International Journal of Research in Engineering, Science and Management (IJRESM)
www.ijresm.com *Volume-1, Issue-3, March 2018*

Examining the Performance of K-Means Clustering Algorithm

Pascal Maniriho[1], Ari Effendi[2]

[1,2]*Department of Informatics, Institut Teknologi Sepuluh Nopember, Campus ITS Keputih Sukolilo, Surabaya, Indonesia, 60111*

***Abstract*—Clustering process involves finding the homogeneity between data or objects and group them into different categories based on their similarity levels. That is, the dissimilarity between objects from different categories is high. In this paper, a brief overview on the existing clustering methods is given. Additionally, the performance of the well-known clustering algorithm namely K-Means algorithm is examined. This algorithm is implemented and evaluated using Fisher's iris data set available at UCI online data repository. The performance is evaluated in terms of execution time and the number of instances that are miss-clustered. During the experiment, the data normalization is performed and the confusion matrix is further utilized to compare the predicted cluster and the actual class of each instance.**

Index Terms— Clustering, instance, k-Means, confusion matrix, data normalization, cluster centroid, Data point

I. Introduction

With the development of the internet and the introduction of new technologies, a huge amount of data is being generated by human being, making it to increase at a large scale as well as causing its mining and manipulation to be more complex. Moreover, new challenges and requirements are being brought by this massive growth of data which creates research in many different areas. To effectively and efficiently examine structures in the data, clustering techniques are employed. However, the most challenging question is how these data can be mined and get clustered into coherent groups. Nevertheless, many research have been already carried out to address this issue. A study on K-means, FKM and IRP-K-means clustering techniques were conducted on image data and the performance of these three algorithms was further investigated [1].

To detect social media data from the community users, a new scheme combining K-Means and Genetic Algorithm was developed [2]. The main purpose of their framework was to cluster these social data by providing the best way for initializing the cluster centroid. Besides, the optimization technique was used in order to get clusters that are accurate and three factors (‘’attitudes scores, leadership and follower’’) were considered during the clustering process. In order to overcome the problem of convergence rate and global search related to K-Means, Lashkari and Hussein [3] introduced a new technique that achieves the best solutions. To evaluate the performance of K-Means, the analysis of the patient records was done in [4] in order to group patients based on their risk levels of epilepsy from EEG signals. During their study, the obtained results were further compared with the one from KNN classifier and K-means performed well over KNN. The data set having small number of instances that are labelled was used during the semi-supervised experiment that was elaborated in [5]. The enhancement of K-means algorithm was done by employing the concept of Parallelism in both CPU and CUDA [6]. Further exploration on clustering methods was carried out in the research presented by Praveen and Rama [7]. Their study was extended by elaborating how well K-Means does perform on the given data set as well as analysing how it gets affected by the selection of the initial seed. Besides, K-means was applied in categorizing spatial data based on Hadoop [8]. To improve the membership of objects to be assigned into clusters, a new approach which compares K-Means and Fuzzy C-means was designed [9]. With this approach data objects can be assigned into clusters based on their degree of belongingness which is highly depending upon the selected fuzziness factor. The analysis of multidimensional data related to students' performance was achieved by adapting K-Means clustering approach via R-statistical tool [10].

The following points will be covered on the next parts of this paper: Clustering concept, K-Means clustering technique, results from the experiment, the results discussion and finally conclusion and future work will be presented at the end.

II. Clustering

A. Clustering Concept

Clustering is considered as the well-known unsupervised learning technique that deals with data that are unstructured. Given unlabeled data set, it aims at categorizing the data objects into different groups (also known as clusters) based on some similarity measures like distance between data points, characteristics that describes objects etc. That is, all the data objects are assigned into groups in such a way that the similarity between data belonging into the same group is high. Notice that the formed groups are called clusters and they are said to be compact, if the intra-cluster similarity is high while the inter-cluster similarity is low. Fig. 1 illustrates the clustering process.

