International Journal of Civil Engineering and Technology (IJCIET)
Volume 9, Issue 2, February 2018, pp. 705–714, Article ID: IJCIET_09_02_067
Available online at http://www.iaeme.com/ijciet/issues.asp?JType=IJCIET&VType=9&IType=2
ISSN Print: 0976-6308 and ISSN Online: 0976-6316

AN EFFICIENT APPROACH TOWARDS CLUSTERING USING K-MEANS ALGORITHM

Suraj Nair Aiyappa, Ramamurthy B
Department of Computer Science,
CHRIST (Deemed to be University), Bangalore, India

ABSTRACT

Cluster analysis is one of the major knowledge mining methods in the field of data analytics; the approach used for clustering will influence the accuracy of the results and quality of the obtained clusters. A good clustering process or algorithm is one which increases the fit of the data points in each cluster and which satisfies the clustering criteria, if these measures are not met adequately the desired pattern will not be seen and the patterns obtained for analysis may turn out to be inaccurate or insufficient. This paper discusses the standard k-means clustering algorithm and provides an efficient approach towards clustering using the standard global K-means algorithm; the process eliminates the need for initializing random number of clusters multiple times which is followed as the standard process in the field. The effectiveness of the proposed approach was analyzed using the benchmark dataset and the implementation was performed using the well-known analytic tool R Studio and supporting packages.

Key words: Partition clustering, Distance Measure, Hopkins Index, Elbow method, Gap-statistic method, standard K-means process.

Cite this Article: Suraj Nair Aiyappa, Ramamurthy B, An Efficient Approach Towards Clustering Using K-Means Algorithm, International Journal of Civil Engineering and Technology, 9(2), 2018, pp. 705–714.
http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=9&IType=2

1. INTRODUCTION

Cluster analysis is one of the most conventional and studied data analysis technique majorly used in field like recognition of hidden patterns in the data, data compression, image processing and various other fields[1]. It is a process that divides the data into many sub-clusters and these clusters exhibit a property of high intra-cluster similarity and low inter –cluster similarity, i.e. the data objects or data points within a cluster have high similarity and data points in one cluster when compared to the points in another cluster have very low similarity [2]. The aforementioned property of clustering is what defines the quality of the clusters and efficiency of the clustering algorithm that will be used. Clustering is considered as an unsupervised learning technique as in this case, the method brings out any hidden patterns and relation between the data points unlike supervised learning where classes are defined and the process of

classification is performed based on the class labels [3]. The process of clustering can be subdivided broadly into two categories, i.e. clustering based on the method of partitioning and based on hierarchy known commonly as hierarchical clustering [4]. The process in which the data sets are divided or merged known as divisive and agglomerative approaches, according to the similarity of the data points, is known as hierarchical clustering method and the outcome of the above process can be represented in a tree like structure known as a dendrogram. This method of clustering requires a lot of memory space and time to process high dimensional data.

The other type of approach towards clustering of data is known as partitioning method wherein the data points are partitioned based on the given input of the number of clusters and an outcome consists of clusters with high cohesion and low coupling. The most famous and basic algorithmic approach towards partition clustering is performed using the global K-means algorithm [5], a recursive process that takes as input, from the user; the dataset consisting of *n* objects and the initial number of clusters *k*. It then partitions the set of *n* objects into *k* clusters by calculating the distance between the data points and the cluster centres and assigning them to the nearest cluster centre, various distance measures can be used for the process [6]. At each step a new set of cluster centres are recalculated. The algorithm keeps calling itself recursively until and unless there is no change in the centroids. The main goal of this algorithm is to minimize the function of squared errors as shown in Eq. (1)

$$Z= \sum_{i=1}^{j} \sum_{i=0}^{n} ||x_i^j - c_j|| \quad (1)$$

Where X_i *is* the data point and C_j is the respective cluster center.

The basic flow of the algorithm is represented as a floe diagram in Fig. 1.

