International Journal of Current Trends in Engineering and Technology
www.ijctet.org, ISSN: 2395-3152
Volume: 09, Issue: 05 (September -October, 2023)

Improved Data Analysis Using K-Means Algorithm with Quantum Particle Swarm Optimization Algorithm and PVSFCA

Shilpa Meshram, Jayshree Boaddh
Computer Science and Engineering Department
Mittal Institute of Technology, Bhopal, India
shilpameshram028@gmail.com, jayshree.boaddh@gmail.com

Abstract: Data analysis based on the K-Means algorithm has been enhanced by integrating the Quantum Particle Swarm Optimization (QPSO) approach and the proposed Vector Space Function Clustering Algorithm (PVSFCA). Data mining aims to extract valuable insights from vast datasets and present them in an understandable format for further utilisation. Clustering involves grouping objects so that objects within the same cluster exhibit greater similarity to each other than those in other clusters. K-Means clustering involves partitioning a dataset into K clusters. This study explores the significance and prevalent data mining techniques, investigates clustering's role in data mining, and delves into the characteristics, fundamental principles, and execution process of the K-means algorithm. Although K-means clustering is widely used for its simplicity, efficiency, and empirical success, classic K-means has shortcomings like predefining K, random initial centre selection, and more, impacting its performance. Numerous variations of K-means have emerged to address these limitations. K-means is often utilised to minimise the squared distance between feature values of points within the same cluster. The Quantum Particle Swarm Optimization (QPSO) algorithm, integrated with K-means (QPSO-K-means), is an evolutionary computation technique to find sub-optimal solutions in various scenarios. This approach simulates cluster centres as particles to obtain stable and suitable clusters, leading to effective clustering outcomes. The proposed algorithm (PVSFCA) is analysed using the UCI healthcare dataset, demonstrating its efficiency and accuracy. The algorithm's outcomes provide a reliable foundation for enhancing clustering strategies by considering factors such as iterations, error rate, and the optimal creation of cluster centres.

Keywords: Data Mining, Unsupervised Learning, Clustering, Data Mining Techniques, QPSO-K-means Clustering Algorithm, Datasets, Data Analysis, Intelligent Data Analysis, Error Rate, PVSFCA.

I. INTRODUCTION

The information industry has witnessed a substantial surge in interest towards data mining, largely due to the abundance of extensive datasets and the pressing necessity to transform such data into actionable insights and knowledge. In the past decade, the significance of data mining has grown exponentially, especially in the face of intense market competition. The timely delivery of high-quality information has emerged as a pivotal factor in effective decision-making. However, the real world is inundated with copious amounts of data, making it challenging to sift through this vast sea of information to extract what is relevant and present it in a timely and organised manner. Recent advancements in information technology have resulted in the generation of enormous quantities of data. These datasets often lack structured formats, making analysis a complex task. Data clustering has emerged as a widely employed technique to derive meaningful insights, summarise information, uncover natural patterns, and identify hidden relationships within the data. Data clustering, also called clustering analysis, involves segregating a dataset into coherent groups (clusters), where objects within the same cluster exhibit similarity while differing from objects in other clusters. It represents an unsupervised classification endeavour where predefined classes are absent. The pivotal role of clustering transcends various domains, including machine learning, making it one of the cornerstones of data mining. This method partitions data objects into groups based on information encapsulated within the data, leveraging the relationships among objects and their associated feature values. This technique finds applications in diverse realms, such as knowledge discovery, vector quantisation, pattern recognition, data mining, and data dredging. The central objective of clustering is to segment extensive datasets into meaningful clusters. Two primary approaches for clustering are hierarchical clustering and partitioned clustering. The

K-means algorithm falls within the domain of partitioned clustering, producing a localised suboptimal solution. Conversely, the Quantum Particle Swarm Optimization based on the K-means clustering algorithm (QPSO-KMCA) employs a globalised search methodology, enhancing the K-means algorithm's capability to reach global suboptimal solutions. By synergising the strengths of both algorithms, it becomes possible to create a novel algorithm that mitigates the limitations of each algorithm, resulting in a comprehensive and combined solution.