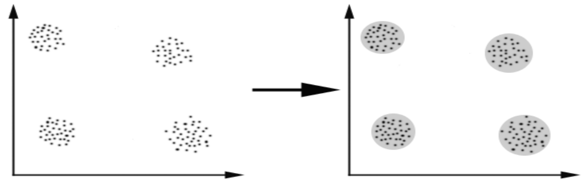

Fig. 1. Data objects (a) Before clustering (b) After clustering

B. Clustering Concept

Several clustering approaches have been already presented in the literature. Besides, since most of these approaches may have

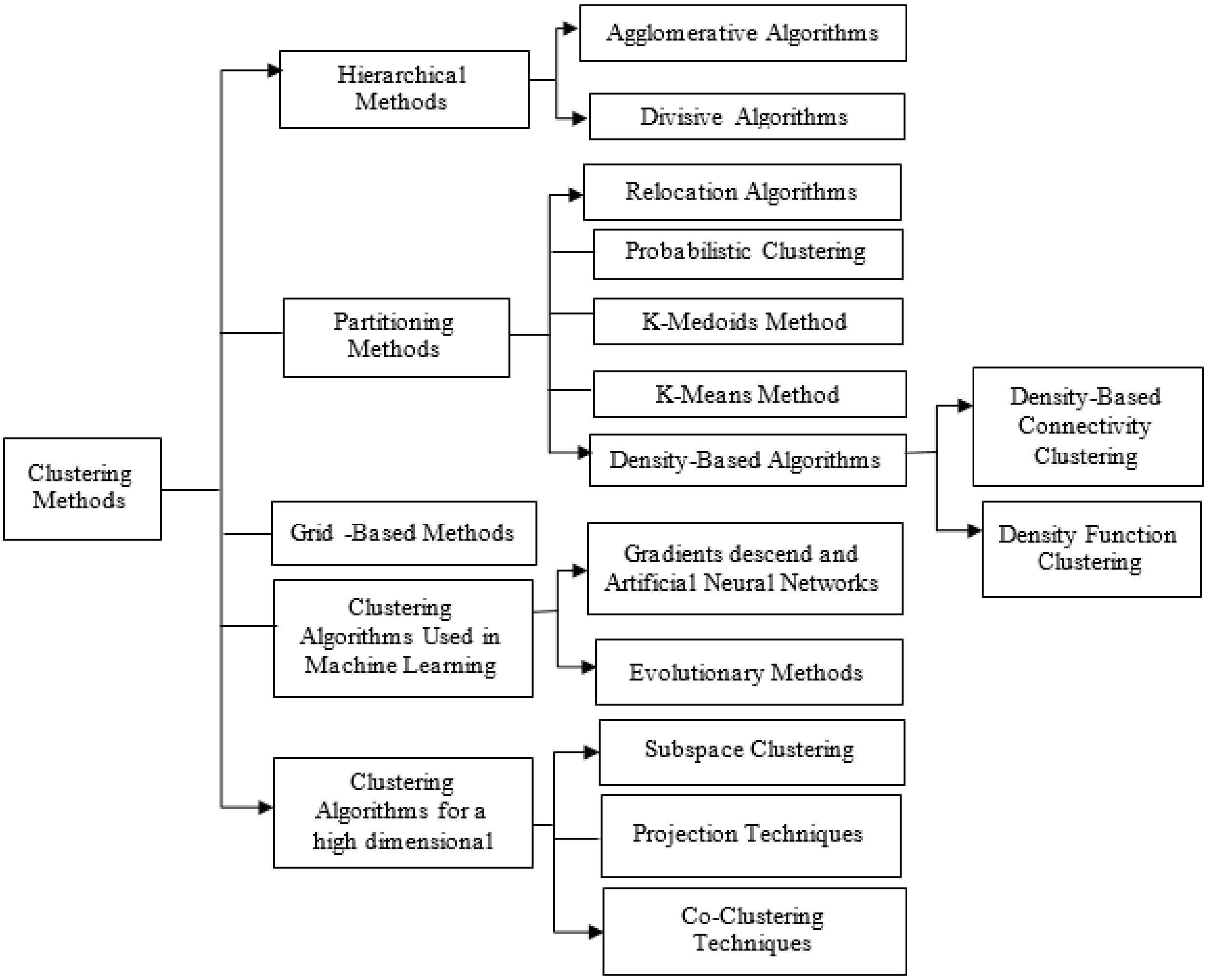

Fig. 2. Clustering Algorithms categorizations

some common features, providing a crisp categorization is somehow challenging.

Nonetheless, it is ideal to illustrate some of the existing clustering approaches. Hence, Fig. 2 above provides various clustering techniques.

III. K-Means Clustering

K-means Algorithm is among the most famous, fast and robust unsupervised clustering methods [14][15]. Additionally, it falls under the category of partitioning methods. By employing K-Means method, data objects can be clustered as follows. Given a set of data points X, based on some defined criterions K-Means groups them into different coherent groups. The main goal of this algorithm is to group data by maximizing the similarity between the data objects belonging to the same group while minimizing the similarity between data objects that belongs into different groups or clusters. K-Means performance is greatly influenced by the initial cluster centroids which are chosen randomly.

