International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 2 Feb. 2023, pp: 799-806 www.ijaem.net ISSN: 2395-5252

A Comprehensive Study of Different Clustering Algorithms Based on Big Data Analysis

Lipika Barua[1*], Sharif Ahamed[2], Md. Atikur Rahman[3], Md. Rasel Mia[4], Md. Karam Newaz[5]

[1,2,3,4,5]*Department of Computer Science and Engineering, Gono Bishwabidyalay, Savar,Dhaka 1344, Bangladesh*

Date of Submission: 15-02-2023 Date of Acceptance: 25-02-2023

ABSTRACT
There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information from the rapidly growing volumes of digital data. Clustering is a well-known fundamental task of data mining to extract information. However, several researchers have developed and have provided many clustering algorithms for various domains. This complexity makes it difficult for researchers and practitioners to keep up with clustering algorithms development. As a result, finding appropriate algorithms helps significantly to organize information and extract the correct answer from different queries of the databases. In this case, the main objective of this paper is to find the appropriate clustering algorithm for the sparse industrial dataset. To achieve this goal, we first represent related work that focuses on comparing different clustering algorithms over the past 20 years. After that, we provide a categorization of different clustering algorithms found in the literature by matching their properties to the 4V's challenges of big data which allow us to select the candidate clustering algorithm. Finally, using internal validity indices, K-means, agglomerative hierarchical, SOM, and DBSCAN have been implemented and compared on 4 datasets. In addition, we highlighted the best-performing clustering algorithm that gives us the efficient clusters for each dataset.

Keywords: Clustering algorithms, Unsupervised learning, Big data, Sparse dataset, Validity metrics.

I. INTRODUCTION

In this digital era, according to massive progress and development of the internet and online world technologies, data generated by machines and devices, product lifecycle management solutions, production planning systems have reached a huge volume of more than a thousand Exabyte day by day and is expected to increase in the next years. To capture long-term revenues and advantages, companies must manage the knowledge and must have the right information at the right time and under the right format. Hence, we use the data mining process which is at the intersection of AI, statistics, ML, and database systems to extract implicit, previously unknown, and potentially useful information from data.

Clustering analysis is used to classify cases into similar significant groups based on their distinct instances. It is one of the most widely fundamental tasks of data mining for exploratory data analysis. Furthermore, from an optimization perspective, the main goal of clustering is to maximize both the homogeneity within a cluster and the heterogeneity among different clusters [1]. However, clustering is deemed as a form of an unsupervised task, which calculates the similarity between objects without having any information about their correct distribution. The different algorithms developed over the years by the researchers. However, with the vast number of surveys and comparative studies concerning the clustering algorithms, exploring the algorithm that cluster industrial sparse dataset remains an open issue. Therefore, the main goal of this paper is to provide comprehensive reviews of the clustering algorithms that optimally cluster the sparse industrial datasets. To achieve this goal, we seek to reflect the profile of corresponding algorithms by making the first analysis and comparison between five categorized groups. They are partitioning, hierarchical, density, grid, and model-based algorithms. Second, we provide the readers with a proper analysis of the selected algorithms by experimentally comparing them to real datasets concerning internal clustering validation.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 2 Feb. 2023, pp: 799-806 www.ijaem.net ISSN: 2395-5252

This paper is structured as follows, section 2 presents literature reviews over the past 20 years on research and review articles with a focus on comparative research. Section 3 provides a review of clustering algorithms. Section4, introduces the datasets and clustering evaluation measurements. Section 5, concludes the paper and discusses future research.

II. LITERATURE REVIEW

Over the past twenty years, clustering publications became increasingly important, which proves that researchers are paying more and more attention to this problem. To deeply study these works, among 30 papers selected in the literature review, 20 were ultimately used for comparative studies, while 10 were used for applying the clustering method in the industry sector.