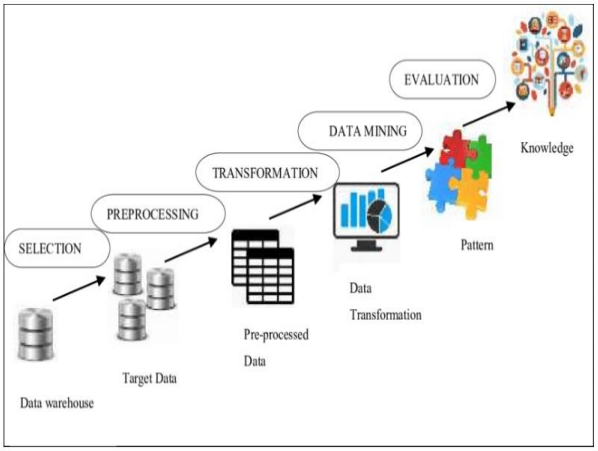

Figure 1: Knowledge Discovery Database (KDD) Process

1.1 Clustering

Clustering constitutes a fundamental category within unsupervised learning. In unsupervised learning, insights are drawn from datasets without predefined labelled responses. Clustering aims to uncover meaningful structures, underlying processes, generative attributes, and inherent groupings within a collection of examples. The essence of clustering lies in segregating a population or data points into distinct groups, where data points within the same group exhibit greater similarity while being dissimilar to data points in other groups. The concept of clustering revolves around grouping objects based on their similarities and dissimilarities. For instance, data points on a graph that cluster can be categorised into a single group. In the graphical representation below, we can discern the presence of three distinct clusters [3].

1.2 Quantum Particle Swarm Optimisation based on K-means Clustering Algorithm (QPSOKMCA)

Initially proposed by M QPSOKMCA, the K-means algorithm has found extensive application in solving clustering problems, playing a vital role in knowledge discovery and data mining. Over time, however, researchers have unveiled several shortcomings within the K-means algorithm. The proliferation of diverse data types and the exponential growth in data volume, particularly fueled by the expansion of computer information networks, have presented new challenges. While the K-means algorithm boasts operational swiftness and simplicity, it is sensitive to initial values, often succumbing to the drawback of local optima. Enter Quantum Particle Swarm Optimisation based on K-means Clustering Algorithm (QPSOKMCA), a relatively recent algorithm developed to address these limitations. It inherits the essence of the genetic algorithm. Still, it streamlines the operation process by eliminating the "selection" and "variation" stages, instead leveraging the group optimal solution to determine the global optimal solution [5]. This bionic optimisation algorithm is grounded in iterative principles and stands out for its rapid execution, uncomplicated steps, and impressive precision. Upon examining relevant data, it becomes evident that QPSOKMCA excels in global search capabilities and local optimisation, thereby rectifying the operational shortcomings of the K-means algorithm [6].

Figure 2: Clustering Process in Different Clusters

II. RELATED WORK

Lili Bai et al. [7] proposed an enhanced K-means clustering algorithm based on improved quantum particle swarm optimisation. In this approach, the cluster centre is simulated as a particle, and cloning and mutation operations are utilised to enhance the