Overall, the functionality of K-Means algorithm can be summarized as follows.

1. First, decide K number of clusters
2. Randomly select K data objects as the initial cluster centroids.
3. Compute the distance between each data object and cluster centroid.
4. Assign each data object to its closest centroid.
5. Recalculate each cluster centroid.
6. Is the centroid's position still changing? if yes,
7. Repeat step 3-5 until the last iteration is reached. Notice that at this stage the convergence has been reached.

The convergence means that the data points' clusters are no longer changing from one iteration to another. Besides, with regard to the distance, Euclidean distance is mostly used. The entire process of K-means algorithm is depicted in Fig. 3.

IV. Methodology

During the experiment the Fisher's iris data set available at UCI machine learning data repository is utilized to evaluate the performance of K-means clustering approach. It is worth mentioning that two different phases are considered in the experiment. Furthermore, it is also important to note that before feeding the data set to the clustering algorithm, the data set is first normalized (second phase). The main reason of this normalization is that when we look at the distribution of the iris data set we could see that the data points are somehow very dispersed which can affect the performance of the algorithm. Owing to this matter, it is ideal to bring all the data points into the same interval which is between zero and one in order for the data to be more scalable. The distribution of the data points (form iris data) can be viewed from Fig. 3 below. In addition, MATLAB programming language is utilized to implement and evaluate the performance of this algorithm.

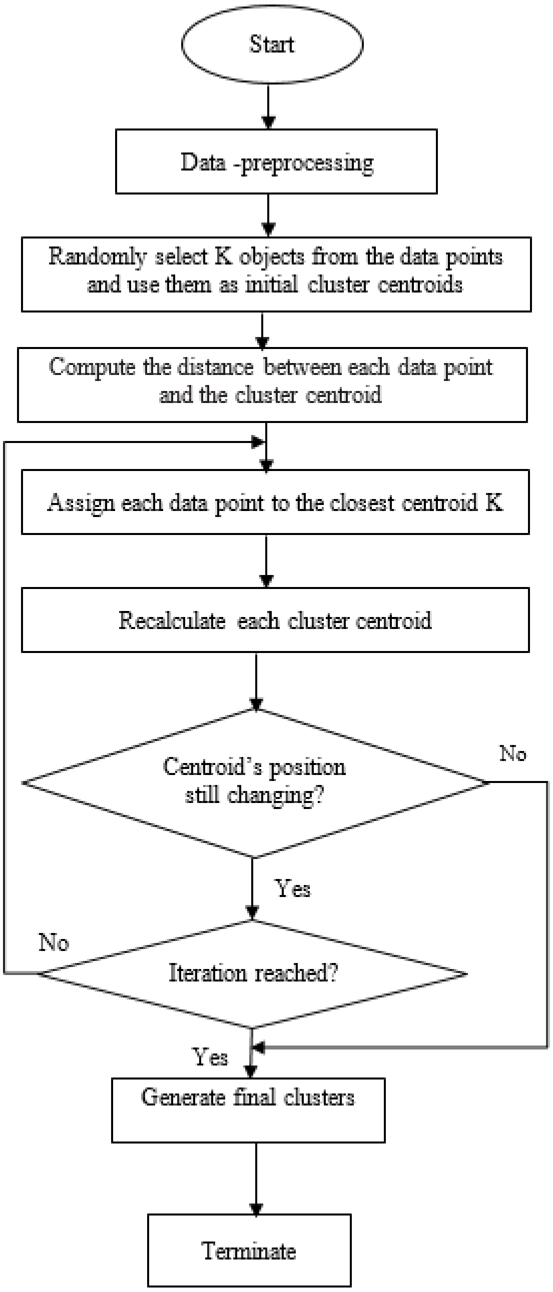
Start
Data -preprocessing
Randomly select K objects from the data points and use them as initial cluster centroids
Compute the distance between each data point and the cluster centroid
Assign each data point to the closest centroid K
Recalculate each cluster centroid
Centroid's position still changing?
No
Yes
No
Iteration reached?
Yes
Generate final clusters
Terminate

Fig. 3. K-Means Clustering Process

V. Experimental Results

A. Phase1: Clustering data using unlabeled data set

During this phase, the labels are first removed before feeding the data point to the algorithm, thereafter all instances are clustered into two clusters. Additionally, the number of instance in each cluster and the running time are further recorded and the results are shown in Table I.