There are several research that has extensively studied popular and known algorithm such as K-means, DBSCAN, DENCLUE, K-NN, fuzzy k-means, and SOM to discuss their advantages and disadvantages with taking into account several factors which may influence the criterion in choosing an appropriate clustering algorithm for a given dataset[2, 3, 4, 5]. While other research has looked at providing clustering algorithms surveys based on different criteria such as their score, their applicability, their knowledge about the domain and also based on the size of the dataset, the number of clusters, the type of dataset, time complexity, stability, and so on[6-19,10]. In other research, the authors provide a categorizing framework that systematically groups a collection of existing clustering algorithms into categories concerning the 4V's of big data and concludes the suitable algorithm for a variety of big datasets concerning different measurements types [11, 12, 13, 14, 15].In the industry domain, researchers have looked at different algorithms such as K means, DBSCAN, agglomerative hierarchical clustering, and SOM algorithm to respectively cluster packaging and environmental risk, financial, female workers, customer preferences, industrial hygiene, and forest industry datasets[15, 16, 17, 18]. However, with all these surveys and comparisons found in the literature, there exist some limitations such as the characteristics of the algorithms are not well studied, no rigorous empirical analysis has been carried out to ascertain the benefit of one algorithm over another for one specific type of dataset.

In addition, two types of research, which compare respectively DBSCAN and K-means for financial datasets and agglomerative hierarchical clustering and SOM for packaging modularization datasets[19, 20]. No paper deals with different algorithms properly evaluated and compared on real industrial datasets. Therefore, overviewing and exploring the algorithms that determine the best clusters for sparse industrial datasets remains an open issue. As a consequence, and motivated by these reasons, the next section proposes a categorization framework that groups a collection of existing algorithms into categories.

III. OVERVIEW OF CLUSTERING ALGORITHMS

Clustering algorithms have a strong relationship with many fields, especially statistics and science. A rough but widely agreed frame is to classify clustering techniques into several groups. In this paper, we allocate the clustering algorithms according to five categories: **(1) Partitioning-based algorithms** which regard the center of data points as the center of the corresponding cluster when initial groups are specified and reallocated towards a union; **(2) Hierarchical-based algorithms** shows the relationship between each pair of clusters depending on the medium of similarity or dissimilarity in a hierarchical manner called dendrogram; **(3) Density-based algorithms** separate data objects based on their regions of density, connectivity, and boundary. The data which is in the region with a high density of the data space is considered to belong to the same cluster; **(4) Grid-based algorithms** change the original data space into a grid structure with a definite size of clusters to collect regional statistical data, and then perform the clustering on the grid, instead of the database directly;**(5) Model-based algorithms** select a particular model for each cluster and find the best fitting for that model. There are mainly two kinds of model-based clustering algorithms, one based on the statistical learning method and the other one based on the neural network learning method.

Table 1: Categorization of clustering algorithm concerning 4V of big data.

Categories	Algorithms	Volume		Variety			Velocity	Value
		Dataset Size	HighDimensionality	Noisy Data	DatasetType	ClusterShape	Time complexity	Inputs
Partitioning-	K-means	Large &Small	No	No	Numerical	Non-convex	O(nkd)	1

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 2 Feb. 2023, pp: 799-806 www.ijaem.net ISSN: 2395-5252

based	K-modes	Large	Yes	No	Categorical	Non-convex	$O(n)$	1
	K-medoids	Small	Yes	Yes	Categorical	Non-convex	$O(n^2dt)$	1
	PAM	Small	No	No	Numerical	Non-convex	$O(k(n-k)^2)$	1
Hierarchical-based	Ward	Large &Small	No	No	Numerical	Non-convex	$O(n)$	1
	BIRCH	Large	No	No	Numerical	Non-convex	$O(n)$	2
	CURE	Large	Yes	Yes	Numerical	Arbitrary	$O(n^2 \log n)$	2
	Chameleon	Large	Yes	No	All type	Arbitrary	$O(n^2)$	3
Density-based	DBSCAN	Large	No	No	Numerical	Arbitrary	$O(n \log n)$	2
	OPTICS	Large	No	Yes	Numerical	Arbitrary	$O(n \log n)$	2
	DENCLUE	Large	Yes	Yes	Numerical	Arbitrary	$O(log\|D\|)$	2
Grid-based	Wavecluster	Large	No	Yes	Spatial	Arbitrary	$O(n)$	3
	STING	Large	No	Yes	Spatial	Arbitrary	$O(k)$	1
	CLIQUE	Large	Yes	No	Numerical	Arbitrary	$O(ck+mk)$	2
	OPTIGRID	Large	Yes	Yes	Spatial	Arbitrary	$O(nd)$	3
Model-based	EM	Large	Yes	No	Spatial	Non-convex	$O(knp)$	3
	COBWEB	Small	No	No	Numerical	Non-convex	$O(n^2)$	1
	SOM	Small	Yes	No	Multivariate	Non-convex	$O(n^2m)$	2