diversity and global search capabilities of the Quantum Particle Swarm Optimization (QPSO) method. This approach yields stable cluster centres, leading to effective clustering results. Semeh Ben et al. [8]** explored the efficiency of partitional clustering algorithms in pattern recognition. While the classic K-means algorithm is proficient in numeric clustering, it falls short in handling categorical clustering. The authors introduced a new categorical method called MFk-M, which transforms categorical data into numeric values using relative frequency for each attribute modality. RSM Lakshmi et al. [9] highlighted the emergence of clustering algorithms as learning aids in handling extensive data volumes. Clustering serves to group similar data for various applications. However, the challenges of applying clustering algorithms to big data, including uncertainty and classification deficiencies, have prompted further exploration into their properties and definitions. Rezaee Jordehi et al. [8] delved into particle swarm optimisation (PSO) techniques for addressing discrete optimisation problems. They explored strategies within PSO for handling discrete variables and thoroughly examined the strengths and drawbacks of each strategy. Poli et al. [11] discussed the substantial success of Particle Swarm Optimization (PSO) applications, acknowledging the proliferation of successful PSO implementations. They aimed to categorise many publications dealing with PSO applications, providing a comprehensive overview of the field's advancements. Kohei Arai et al. [12] recognised the challenge of K-means' initial starting points potentially leading to local optima. They proposed a new approach to optimise initial centroids by leveraging multiple K-means clustering results. By incorporating Hierarchical algorithms, the method demonstrates enhanced clustering outcomes. Chouhan et al. (2018) [16] introduced a method combining Particle Swarm Optimization (PSO) and K-means for document clustering. PSO locates optimal points in the search space, which then serve as initial cluster centroids for K-means, leading to improved document clusters. Janani et al. (2019) [17] presented a novel approach by combining Spectral Clustering with PSO (SCPSO) to enhance text document clustering. Through a combination of global and local optimisation functions, SCPSO successfully manages extensive text document datasets. Barakbah et al. [12] evaluated the performance of the K-means algorithm based on various datasets and algorithms. QPSOKMCA and PVSFCA algorithms were tested for time taken, error rate, and number of iterations, with PVSFCA demonstrating superiority in certain cases. Barakbah et al. [12] introduced the Centronit algorithm for optimising initial centroids in K-means clustering. The proposed method is based on calculating the average distance of the nearest data within the minimum distance region, leading to more effective clustering results. Caron, Mathilde et al. [18] highlighted that although widely studied in computer vision, clustering has not been extensively adapted for the end-to-end training of visual features on large-scale datasets.

III. SIMULATION TOOL

The simulation tool employed for this research is MATLAB, which stands for "Matrix Laboratory." MATLAB is a proprietary programming language and numeric computing environment developed by MathWorks. It offers a versatile platform for performing matrix manipulations, generating function plots, implementing algorithms, creating user interfaces, and interacting with programs written in other languages. While MATLAB is primarily designed for numeric computations, it includes an optional toolbox that utilises the MuPAD symbolic engine, enabling symbolic computing capabilities. The Simulink package is also available, allowing for graphical multi-domain simulation and model-based design in dynamic and embedded systems. With over 4 million users worldwide as of 2020, MATLAB caters to diverse fields, including engineering, science, and economics. Furthermore, as of 2017, more than 5,000 colleges and universities globally integrate MATLAB into their curriculum to facilitate instruction and research.

III. RESULT ANALYSIS

The research delves into data mining by integrating quantum particle swarm optimisation with K-means clustering. While the K-means algorithm is inherently inclined towards generating local suboptimal solutions, the Quantum Particle Swarm Optimization based on K-means Clustering Algorithm (QPSOKMCA) introduces a globalised search methodology capable of seeking global suboptimal solutions with higher error rates. The proposed algorithm presents a unique amalgamation that capitalises on the strengths of both individual algorithms, effectively overcoming their drawbacks to

provide a comprehensive solution. By applying this combined approach, the study aims to identify the minimum attainable error rate for healthcare dataset analysis while striving to determine the optimal solution.

Table 1: Breast Cancer Dataset Analysis (Case 1)

Dataset	Random values	Algorithm	Time (sec)	ER (%)	iteration
Breast cancer dataset	0.3287	QPSOKMCA	1.25	4.36	4
		PVSFCA	1.05	1.70	5
	0.0967	QPSOKMCA	0.13	2.65	3
		PVSFCA	0.08	0.09	5
	0.5201	QPSOKMCA	0.97	4.60	4
		PVSFCA	0.44	1.70	5
	0.5645	QPSOKMCA	1.05	4.63	4
		PVSFCA	1.00	1.72	5

Case 1: Data Analysis Enhancement

The improvement of data analysis was realised by synergising the K-means algorithm with the Quantum Particle Swarm Optimisation algorithm and the proposed Vector Space Function Clustering Algorithm (PVSFCA). This amalgamation culminated in the attainment of effective clustering outcomes. The UCI healthcare dataset is the testing ground for evaluating the method's performance. Applying our recommended algorithm, PVSFCA proved swift and efficient in the analytical process. The results yielded from the analysis were found to be accurate and dependable.