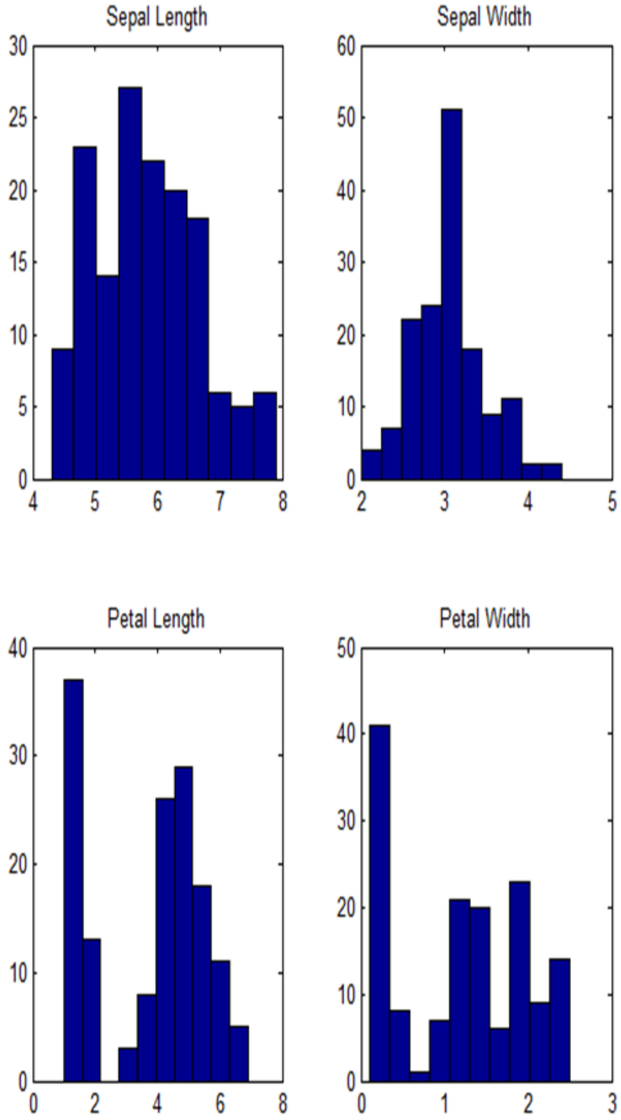

Fig. 4. Fisher's Iris Data visualization

TABLE I
CLUSTERING USING K-MEANS ALGORITHM

Clusters	Number of instances clustered in each cluster	Running time in second
Cluster 1	50	3.534608
Cluster 2	100	

B. Phase 2: Clustering using labeled data set

In this phase, the labelled data is used. Furthermore, the performance is measured and evaluated using the confusion matrix. The confusion matrix is employed in order to keep track of the actual class and predicted cluster of each instance (or data point). Note that the running time is also investigated and the results obtained are compared with the ones presented in [16].

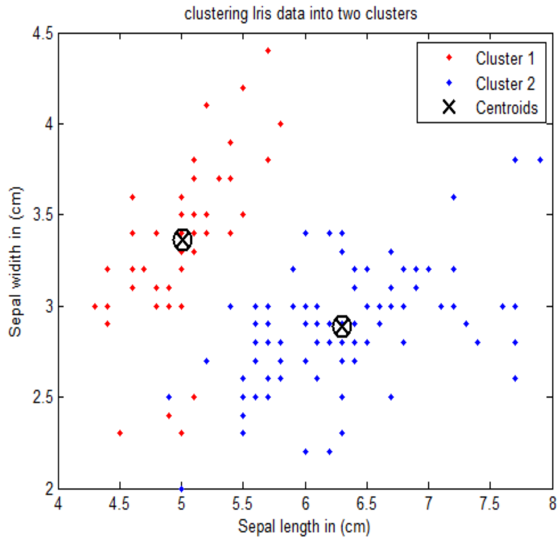

Fig. 5. Separation of clusters obtained using K-means before normalizing the data set