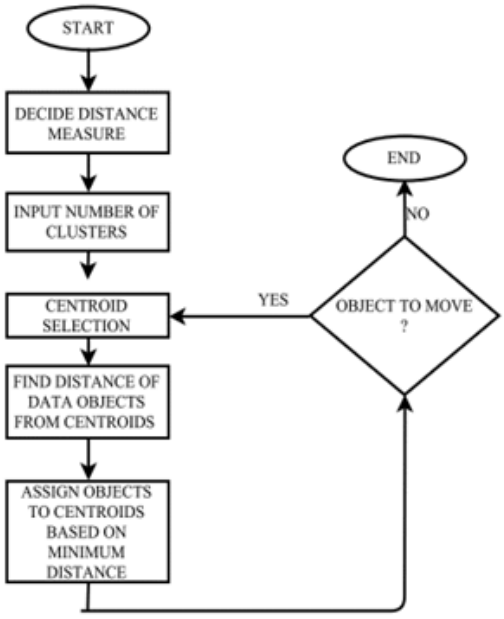


Figure 1 Flow of the K-Means Algorithm

The K-means algorithm has been considered as the benchmark partition-clustering algorithm as it is easy to understand as well as implement. However, the major drawbacks of this algorithm are the algorithm takes as input the initial number of clusters *k*, which will lead to, repetitive running of the algorithm with various *k* as input, that can lead to wastage of time and increased complexity. Another drawback is that it is susceptible to outliers as is considers the mean of the cluster to assign centroids and if the data set, consists of objects that are highly variated, it can cause irregular cluster assignments.

The k-means algorithm was chosen, as it is easy to implement and understand for novice users, the proposed approach in this paper can be utilized in conjunction with other clustering algorithms that has the drawback of initializing number of clusters before the application of the algorithm. The solutions to the above drawbacks are discussed in this paper by providing an efficient approach towards clustering using the k means algorithm. K-means algorithm has been taken into consideration, as it is the most basic algorithm that is available for clustering.

The rest of the paper discusses about few related works on this algorithm and proceeds towards the proposed approach and a brief explanation on the implementation of the approach.

2. RELATED WORK

K-means is a flexible and easy to implement algorithm, there were many approaches, which were performed by many research scholars that aim at increasing the proficiency and effectiveness of the k-means algorithm. Few of the works by different scholars are discussed, in the published writing [7], an enhanced K-means is proposed which makes use of uniform distribution of the data objects. It also follows a specified approach for selection of optimal number of clusters which is passed on to the second juncture of the algorithm wherein the data objects are assigned to the specific clusters based on the distance measure utilized. This approach was experimentally proved to provide accurate clustering results with a reduced time complexity when a high dimensional data is considered. Another literature [8] also provides an improved version of the K-means algorithm that has improved the clustering accuracy by eliminating the drawback of repeatedly calculating the centres. The proposed method was implemented by making use of two simple data structures that store the distance of the datum from the cluster centres and use it in the next iteration by comparing it to the calculated cluster centre. In the work [9], an interesting method known as 2 layer k-means is proposed which according to F-measure has shown an impressive accuracy when compared to other algorithms. The process is divisive where each cluster has been partitioned into several clusters and then combines them utilizing the global k-means algorithm; this method has ensured that any data point that may lie in between two clusters if the standard k-means is applied is also clustered accurately in this approach. Scholars have also approached to increase the efficiency of K-means algorithm by making use of optimization algorithms that was designed by observing the natural behaviour of animals in the environment. One such algorithm is the artificial bee colony algorithm most commonly known as ABC algorithm [10], applied to clustering by understanding and implementing the concept of dance area, and food foraging behaviour that bees use for gathering food and communicating of the optimal location of honey. The accuracy is measured using the classification error percentage or the CEP that signifies the error percentage of classification that has shown a lower error percentage for the ABC algorithm when compared to the PSO (Particle Swarm optimization) algorithm. Considering the domain specific approaches in the field, an approach to classify loyalty of a customer to a product a method has been proposed [11] by including few additional criteria's i.e. joining WRFM-based method to K-means algorithm applied in data mining with K-optimum according to Davies–Bouldini Index. The paper has proposed a new procedure, based on the expanded RFM model to classify customer product loyalty. The proposed method has been implemented for an industry named SAPCO Co. Based in Iran, which showed success than the other companies in Iran that commonly used random selection. This paper provides an efficient and effective methodology to be used for implementing the firm's objective. The authors referred to the RFM model and the k means algorithm of clustering to propose this paper. An application of clustering has also been studied for pattern discovery from web usage data [12] and designing an application this paper proposes an up-to-date survey of Web Usage mining, including academic and commercial efforts applied to research. This paper applies the concepts of clustering in pattern discovery (ex: clustering of pages will find groups of pages having similar