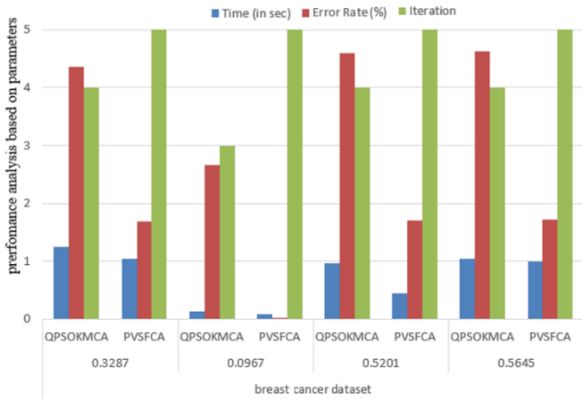

Figure 3: Comparative Analysis between QPSOKMCA and PVSFCA

Furthermore, this study provides a robust and valuable foundation for implementing enhanced clustering strategies. This research empowers the deployment of effective clustering improvement measures by quantifying parameters such as the number of iterations, error rate, and the optimal configuration of cluster centers for generating optimal clusters. In Case 1, the existing algorithm (QPSOKMCA) and the proposed algorithm (PVSFCA) were employed to analyse a breast cancer dataset. Key metrics such as error rate, data analysis time, and the number of iterations were meticulously calculated and compared. The results are depicted in Figure 3 below, which unmistakably illustrates the superiority and reliability of our proposed algorithm.

Figure 3 reinforces the superior performance of the proposed algorithm, PVSFCA, confirming its status as the optimal and dependable solution for enhanced data analysis.

Case 2: Thyroid Dataset Analysis

Table 2: Thyroid Dataset Analysis

Dataset	Random values	Algorithm	Time (sec)	ER (%)	Iteration
Thyroid Dataset	0.945	QPSOKMCA	0.15	2.65	3
		PVSFCA	0.09	0.01	4
	0.053	QPSOKMCA	0.73	3.73	5
		PVSFCA	0.70	1.03	6
	0.575	QPSOKMCA	1.06	4.02	4
		PVSFCA	0.55	1.24	5
	0.845	QPSOKMCA	1.08	4.22	4
		PVSFCA	1.09	1.37	5

Case 2 showcases the enhancement of data analysis by integrating the K-means approach with the Quantum Particle Swarm Optimisation algorithm and the proposed Vector Space Function Clustering Algorithm (PVSFCA). As a result, effective clustering outcomes are achieved. The UCI healthcare dataset is again employed to evaluate the method's performance, confirming the versatility of the proposed algorithm, PVSFCA, which boasts swift and efficient analytical capabilities. The accuracy and reliability of the analysis findings are reinforced, setting a solid foundation for implementing effective clustering improvement strategies. This research presents a robust reference base by quantifying key characteristics such as the number of iterations, error rate, and optimal configuration of cluster centres to achieve optimal clusters.

Additionally, echoing the methodology utilised in Case 1, the existing algorithm (QPSOKMCA) and the proposed algorithm (PVSFCA) were utilised to analyse the thyroid dataset. Essential metrics were

meticulously evaluated and compared, including error rate, data analysis time, and iteration count. The comparison between the previous algorithm (QPSOKMCA) and the proposed algorithm (PVSFCA) is illustrated in Figure 4 below, further corroborating the superiority and reliability of the proposed algorithm.