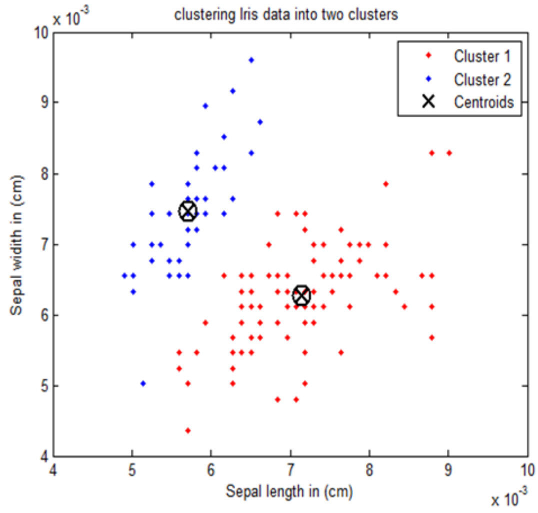

Fig. 6. Separation of clusters obtained using K-Means Algorithm after Normalizing data

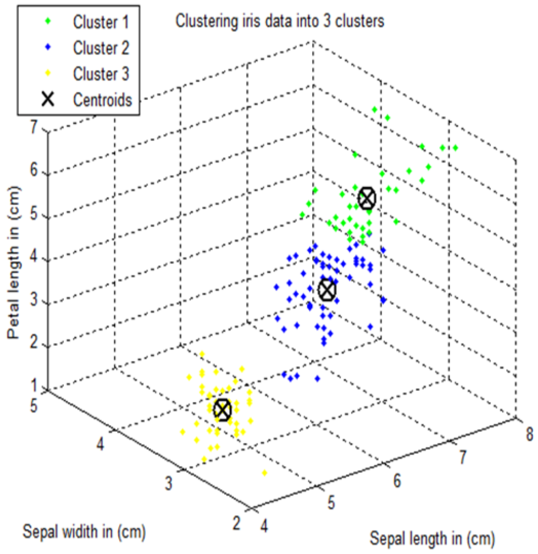

Fig. 7. Clusters obtained after applying K-Means algorithm using non-normalized the data

TABLE II
RESULTS FROM THE CONFUSION MATRIX OBTAINED USING K-MEANS ALGORITHMS DURING THE EXPERIMENT CARRIED OUT IN [16]

Cluster	The species of the iris		
	Setosa	Versicolor	Virginica
Cluster 1	0	3	36
Cluster 2	0	47	14
Cluster 3	50	0	0

TABLE III
RESULTS FROM THE CONFUSION MATRIX OBTAINED USING UN-NORMALIZED DATA SET DURING OUR EXPERIMENT

Cluster	The species of the iris			
	Setosa	Versicolor	Virginica	Running time in second
Cluster 1	50	0	0	1.433743
Cluster 2	2	48	0	
Cluster 3	0	14	36	

TABLE IV
RESULTS FROM THE CONFUSION MATRIX OBTAINED USING NORMALIZED DATA SET DURING OUR EXPERIMENT

Cluster	The species of the iris			
	Setosa	Versicolor	Virginica	Running time in second
Cluster 1	50	0	0	1. 573070
Cluster 2	0	48	2	
Cluster 3	0	4	46	

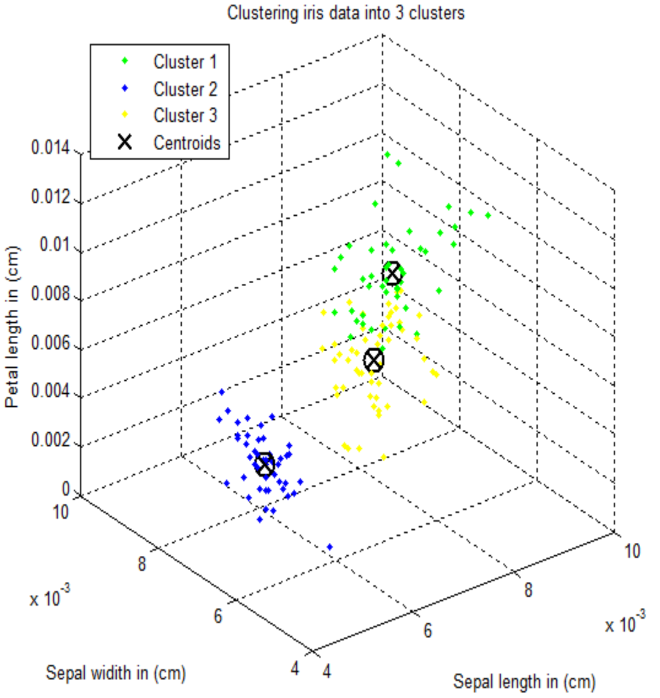

Fig. 8. Clusters obtained after applying K-Means algorithm using non-normalized the data

In addition, the separation of clusters presented in Table III and Table IV is depicted in Fig. 7 and Fig. 8 above.

