	High	Low	Low	Low	Low	Low	No	No
Scalability for Large Data Sets	No	No	Yes	Yes	No	No	No	Yes
Scalability for High Dimensionality	Yes	No	No	Yes	Yes	Yes	Yes	Yes
Time Complexity	$O(n^2)$	$O(n^2)$	$O(n)$	$O(n^2 logn)$	$O(n^2 logn)$	$O(n^2)$	$O(n^2 logn)$	$O(n^2 logn)$

TABLE II. COMPARISON AMONG CLUSTERING ALGORITHMS BASED ON PARTITIONED-BASED METHOD

Algorithm / **Basis**	**K-means**	**K-medoids**	**PAM**	**CLARA**	**CLARANS**
Reference	[28]	[30]	[26]	[25]	[29]
Type of Data	Numerical	Numerical / Categorical	Numerical	Numerical	Numerical
Shape of Clusters	Convex	Convex	Arbitrary	Arbitrary	Arbitrary
Result	Centre of Clusters	Medoids of Clusters	Medoids of Clusters	Medoids of Clusters	Medoids of Clusters
Sensitive to Outliers / Noise	More	Less	Less	Less	Less
Scalability for Large Data Sets	Yes	No	No	Yes	No
Scalability for High Dimensionality	No	No	No	No	No
Time Complexity	O(knt)	$O(k(n-k)^2)$	$O(k(n-k)^2)$	$O(ks^2+k(n-k))$	$O(kn^2)$

TABLE III. COMPARISON AMONG CLUSTERING ALGORITHMS BASED ON DENSITY-BASED METHOD

Algorithm / **Basis**	**DBSCAN**	**OPTICS**	**Mean-shift**	**DENCLUE**	**RDBC**
Reference	[9]	[2]	[6]	[17]	[44]
Shape of Clusters	Arbitrary	Arbitrary	Arbitrary	Arbitrary	Arbitrary
Sensitive to Outliers / Noise	No	No	No	Low	No
Scalability for Large Data Sets	Yes	Yes	No	Yes	Yes
No of parameters	Small	Large	Small	Large	Small
Time Complexity	O(nlogn)	O(nlogn)	(kernel)	$O(n^2)$ or nlogn	$O(n^2)$

TABLE IV. COMPARISON CLUSTERING ALGORITHMS BASED ON GRID-BASED METHOD

Algorithm / **Basis**	**STING**	**CLIQUE**	**OptiGrid**	**GRIDCLUS**	**GDILC**	**WaveCluster**
Reference	[48]	[1]	[18]	[36]	[54]	[39]
Shape of Clusters	Arbitrary	Convex	Arbitrary	Convex	Arbitrary	Arbitrary
Sensitive to Outliers / Noise	Low	Moderate	No	Yes	No	Yes
Scalability for Large Data Sets	Yes	No	Yes	Yes	Yes	Yes
Scalability for High Dimensionality	Yes	Yes	Yes	Yes	No	No
Time Complexity	O(n)	$O(n+k^2)$	Between O(nd) and O(d n log n)	O(n)	O(n)	O(n)

TABLE V. COMPARISON CLUSTERING ALGORITHMS BASED ON FUZZY METHOD

Algorithm / **Basis**	**Fuzzy k-means**	**Fuzzy k-modes**	**FCM**	**FCS**	**MM**	**MEC**
Reference	[3]	[19]	[4]	[7]	[51]	[35]
Carried out using	Fuzzy k-partition of data set and k-prototypes	Fuzzy k-partition	Nondegenerate fuzzy c-partitions	Fuzzy scatter matrix	Graph	Statistical physics; Maximizing the entropy at a given average variance
Shape of Clusters	Convex	Convex	Convex	Arbitrary	Arbitrary	Convex
Sensitive to Outliers / Noise	Yes	No	High	High	Low	Low
Type of Data	Numerical	Numerical / Categorical	Numerical	Numerical	Numerical	Numerical
Time Complexity	O(knt)	$O(k(n-k)^2)$	O(n)	(kernel)	$O(v^2n)$	O(en)