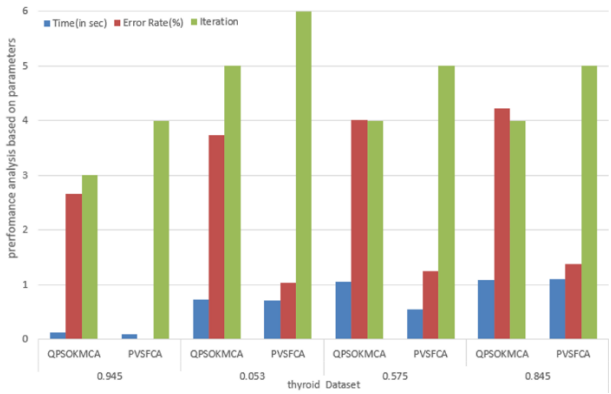

Figure 4: Comparative Analysis between QPSOKMCA and PVSFCA for Thyroid Dataset

Figure 4 conclusively supports the superiority of the proposed PVSFCA algorithm, reaffirming its standing as an optimal and dependable solution for improved data analysis, particularly in the context of the thyroid dataset.

Case 3: E. coli Dataset Analysis

Table 3: E. coli Dataset Analysis

Dataset	Random Values	Algorithm	Time (sec)	ER (%)	Iteration
E. coli Dataset	0.0173	QPSOKMCA	0.23	3.09	5
		PVSFCA	2.59	0.85	6
	0.3731	QPSOKMCA	0.26	0.45	5
		PVSFCA	0.31	0.45	5
	0.9784	QPSOKMCA	0.25	2.65	3
		PVSFCA	0.08	0.01	4
	0.2376	QPSOKMCA	1.14	4.59	10
		PVSFCA	2.22	1.78	10

In Case 3, the augmentation of data analysis was achieved through the fusion of the K-means algorithm with the Quantum Particle Swarm Optimisation algorithm and the proposed Vector Space Function Clustering Algorithm (PVSFCA). This integration resulted in the acquisition of effective and improved clustering outcomes. The evaluation process was again conducted using the UCI healthcare dataset, emphasising the versatility and efficiency of the suggested algorithm, PVSFCA. The precision and reliability of the analysis findings were confirmed, laying a strong groundwork for the application of impactful clustering enhancement strategies. This research imparts a sturdy reference framework by quantifying crucial attributes such as the number of iterations, error rate, and optimal configuration of cluster centres to attain optimal clusters.

Furthermore, analogous to the methodology employed in Cases 1 and 2, the existing algorithm (QPSOKMCA) and the proposed algorithm (PVSFCA) were leveraged to analyse the E. coli dataset. Fundamental metrics underwent rigorous scrutiny and comparison, including error rate, data analysis time, and iteration count. The comparison between the previous algorithm (QPSOKMCA) and the proposed algorithm (PVSFCA) is visually depicted in Figure 5 below, unequivocally underscoring the prowess and dependability of the proposed algorithm.

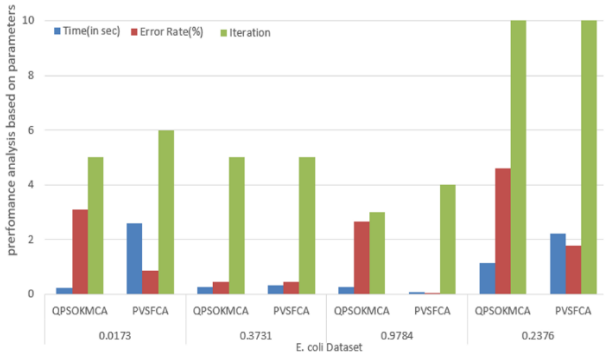

Figure 5: Comparative Analysis between QPSOKMCA and PVSFCA for E. coli Dataset

Figure 5 substantiates the excellence of the proposed PVSFCA algorithm, serving as a conclusive testament to its status as a superior and trustworthy solution for advancing data analysis, specifically in the context of the E. coli dataset.

Case 4: Yeast Dataset Analysis

In Case 4, data analysis advancement was achieved by amalgamating the K-means algorithm with the Quantum Particle Swarm Optimisation algorithm and the proposed Vector Space Function Clustering Algorithm (PVSFCA). The result was the acquisition of clustering outcomes that demonstrated effectiveness and enhancement. Once more, the evaluation process leveraged the UCI healthcare dataset, reaffirming the adaptability and efficiency of the suggested algorithm, PVSFCA. The precision and

reliability of the analysis findings were verified, establishing a robust foundation for implementing influential clustering enhancement strategies. By quantifying vital attributes such as the number of iterations, error rate, and the optimal arrangement of cluster centres to achieve optimal clusters, this research furnishes a substantial reference framework.