VI. Results and Discussion

In the experimental results presented above, 150 instances from Fisher's iris data set were clustered into different clusters. Table I shows the results obtained after clustering data into two clusters using unlabeled data set and the separation between these two clusters is depicted in Figs. 5 and 6. Table II shows the experimental results generated during the experiment carried out in [16] while table III and IV present the results from our experiment. The confusion matrix demonstrates the number of instances that are clustered in each cluster.

If we look at the results from Table III, we could see that by using unnormalized data, 50 instances of iris Setosa are well clustered, 48 instances which are also well clustered as versicolor while 2 instances are misclustered as Setosa. In addition, 36 instances are correctly clustered as Virginica while 14 are misclustered as Versicolor. However, Table IV shows that after normalizing the data set, the number of misclustered instances is greatly reduced. That is, good results are achieved while using the normalized data set. Figs. 7 and 8 depicts the separation of three clusters for both cases (using the normalized, Table III and un-normalized dataset, Table IV). Generally, good results are achieved compared to the ones in [16].

VII. Conclusion and Future Work

Clustering is unsupervised learning technique that is employed to find a well structure on a given unstructured data objects. In this Paper, we have presented a brief overview on the existing clustering approaches, thereafter the performance of K-Means clustering algorithm was examined by clustering the most famous data set namely Fisher's iris data set. Different metrics such as the running time and number of miss-clustered instances were investigated. Additionally, the data normalization was carried out due to its feature of bringing the attributes' values into the same interval which is ideal for data objects whose distribution is very dispersed.

Besides, the performance was evaluated using both normalized and non-normalized data sets to see the impact on the generated clusters. The confusion matrix was utilized to get the number instances that are miss-clustered. Since several clustering algorithms have been already proposed by many researchers, this study will be extended to comparing the performance of K-Means with other clustering algorithms in the future work.

References

[1] S. Bettoumi, C. Jlassi, and N. Arous, "Comparative Study of k-means Variants for mono-view clustering," in *International Conference for Signal and Image Processing -ATSIP*, 2016, pp. 183–188.

[2] A. Alsayat, "Social Media Analysis using Optimized K-Means Clustering," in IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA), 2016.

[3] M. Lashkari and M. Hossein Mottar, "The Improved K-means Clustering algorithm using the proposed Extended PSO algorithm," in *International Congress on Technology, Communication and Knowledge (ICTCK),* 2015, no. Ictck, pp. 11–12.

[4] M. Manjusha and A. E. E. G. D. Acquisitiom, "Performance Analysis of KNN Classifier and K-Means Clustering for Robust Classification of Epilepsy from EEG Signals," in *International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET),* 2016, pp. 2412–2416.

[5] X. Cui and F. Wang, "An Improved Method for K-Means Clustering," in *International Conference on Computational Intelligence and Communication Networks An*, 2015, no. 1.

[6] M. Baydoun, M. Dawi, and H. Ghaziri, "Enhanced Parallel Implementation of the K-Means Clustering Algorithm," in *2016 3rd International Conference on Advances in Computational Tools for Engineering Applications (ACTEA)*, pp. 7–11.

[7] P. Praveen and B. Rama, "An Empirical comparison of Clustering using Hierarchical methods and K-means Cluster Analysis Types :," *2nd Int. Conf. onAdvances Electr. Electron. Information, Commun. Bio-Informatics*, pp. 1–5, 2016.

[8] Y. Zhong and D. Liu, "The Application of K-Means Clustering Algorithm Based on Hadoop," in *EEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA),* 2016, pp. 88–92.

[9] C. T. Baviskar and S. S. P. Associate, "Improvement of Data Object's Membership by using Fuzzy K-Means Clustering Approach," in *International Conference on Computation of Power, Energy Information and Communication (ICCPEIC) Improvement*, 2016.

[10] Agnivesh and R. Pandey, "Elective Recommendation Support through K-Means Clustering using R-Tool," in *2015 International Conference on Computational Intelligence and Communication Networks Elective*, 2015, pp. 851–856.

[11] "Clustering-Introduction."[Online].Available: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/index.html. [Accessed: 19-Jan-2017].

[12] "Clustering."[Online].Available: http://www.slideshare.net/mrizwanaqeel/clustering-54063985. [Accessed: 02-Feb-2017].