content). From the literature review performed, it is noticed that the above papers have utilized benchmark datasets from the UCI repository of machine learning has been used for running the algorithm on and to establish the accuracy of the proposes approach.

3. PROPOSED APPROACH

The standard k-means algorithm has many drawbacks but is still the most preferred algorithm to understand the basics of clustering and it provides accurate results with an addition of fine-tuning like enhancing the efficiency of the algorithm or by specifying an approach that eliminates the drawbacks in a systematic manner.

The drawbacks that are addressed in this paper are, the algorithm takes as input the initial number of clusters *k*, which will lead to, repetitive running of the algorithm with various *k* as input, that can lead to wastage of time and increased complexity. The algorithm is also susceptible to outliers as K-means algorithm considers the mean of the cluster to assign centroids and if the data set consists of objects that are highly variated, it can cause irregular cluster assignments. This paper provides an efficient approach towards the process of clustering by integrating many methods that eliminates the aforementioned drawbacks. Diagrammatic representation of proposed approach is as follows.

The different phases of the proposed approach are to be followed to eliminate the drawbacks of the global K-means algorithm. The phases of the proposed approach is represented in Fig. 2, and a brief explanation of each of the phases is provided.

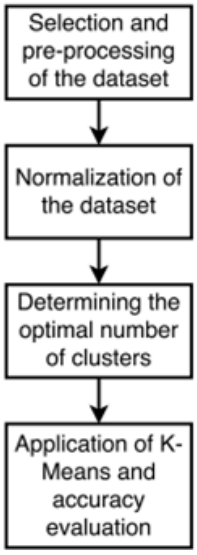


Figure 2 The proposed Approach.

Phase 1- Selection and pre-processing of the dataset

Pre-processing of data is an important aspect in the process of knowledge mining. The dataset acquired for analysis may contain a lot of noise or anomalies that may cause irregularities while execution of the algorithm and yield inaccurate results. To eliminate this, the datset must be pre-processed and cleaned with accurate and appropriate methods that is apt for the dataset, like removal of missing values and conversion of data from categorical to nominal/ordinal as K-means algorithm works well with only numerical data. When high dimensional data is taken into consideration, dimensionality reduction methods like principal component analysis and multi-dimensional scaling can be employed to reduce the dimensions and not to lose the entire meaning of the data.

Phase 2- Normalization of the dataset

Normalization is the process where the data objects are scaled down to a particular range. Based on the users or on the measures of dispersion, this phase can be considered as a step in pre-processing, but is considered as another phase here as it is one of the major steps to eliminate the effect of outliers by scaling the data down to a lower range. The normalization technique used in the implementation of this approach is the z-score normalization that scales down the value of the data points using its mean and standard deviation.

$$\boldsymbol{Y = (data\_object - \mu)/\sigma} \quad (2)$$

Where, μ stands for the mean of the data column and σ signifies standard deviation.

Phase 3- Deduce the optimal number of clusters

There are many methods in literature for ascertaining the optimal number of clusters but the two most prominent and efficient methods that can be utilized are elbow method and the gap statistic method. An overview of the methods is as follows.