However, to facilitate the choice of the appropriate clustering algorithms, table 1 provides a summary of the clustering algorithm concerning the relative strength and weakness of each five categorizations described above and also by matching the considered factors to the 4V's of big data namely Volume, Variety, Velocity, and Value. According to the table, we can state that the algorithms that are chosen are: the k-means algorithm, the agglomerative hierarchical algorithm with ward distance, the Self-Organization Map (SOM),and Density-based Spatial Clustering of Application with Noise (DBSCAN). The general reasons for selecting these four algorithms are popularity, flexibility, and applicability to industrial datasets. Therefore, we haven't selected an algorithm in the grid-based clustering algorithm due to its incapacity to locate clusters in a low dimensional subspace and also to the required type of dataset which is spatial in this case. Thus, the main focus of the next section is to investigate the behavior of the selected algorithm concerning four industrial validated datasets.

IV. MATERIALS AND METHODS

4.1 Dataset Description

The experiments were carried out on four different industrial datasets. The first dataset is gathered from the paper where the author suggests that the design for logistics encompasses four essential subsystems that interact to determine the content of the design for logistics [21]. The aim of this logistics dataset (DS-1) is to regroup modules into homogenous clusters to decrease complexity and enhance efficiency. The second dataset is generated from a list of surveys and studies collected from several papers which deal with the quality and safety requirements [22-24]. The customer requirement dataset (DS-2) contains 98 guidelines decomposed into 16 different modules. The third dataset is generated from the quality requirements of the automotive sector following the International Automotive Task Force IATF/ISO 16949. This automotive-quality system dataset (DS-3)contains 10 categories of requirements based on 210 guidelines. Finally, the fourth dataset is the aircraft dataset (DS-4) extracted from the paper [25]. This dataset aims to cluster 53 different transport aircraft into groups based on 37 applications.

4.2Data Preprocessing and Analysis

After presenting the selected dataset used in the experimental comparison of our algorithms and to improve the quality of the datasets, one important step to achieve in the data mining process is data-preprocessing. This critical step deals with the preparation and transformation of the initial dataset, it is divided into four categories: Data cleaning, Data integration, Data transformation, and Data reduction. In addition, the formulated incidence matrix for the three first datasets is constructed only with two

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 2 Feb. 2023, pp: 799-806 www.ijaem.net ISSN: 2395-5252

entries Entry 1 indicates that a particular design factor does belong to a module whereas entry 0 indicates that the design factors do not belong to the module, while the matrix in the fourth dataset is subjectively on the interval of [0, 2]. Hence, we have conducted these steps to enhance the quality of our four datasets.

4.3Validity Metrics

Several authors have suggested various indices to investigate cluster validity[26-28]. External and Internal clustering validation are the two main categories of cluster validity. In this work there is no external information available, internal validation measures are the only option for cluster validation. To detail the 10 internal measures that are considered in this paper,we shall from now that n is the number of observations, p the number of variables, q the number of clusters, X the n × p matrix of p variables measured on n independent observations.

- **C-index –** is calculated using the equation:

$$Cindex = \frac{S_w - S_{min}}{S_{max} - S_{min}}$$

The minimum value of the index is used to indicate the optimal number of clusters.

- **CH index – is** a popular index using a ratio of the between clustermeans and the within-cluster sum of squares statistic. The equation is:

$$CH(q) = \frac{trace\ (B_q)\ /\ (q-1)}{trace\ (W_q)\ (n-q)}$$

- **Dunn index –**is the ratio between the minimal inter-cluster distances to maximal intra-cluster distance. It is computed as:

$$Dunn = \frac{min_{1 \le i < j \le q}\ d(C_i\ , C_j)}{max_{1 \le k < q}\ d(C_k)}$$

The number of clusters that maximize Dunn is taken as the optimal number of clusters and indicates that the clusters are compact and well separated.