Table 4: Yeast Dataset Analysis

Dataset	Random values	Algorithm	Time (sec)	ER (%)	Iteration
Yeast dataset	0.0129	QPSOKMCA	0.24	3.32	4
		PVSFCA	0.16	0.67	5
	0.5058	QPSOKMCA	0.60	4.57	4
		PVSFCA	0.70	1.70	5
	0.4302	QPSOKMCA	0.60	4.41	4
		PVSFCA	1.12	1.72	5
	0.9342	QPSOKMCA	0.10	2.65	3
		PVSFCA	0.06	0.01	4

Furthermore, mirroring the methodological approach utilised in the previous cases, the pre-existing algorithm (QPSOKMCA) and the proposed algorithm (PVSFCA) underwent scrutiny during the analysis of the Yeast dataset. Fundamental metrics were rigorously assessed and juxtaposed, including error rate, data analysis time, and iteration count. The juxtaposition between the previous algorithm (QPSOKMCA) and the proposed algorithm (PVSFCA) is visually presented in Figure 6 below, definitively highlighting the competence and dependability of the proposed algorithm.

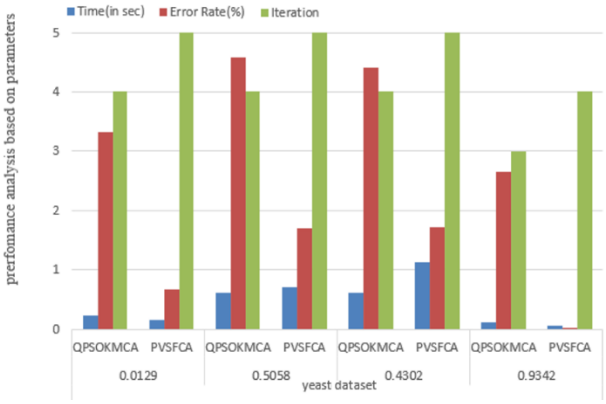

Figure 6: Comparative Analysis between QPSOKMCA and PVSFCA for Yeast Dataset

Figure 6 serves as compelling visual evidence, affirming the supremacy of the proposed PVSFCA algorithm. This evidence underscores its efficacy as a superior and trustworthy solution for advancing data analysis, specifically concerning the Yeast dataset.

IV. CONCLUSION

In conclusion, this study has showcased the significant enhancement of data analysis by integrating the K-Means algorithm with the Quantum Particle Swarm Optimization approach and the proposed Vector Space Function Clustering Algorithm (PVSFCA). As data generation becomes increasingly prevalent, particularly through computer programs that enable forecasting and future insights, the role of Machine Learning has emerged to interpret and predict from diverse data inputs. The popular and efficient K-means algorithm addresses clustering, a vital component in various applications. This research has provided an overview of the latest scientific advancements in this domain. The evolution, limitations, and applications of the K-means algorithm have been explored, emphasising the continuous efforts to enhance cluster efficiency and accuracy, especially for vast datasets. However, a common challenge in applying clustering techniques to extensive datasets is the lack of consensus on attribute definitions and formal classification. To counter this challenge, this paper delves into concepts and techniques pertinent to clustering. It concisely evaluates existing theoretical and empirical algorithms while presenting a comparative analysis to demonstrate their effectiveness. The proposed Vector Space Function Clustering Algorithm (PVSFCA) has been proven swift and efficient in its analytical process. The findings validate the algorithm and establish a robust reference for implementing efficient clustering improvements. Through a meticulous assessment of characteristics such as iteration count, error rate, and the optimal arrangement of cluster centres, the proposed method serves as a foundation for effective cluster optimisation.

REFERENCES

[1]. Jayamalini, K., and M. Ponnavaikko. “Research on web data mining concepts, techniques and applications.” In 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), pp. 1-5. IEEE, 2017.