The elbow method [13] takes into consideration, the within sum of squares and plotting a curve of the within sum squares in contrast to the number of clusters and the elbow of the curve in the plot is the exact location that signifies the number of clusters that is optimal and can be chosen for the next phase. The point that signifies the apt clusters must be chosen in a way such that adding another cluster will not increase the total within sum of squares.

The gap statistic method, collates the total intra-cluster dissimilarity for various values of k with their assumed values under null reference distribution of the data [14]. The value of amount of clusters that is most favourable will be that which increases the gap statistic measure (i.e., the value that produces the biggest gap statistic). The equation to calculate the gap statistic is mentioned in Eq. (3)

$$\boldsymbol{Gap(k) = \frac{1}{X}\sum_{x=1}^{X} \log(W_{kx}) - \log(W_k)} \quad (3)$$

Where, k is the number of clusters, X is the set of reference data chosen, W_{kx} is the intra cluster distance of k with respect to the reference datset X and W_k is the intra cluster distance in reference to the data set chosen.

Phase 4- Application of K-means and accuracy evaluation

The final phase of the proposed approach is to apply the k-means algorithm by utilizing the data from the previous phase. The accuracy/percentage of fit can be estimated by taking the correlation of between sum of squares against the total sum of squares.

The above process has been proven effective and yields optimal results. If the dataset is not fit for applying the process that is described above, then it would be a waste of time and computational resources .To eliminate this drawback, a method known as the Hopkins index [15] can be utilized which gives a numerical value that depicts the clustering tendency of the dataset. Utilizing the value, it can be deduced if the dataset is fit for clustering or whether, it must be considered for some other data analysis technique

4. EXPERIMENTAL RESULTS

The validity of the proposed approach is proved using the well-known statistical computing language-R, an integrated development environment known as R studio [16] is used for the implementation of the functionalities of the programming language. The tool is well known for its flexibility in allowing users to program their own procedures and provides a wide array of packages that can be used to make the approach efficient and less time consuming.

R studio provides many inbuilt packages that aid in the statistical programming of the functionalities; few of these packages are used for the implementation of the proposed approach.

4.1. Packages utilized

For the phase of normalization a package known as clustersim is used that provides a function used for normalization of the data, the function takes as parameters; the data frame which has been reprocessed and a value ranging from n0 to n13 where each value stands for a different type of normalization like positional standardization, standardization, positional quotient transformation etc. For utilization of the Z-score normalization technique, the function is as described below.

Z=data.Normalization (datum, type="n1", normalization="column")

The phase of determining the number of clusters which is optimal for the chosen dataset was performed using the package, Factoextra [17], that helps in a visualization of the results and Nbclust [18] package, it provides over 30 indices to predict the optimal number of clusters and out of which two indices are chosen, the elbow and the gap statistic method respectively.

4.2. Datasets utilized

As discussed in the literature review it was found that most of the literatures used the standard benchmark datasets drawn from an online repository of machine learning [19] for establishing the accuracy of their respective enhanced algorithms and few other published works also utilized the benchmark datasets [20,21].

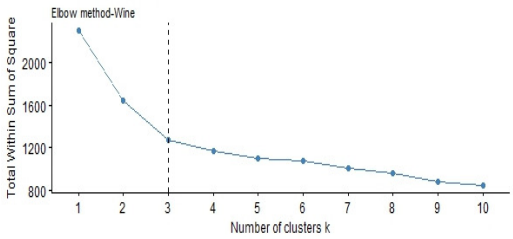
Table 1 Dataset Information

Dataset	Classes	Instances	Features
E.COLI	4	336	7
WINE	3	178	13
GLASS	7	214	11

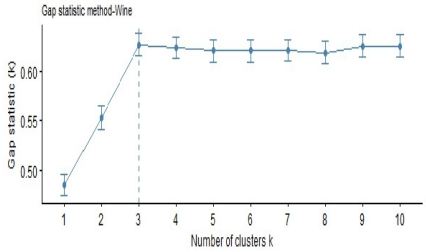
As seen in the above table, the datasets utilized have a predefined set of classes available. These datasets are stripped of their class column during the stage of pre-processing and then subjected to the further phases of the proposed approach. The output of phase 3.i.e. the determination of the apt number of clusters must produce results that are close to or equal to the number of classes in the dataset.