- **Gamma index**– is based on the comparison between all within-cluster dissimilarities and between-cluster dissimilarities. The equation is:

$$Gamma = \frac{s(+) - s(-)}{s(+) + s(-)}$$

The maximum value of the index is taken to present the correct number of clusters.

- **The Ball-Hall index –** is based on the average distance of the items to their respective cluster centroids. It is computed using the formula:

$$BH = \frac{W_q}{q}$$

The largest difference between levels is used to indicate the optimal solution.

- **Davies-Bouldin (DB) index –** is a function of the sum ratio of within-cluster scatter to between cluster separations. It is calculated using the equation:

$$\mathrm{DB(q)} = \frac{1}{\mathrm{q}}\sum_{\mathrm{k}=1}^{\mathrm{q}} \max_{\mathrm{k}\neq\mathrm{l}} \left(\frac{\delta_\mathrm{k} + \delta_\mathrm{l}}{\sqrt[\mathrm{v}]{\sum_{\mathrm{k}=1}^{\mathrm{q}} \left|\mathrm{c_{kj}} - \mathrm{c_{lj}}\right|^{\mathrm{v}}}} \right)$$

DB values close to 0 indicate that the clusters are compact and far from each other.

- **Tau index –** is computed between corresponding entries in two matrices. This index is computed using the equation:

$$\mathrm{Tau} = \frac{\mathrm{s(+)} - \mathrm{s(-)}}{\sqrt{\frac{\mathrm{N_t(N_t-1)}}{2-\mathrm{t}} \times \frac{\mathrm{N_t(N_t-1)}}{2}}}$$

The maximum value of the index is taken as an indication of the correct number of clusters.

- **Connectivity index**– measures the distance between observations placed in the same cluster as their nearest neighbors. The connectivity is defined as:

$$\mathrm{Conn}\ (\zeta) = \frac{1}{\mathrm{B}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{L}} \mathrm{x_{i,nn_i(j)}}$$

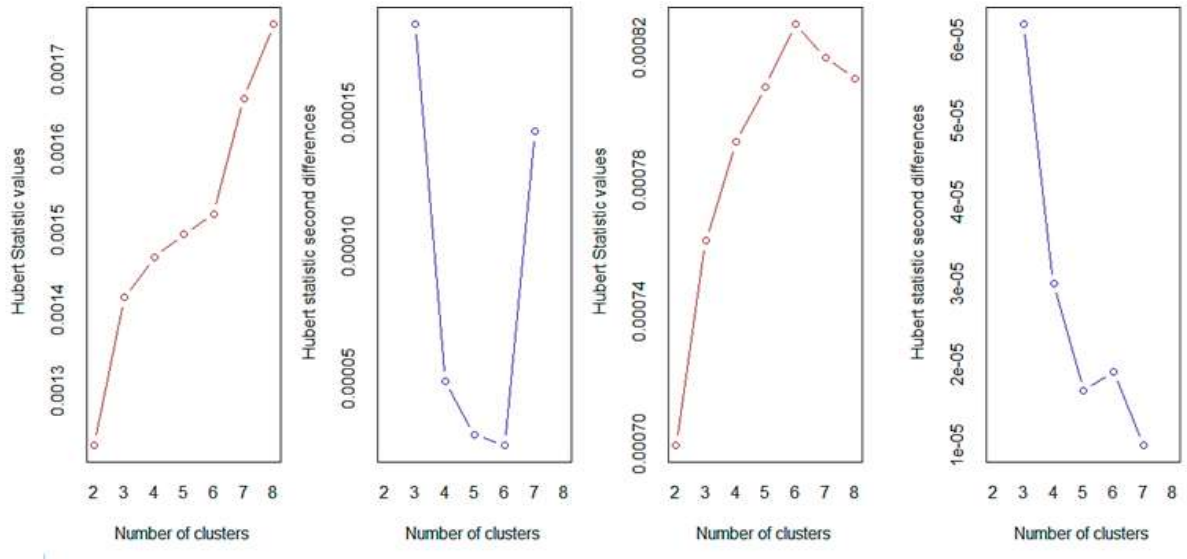
The connectivity has a value between zero and ∞ should be minimized.