[2]. Gera, Mansi, and Shivani Goel. “Data mining-techniques, methods and algorithms: A review on tools and their validity.” International Journal of Computer Applications 113, no. 18, 2015.

[3]. Sisodia, Deepti, Lokesh Singh, Sheetal Sisodia, and Khushboo Saxena. "Clustering techniques: a brief survey of different clustering algorithms." International Journal of Latest Trends in Engineering and Technology (IJLTET) 1, no. 3: 82-87, 2012.

[4]. Zou, Hailei. "Clustering algorithm and its application in data mining." Wireless Personal Communications 110, no. 1: 21-30, 2020.

[5]. Mythili, S., and E. Madhiya. "An analysis on clustering algorithms in data mining." International Journal of Computer Science and Mobile Computing 3, no. 1 334-340, 2014.

[6]. J. Liu, L. Han and L. Hou, "K-Mean Clustering Algorithm Based on Particle Swarm Optimisation," System Engineering Theory and Practice, vol. 06, pp. 54-58, 2005.

[7]. Bai, Lili, Zerui Song, Haijie Bao, and Jingqing Jiang. "K-Means clustering based on improved quantum particle swarm optimisation algorithm." In 2021 13th International Conference on Advanced Computational Intelligence (ICACI), pp. 140-145. IEEE, 2021.

[8]. Salem, Semeh Ben, Sami Naouali, and ZiedChtourou. "A fast and effective partitional clustering algorithm for large categorical datasets using a k-means based approach." Computers & Electrical Engineering 68, 463-483, 2018.

[9]. Patibandla, RSM Lakshmi, and N. Veeranjaneyulu. "Survey on clustering algorithms for unstructured data." In Intelligent Engineering Informatics, pp. 421-429. Springer, Singapore, 2018.

[10]. Rezaee Jordehi, Ahmad, and Jasronita Jasni. "Particle swarm optimisation for discrete optimisation problems: a review." Artificial Intelligence Review 43: 243-25, 2015.

[11]. Poli, Riccardo. "Analysis of the publications on the applications of particle swarm optimisation." Journal of Artificial Evolution and Applications 2008: 1-10, 2008.

[12]. Arai, Kohei, and Ali Ridho. "Hierarchical K-means: an algorithm for centroids initialisation for K-means." Reports of the Faculty of Science and Engineering, Vol. 36, No.1, 2007.

[13]. R. Chouhan and A. Purohit, "An approach for document clustering using PSO and K-means algorithm," in 2018 2nd International Conference on Inventive Systems and Control (ICISC), IEEE, pp. 1380-1384, 2018.

[14]. R. Janani and S. Vijayarani, "Text document clustering using spectral clustering algorithm with particle swarm optimisation," Expert Systems with Applications, vol. 134, pp. 192-200, 2019.

[15]. Barakbah, Ali Ridho, and Kohei Arai. "Centronit: Initial Centroid Designation Algorithm for K-Means Clustering." EMITTER International Journal of Engineering Technology 2, no. 1: 50-62, 2014.

[16]. Caron, Mathilde, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. "Deep clustering for unsupervised learning of visual features." In Proceedings of the European Conference on Computer Vision (ECCV), pp. 132-149. 2018.

[17]. Sunita Kumari, Abha Kaushik, "A Survey on Clustering Problem with Optimised K-medoid Algorithm", ISSN: 2348-4098 Volume 02 ISSUE 04 April-May 2014.

[18]. Preeti Arora, Dr. Deepali, Shipra Varshney, "Analysis of K-Means and K-Medoids Algorithm For Big Data", International Conference on Information Security & Privacy (ICISP2015), 11-12 December 2015.

[19]. [19] Megha Mandloi, " A Survey on Clustering Algorithms and K-Means", International Journal of Research in Engineering Technology and Management ISSN 2347 – 7539.

[20]. Khaled A. Alenezi, Mohammad A. Alahmad, Walid Aljoby, "Propose Parallelization of K-Medoid Clustering Algorithm", Journal of Advanced Computer Science and Technology Research, Vol.4 No.4, 101-107, December 2014.