4.3. Results

The proposed approach was implemented on the dataset mentioned in Table 1 and is executed according to the phases described. The plots for the results of the respective datasets using both elbow method and gap statistic method have been depicted, as observed in the plots it can be seen that the results obtained from the plots is similar to the classes of each of the dataset mentioned in Table 1.



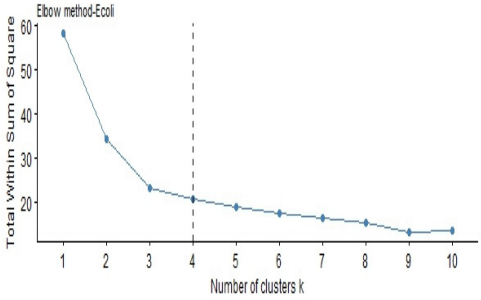
(a)



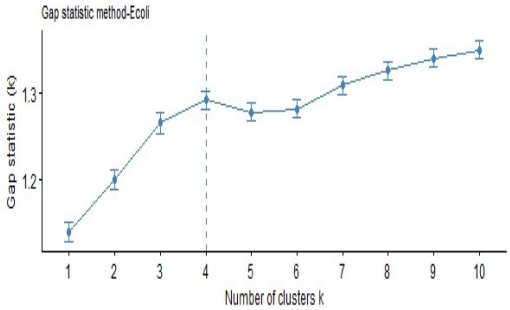
(b)

Figure.3 Obtained results: (a) Optimal clusters for wine dataset using elbow method, (b) Optimal clusters for wine dataset using Gap statistic method.

The above plots are obtained after the application of phase 3, it projects that the optimal number of clusters for the wine dataset is three, which is exactly equal to the number of classes available in the datset according to the repository.



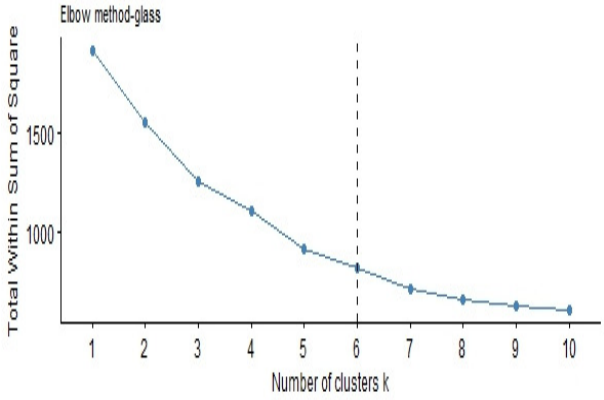
(a)



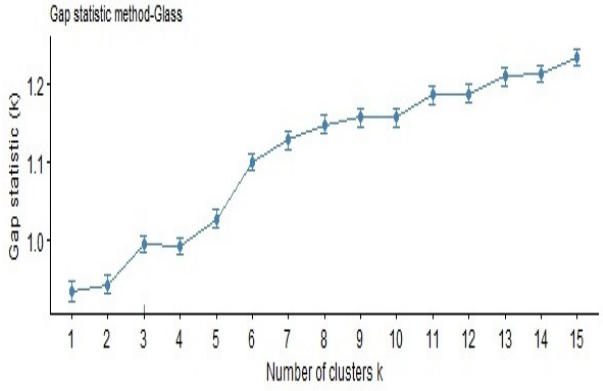
(b)

Figure.4 Obtained results: (a) Optimal clusters for the dataset E.coli using elbow method, (b) Optimal clusters for the dataset E.coli using gap statistic method

The above plots were obtained for the E.coli dataset where the apt number of clusters predicted by the gap statistic method is exactly equal to the number of classes available in the datset according to the repository. The plot obtained for the elbow method shows the elbow bend at 3 and the elbow straightens out at 4, therefore both 3 and 4 can be considered as the optimal number of clusters and a more accurate value can be considered after application of the K-means algorithm.