V. RESULTS AND DISCUSSIONS

Our experiments are divided into two parts. We define first, the appropriate number of clusters that should be considered in each of the fourth datasets using quantitative and graphical interpretation. After that, we compute all the previous indices to compare the selected four algorithms and to select the most appropriate one for clustering small numerical sparse datasets. To find the relevant number of clusters in each dataset, we use the Nbclust package. It provides 30 indices available in SAS and R in one package. Hence and according to the majority rules, 4 would be the best number of clusters for DS-1, 3 for the DS-2, 5 for theDS-3,and 4 for the DS-4. Moreover, if we look at the Hubert index, which is a graphical method, the optimal number of clusters is identified by a significant knee in the plot of index values against

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 2 Feb. 2023, pp: 799-806 www.ijaem.net ISSN: 2395-5252

the number of clusters. This knee corresponds to a significant increase or significant decrease of the index, as the number of clusters varies from the minimum to the maximum. In other words, a significant peak in the plot of second differences values indicates the relevant number of clusters. Hence, as shown in figure 1, for the DS-1 and DS-4 for example, the Hubert index confirms our purpose and proposes 3 and 4 as the best number of clusters for the DS-4andDS-1respectively. Consequently, using this approach, the user faces the dilemma of choosing the best number of clusters.



(a) (b)

Figure 1: (a) Hubert index for DS-4, (b) Hubert index for DS-1.

After selecting the best number of clusters in each dataset, we are able now to run the selected four algorithms to compare first the best clustering algorithm and second to define the appropriate clustering for each dataset. To do so, we compute all the presented internal indices to exploit prior knowledge of data and cluster their label. In this sense, figure 2 andtables2, 3, 4, and 5report the results of the candidate clustering algorithms according to the internal validity measures, from which we can infer several observations.

Table 2:The candidate clustering algorithm'sinternal validity results for DS-1.

Algorithms	**Internal Indices**							
	C-index	**CH**	**Dunn**	**Gamma**	**BH**	**DB**	**Tau**	**Connectivity**
K-means	0.36	1.63	0.69	0.26	79.17	1.69	0.16	18.77
Hierarchical	0.22	1.88	0.71	0.61	79.59	1.93	0.38	18.78
SOM	**0.15**	**1.67**	**0.80**	**0.67**	68.28	**2.66**	**0.45**	**14.98**
DBSCAN	0.17	**1.67**	0.75	0.60	**121.51**	0.76	0.43	26.34

Table 3:The candidate clustering algorithm's internal validity results for DS-2.

Algorithms	**Internal Indices**							
	C-index	**CH**	**Dunn**	**Gamma**	**BH**	**DB**	**Tau**	**Connectivity**
K-means	0.19	7.80	0.29	**0.60**	13.61	**1.67**	0.39	23.62
Hierarchical	0.31	5.77	0.26	0.32	9.16	1.61	0.22	26.50
SOM	**0.16**	**8.20**	**0.35**	**0.60**	11.12	1.42	**0.49**	**16.24**
DBSCAN	**0.16**	8.33	**0.35**	**0.60**	**18.42**	1.45	0.42	16.93

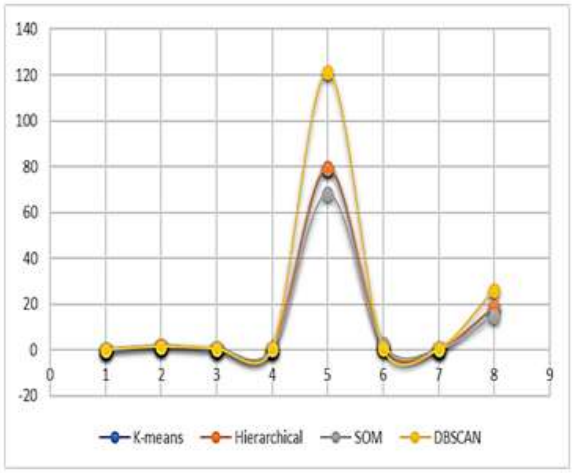
International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 2 Feb. 2023, pp: 799-806 www.ijaem.net ISSN: 2395-5252

Table 4:The candidate clustering algorithm's internal validity results for DS-3.

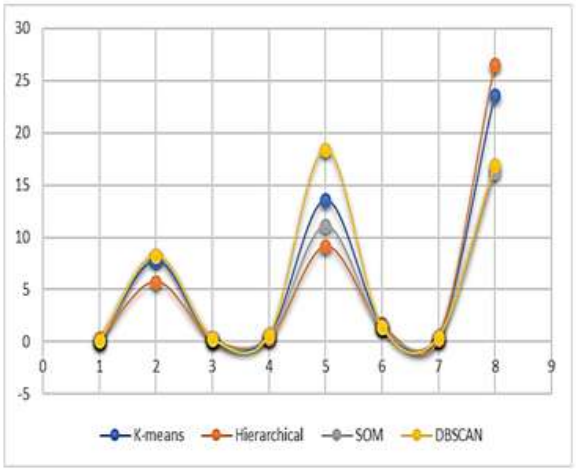
Algorithms	Internal Indices							
	C-index	CH	Dunn	Gamma	BH	DB	Tau	Connectivity
K-means	**0.05**	28.66	**0.67**	0.79	**13.37**	**1.30**	0.39	26.84
Hierarchical	0.12	16.61	0.44	0.69	3.55	0.43	0.47	32.98
SOM	**0.05**	**38.77**	0.43	**0.87**	2.10	0.69	**0.56**	**25.30**
DBSCAN	0.06	22.27	0.56	0.82	1.85	1.01	0.55	27.60

Table 5:The candidate clustering algorithm's internal validity results for DS-4.

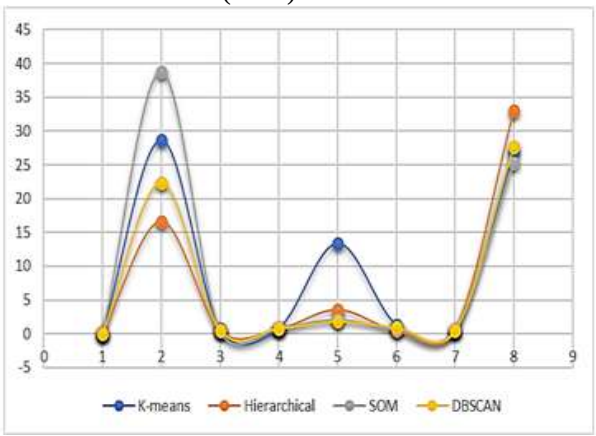
Algorithms	Internal Indices							
	C-index	CH	Dunn	Gamma	BH	DB	Tau	Connectivity
K-means	**0.14**	15.81	**0.47**	**0.87**	24.29	0.98	**0.58**	**10.22**
Hierarchical	**0.14**	14.04	0.33	0.86	24.05	0.98	0.57	11.22
SOM	**0.14**	**13.58**	0.25	0.70	25.81	**1.34**	0.47	10.78
DBSCAN	0.18	18.46	0.35	0.63	**26.19**	**1.34**	0.38	18.09



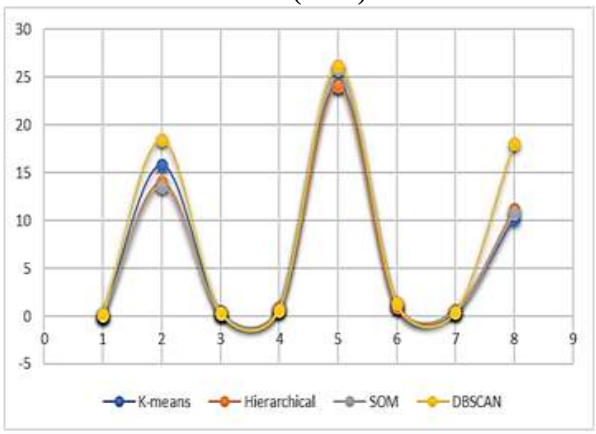
(DS-1)