(a)



(b)

Figure 5 Obtained results: (a) Optimal clusters for the dataset Glass using Elbow method, (b) Optimal clusters for the dataset Glass using Gap statistic method.

It can be observed for the glass dataset the gap statistic as well that the elbow method predicts 6 clusters as the optimal value whereas the dataset has 7 classes. Nevertheless, it can also be observed that the line straightens out at seven, therefore both six and seven can be considered as the optimal number of clusters and the more accurate value can be considered after application of the K-means algorithm, similar is the case with the gap statistic method the graph starts stabilizing at seven number of apt clusters.

The cluster fit can be calculated using the sum of square error that calculates the compactness of the clustering process. The formula that has been used for calculating the fit is,

Percentage of fit =Between_SS/Total_SS (4)

Where, Between_SS stands for between cluster sum of squares and Total_SS stands for sum of squares of all the clusters, the equation indicates the fit of the data objects in each cluster.

Many other available measures can be used to find out the clustering accuracy, like Rand Index, F-measure etc. The above method was chosen to introduce to novice learners the basic property of cluster quality. Table 2 shows the application of the chosen method to the glass dataset.

Table 2 Average Fit for Glass Dataset

Dataset	**GLASS**
Random assignment of initial clusters	3,4,5
Average Fit	37.9%
Optimal cluster prediction by Elbow method	6,7
Average Fit	62%
Optimal cluster prediction by Gap statistic method	6,7
Average Fit	62%

It can be observed from Table 2 that the elbow and the gap statistic method predicts 6 and 7 as the optimal number of clusters for the chosen dataset as observed from Fig. 5. The percentage of fit that is displayed in Table 2 is the average fit of both of the values after executing the k-means with the obtained results from phase 3 the pre determination of optimal number of clusters is efficient and eliminates the drawback of initializing random number of clusters at the start of execution.

5. CONCLUSION

Clustering is widely used and researched as a data mining technique that gives any kind of user the basic insight and reveals interesting patterns in their data set used. There are many tools that provide the facility of clustering where the user need not or only needs a very basic knowledge regarding the field, and the most frequently used algorithm when clustering is considered is the K-means algorithm, although there are many other partition based algorithms like K-medoids, CLARA etc. K-means is most widely used, and this paper proposes an efficient approach towards clustering using this standard algorithm. Other algorithms can also be used which can guarantee an increased efficiency and decreased error rate. The paper was worked on keeping in mind the novice entries into the field of data analysis and this paper gently introduces them to the process of clustering and provides insights into various other terminologies and methods that the novice scholars can refer to for further detailed study. The proposed methodology is easily understandable. A sincere drawback to this work is that each of the steps in the proposed approach has to be performed separately. In future, it would be more efficient if the phase 3 and phase 4 suggested in this work can be combined together for the application of K-means algorithm.

REFERENCES

[1] M.R. Anderberg, “Cluster Analysis for Applications, *Academic Press, New York, NY*, 1973.

[2] A.K. Jain et al, “Data Clustering: A Review”, *ACM Computing Surveys*, vol. 31, pp.264-323, 1999.

[3] Daxin Jiang et al, “Cluster Analysis for Gene Expression Data, *IEEE Transactions on* Data and Knowledge Engineering, 16(11), pp.1370-1386, 2004.

[4] U. M. Fayyad, “Data Mining and Knowledge Discovery: Making sense out of data”, *IEEE Expert*, vol. 11, no. 5, pp.20-25, 1996.