(DS-2)



(DS-3)



(DS-4)

Figure 2: Performance analysis of clustering algorithms internal validity results for four datasets.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 2 Feb. 2023, pp: 799-806 www.ijaem.net ISSN: 2395-5252

First, it can be seen from the table that according to the sizes of the datasets, SOM clustering output for the three first datasets shows good results on almost all internal measures in comparison to the remaining clustering algorithms, while K-means perform well for the DS-4 (K=5). In other words, the quality of the K-means becomes very good when using a huge dataset and it is confirmed by table 2, 3, 4, and 5 presented in the previous section. Furthermore, as the number of instances becomes lower, SOM shows more accuracy in clustering the objects into their suitable clusters than other algorithms. Moreover, most of the time hierarchical clustering and SOM show the same results when the number of instances becomes greater (DS-4). Second, according to the number of clusters, as the value of q becomes greater, the performance of the SOM algorithm becomes lower and the K-means algorithm becomes higher. This is seen in the DS-4which requires a higher number of clusters compared to other ones (k=5). Third, DBSCAN gives better results compared to hierarchical and K-means algorithms when using noisy datasets. Infact, for the DS-2 which is generated from surveys and studies, there were incomplete responses, there were responses that are not compatible, and also there were responses where the responders are not professional in designing for safety and quality. For this reason and because K-means, hierarchical and SOM algorithms are sensitive to noise, this makes it difficult to cluster an object into its suitable cluster and this will certainly affect the result of the algorithm. In addition, we can state that the hierarchical clustering algorithm is the most sensitive for the noisy dataset, more than K-means and SOM.

However, as a general conclusion, SOM is recommended for small datasets while DBSCAN for noisy datasets and K-means for the huge dataset. SOM algorithm differs from other clustering algorithms and especially from other artificial neural networks. In this respect, it is considered an efficient method able to produce not only compact and connected clusters but also a well-separated ones.

VI. CONCLUSION AND FUTURE WORKS

This paper provides a comprehensive survey and intends to compare popular, flexible, and applicable clustering algorithms in the industry field. Through an extensive search, there exists no such comprehensive survey that has intuitively attempted to compare the four clustering algorithms under investigation in this field. We have presented a simple categorization framework that would automatically recommend the most suitable algorithm that will be properly evaluated in a real industrial dataset. Following the empirical study, we can draw the following conclusions: (1) SOM shows in most of the datasets excellent performance and the best clusters in comparison with the reaming clustering algorithms; (2) As the number of clusters q becomes greater the performance of SOM algorithm becomes lower; (3) K-means and SOM are sensitive to a noisy dataset but the hierarchical clustering is the more sensitive one, hence the DBSCAN is the best in this case; (4) No clustering algorithm performs well for all the internal validity measures; (5) K-means give a higher result for huge data than SOM and hierarchical clustering algorithm.

Furthermore, we can extend our comparison to other clustering algorithms with other different industrial datasets. Then, we can develop an algorithm to directly and automatically compare the algorithms based on different validity indices. Based on the promising findings presented in this paper, work on the remaining issues is continuing and will be presented in future papers.

REFERENCES

[1]. Hancer, E., & Karaboga, D. (2017). A comprehensive survey of traditional, merge-split and evolutionary approaches proposed for determination of cluster number. Swarm and Evolutionary Computation, 32, 49-67.

[2]. Chen, G., Jaradat, S. A., Banerjee, N., Tanaka, T. S., Ko, M. S., & Zhang, M. Q. (2002). Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data. Statistica Sinica, 241-262.

[3]. Feizi-Derakhshi, M. R., & Zafarani, E. (2012). Review and comparison between clustering algorithms with duplicate entities detection purpose. International Journal of Computer Science & Emerging Technologies, 3(3).

[4]. Ayed, A. B., Halima, M. B., & Alimi, A. M. (2014, August). Survey on clustering methods: Towards fuzzy clustering for big data. In Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of (pp. 331-336). IEEE.

[5]. Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on neural networks, 16(3), 645-678.

[6]. He, L., WU, L. D., & CAI, Y. C. (2007). Survey of Clustering Algorithms in Data Mining [J]. Application Research of Computers, 1, 10-13.

[7]. Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on neural networks, 16(3), 645-678.
[8]. Treshansky, A., & McGraw, R. M. (2001, September). Overview of clustering algorithms. In Enabling Technology for Simulation Science V (Vol. 4367, pp. 41-52). International Society for Optics and Photonics.
[9]. Abbas, O. A. (2008). Comparisons between Data Clustering Algorithms. International Arab Journal of Information Technology (IAJIT), 5(3).
[10]. Singhal, G., Panwar, S., Jain, K., & Banga, D. (2013). A comparative study of data clustering algorithms. International Journal of Computer Applications, 83(15).
[11]. Sajana, T., Rani, C. S., & Narayana, K. V. (2016). A survey on clustering techniques for big data mining. Indian Journal of Science and Technology, 9(3).
[12]. Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y. & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE transactions on emerging topics in computing, 2(3), 267-279.
[13]. Sherin, A., Uma, D. S., Saranya, K., & Vani, M. S. (2014). Survey On Big Data Mining Platforms, Algorithms And Challenges. International Journal of Computer Science & Engineering Technology, 5.
[14]. Nagpal, P. B., & Mann, P. A. (2011). Survey of Density Based Clustering Algorithms. International journal of Computer Science and its Applications, 1(1), 313-317.
[15]. Simula, O., Vasara, P., Vesanto, J., & Helminen, R. (1999). The self-organizing map in industry analysis. Industrial Applications of Neural Networks, Washington.
[16]. Craigen, D., Gerhart, S., & Ralston, T. (1993). An international survey of industrial applications of formal methods. In Z User Workshop, London 1992 (pp. 1-5). Springer, London.
[17]. Milton, D. K., Hammond, S. K., & Spear, R. C. (1999). Hierarchical cluster analysis applied to workers exposures in fiberglass insulation manufacturing. The Annals of occupational hygiene, 43(1), 43-55.
[18]. Swan, S. H., Beaumont, J. J., Hammond, S. K., Vonbehren, J., Green, R. S., Hallock, M. F., ... & Schenker, M. B. (1995). Historical cohort study of spontaneous abortion among fabrication workers in the semiconductor health study: Agent-level analysis. American journal of industrial medicine, 28(6), 751-769.
[19]. Cai, F., Le-Khac, N. A., & Kechadi, T. (2016). Clustering approaches for financial data analysis: a survey. arXiv preprint arXiv:1609.08520.
[20]. Zhao, C., Johnsson, M., & He, M. (2017). Data mining with clustering algorithms to reduce packaging costs: A case study. Packaging Technology and Science, 30(5), 173-193.
[21]. Dowlatshahi, S. (1999). A modeling approach to logistics in concurrent engineering. European Journal of Operational Research, 115(1).
[22]. Dowlatshahi, D., MacQueen, G. M., Wang, J. F., Reiach, J. S., & Young, L. T. (1999). G Protein-Coupled Cyclic AMP Signaling in Postmortem Brain of Subjects with Mood Disorders. Journal of neurochemistry, 73(3), 1121-1126.
[23]. Zhu, A. Y., von Zedtwitz, M., Assimakopoulos, D., & Fernandes, K. (2016). The impact of organizational culture on Concurrent Engineering, Design-for-Safety, and product safety performance. International Journal of Production Economics, 176, 69-81.
[24]. Goh, Y. M., & Chua, S. (2016). Knowledge, attitude and practices for design for safety: A study on civil & structural engineers. Accident Analysis & Prevention, 93.
[25]. McCormick Jr, W. T., Schweitzer, P. J., & White, T. W. (1972). Problem decomposition and data reorganization by a clustering technique. Operations Research, 20(5), 993-1009.
[26]. Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. Psyc
[27]. Halkidi, M., Vazirgiannis, M., & Batistakis, Y. (2000, September). Quality scheme assessment in the clustering process. In European Conference on Principles Data Mining and Knowledge Discovery (pp. 265-276). Springer, Berlin, Heidelberg.
[28]. Theodoridis, S., & Koutroubas, K. (1999). Feature generation II. Pattern recognition, 2, 269-320.