[5] Macqueen, J. B. (1967). "Some Methods for Classification and Analysis of Multivariate Observations”, In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics *and Probability*, 2009-04-07, pp. 281–297.

[6] Dibya Jyoti Bora et al, “Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Mat lab”, *International Journal of* Computer Science and Information Technologies, Vol. 5(2), 2014, 2501-2506.

[7] D. Napoleon and P. G. Lakshmi, “An efficient K-Means clustering algorithm for reducing time complexity using uniform distribution data points, *Trendz in Information Sciences &Computing (TISC2010)*, 2010.

[8] Shi Na et al, "Research on k-means Clustering Algorithm: an improved k-means Clustering Algorithm", In: *Third International Symposium on Intelligent Information Technology and Security Informatics*, 2010.

[9] C.Liu et al, "A Modified K-Means Algorithm-Two-Layer K-Means Algorithm", In: Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2014

[10] Dervis Karaboga, Celal Ozturk, "A novel clustering approach: Artificial Bee Colony (ABC) algorithm", Applied Soft Computing 11 pp.652–657 2011, Elsevier.

[11] Seyed Mohammad Seyed Hosseini et al, "Cluster analysis using data mining approach To develop CRM methodology to assess the customer loyalty", Expert Systems with Applications (2010) 5259–5264.

[12] Jaideep Srivastava et al, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations. Jan 2000. Volume 1, Issue 2 - page 12.

[13] David J. Ketchen, Jr; Christopher L. Shook (1996). "The application of cluster analysis in Strategic Management Research: An analysis and critique", Strategic Management Journal. 17(6), pp.441–458.

[14] R. Tibshirani et al, "Estimating the number of clusters in a data set via the gap statistic," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 63, no. 2, pp.411–423, 2001.

[15] R. G. Lawson and P. C. Jurs, "New index for clustering tendency and its application to chemical problems," Journal of Chemical Information and Modeling, vol. 30, no. 1, pp.36–41, Jan. 1990.

[16] RStudio,Wikipedia,20-Oct-2017.https://en.wikipedia.org/wiki/RStudio.

[17] A. Kassambara and F. Mundt, "Package Factoextra, *CRAN repository*, Aug. 2017.

[18] Malika Charrad, Nadia Ghazzali et al, "Package NbClust, *CRAN repository,* 2015.

[19] UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php.

[20] Y. S. Thakare and S. B. Bagal, "Performance Evaluation of K-means Clustering Algorithm with Various Distance Metrics", International Journal of Computer Applications, vol.110, no. 11, pp. 12–16, 2015.

[21] M. Thirupathaiah, P. Venkata Prasad and V. Ganesh, Analysis of Various Compensation Devices for Power Quality Improvement in Wind Energy System. International Journal of Electrical Engineering & Technology, 7(3), 2016, pp. 25–39.

[22] M. T. Shah and P. N. Tekwani, Bi-Directional Three-Level Front-End Converter for Power Quality Improvement. International Journal of Advanced Research in Engineering and Technology, 7(4), 2016, pp 17–29.

[23] B.Rajani and Dr.P.Sangameswara Raju, Comparision of Pi, Fuzzy & Neuro-Fuzzy Controller Based Multi Converter Unified Power Quality Conditioner, Volume 4, Issue 2, March – April (2013), pp. 136-154, International Journal of Electrical Engineering and Technology (IJEET)

[24] V. R. Patel and R. G. Mehta, "Data Clustering: Integrating Different Distance Measures with Modified k-Means Algorithm", In: Advances in Intelligent and Soft Computing Proceedings of the International Conference on Soft Computing for Problem Solving), pp. 691–700, 2012.

[25] Shankar Kumar Yadav and Vikas Srivastava, Non-Conventional Materials in Rigid Pavement: Effect on Mechanical Properties. International Journal of Civil Engineering and Technology, 8(4), 2017, pp. 1888–1896.