Evaluating Students Placement Performance Using Normalized K-Means Clustering Algorithm

M. Vasuki[1], Dr.S. Revathy[2]

[1]Research Scholar, Satyabama Institute of Science and Technology, Chennai.
Department of Computer Science Engineering,
dheshna@gmail.com
[2]Associate professor, Satyabama Institute of Science and Technology, Chennai.
Department of Master Computer Application,
ramesh.revathy@gmail.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: Ensemble Cluster is verifiedas a worthy alternative in front of theanalysis of the clustering problems. Constructinga cluster for ausing similar dataset and mergingit into a distinct clustering. The mixturing process is useful toextend theclustering quality.Another name of clustering Ensemble is consensus clustering. ClusterEnsemble providing as a promising solutions for heterogeneous or for multisource data clustering. Spectral ensemble clustering results in used todropped thedifficulty of algorithm.Now weprovide various clusteringmethods applied in same dataset and producedifferent clusteringresults. The several methods feature alldiscussed, it helped in choosing the utmostsuitable one to solve a problem at handOn the preprocessed dataset, clustering isgenerated by using clustering's namely; normalized k-means, to predict the level of student's performance inplacement.

Keywords: Consensus Clustering, k-means, performance.

I.INTRODUCTION

Cluster analysis is that the necessary technique in any field of analysis used foranalyzing variable knowledge. Herewe applying some of clustering algorithms such ask means, KCC++, GKCC and KCC to same dataset and we can get different results. For finding which result is correct one? How we can estimate the best one?

In clustering analysis there are two types of approach have been used for cluster result evaluations.

1.1 Cluster Validity Indexes (CVI)
1.2. Clustering Ensemble Algorithms.

1.1 Cluster Validity Indexes (CVI)

The cluster validity indexes are used to calculate the quality of clustering results.

1.2. Clustering Ensemble Algorithms.

Combining the different clustering results and yield a single results as an another approach for improving the quality of the clustering algorithms results.

Cluster ensemble method has two major steps.

Step 1: Generation

Step 2: consensus function

Step 1: Generation

Various clustering algorithms are applied for same dataset and partition the data objects into different groups. Every group consists of same objects.

.

Step 2: Consensus function

Generation step gives a group of partition results are combined of all partition results into single result called as consensus function.Cluster ensemble algorithm has the property robustness which means the mixture process must have improved performance than the single clustering algorithms.

.

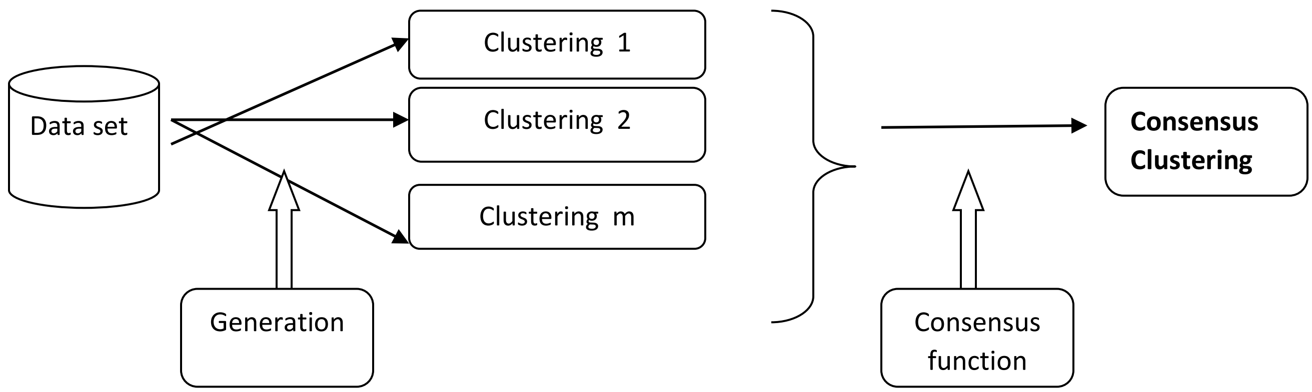


Fig 1.Process of Cluster Ensemble or Consensus Cluster

In Generation Process different clustering algorithms can be used.

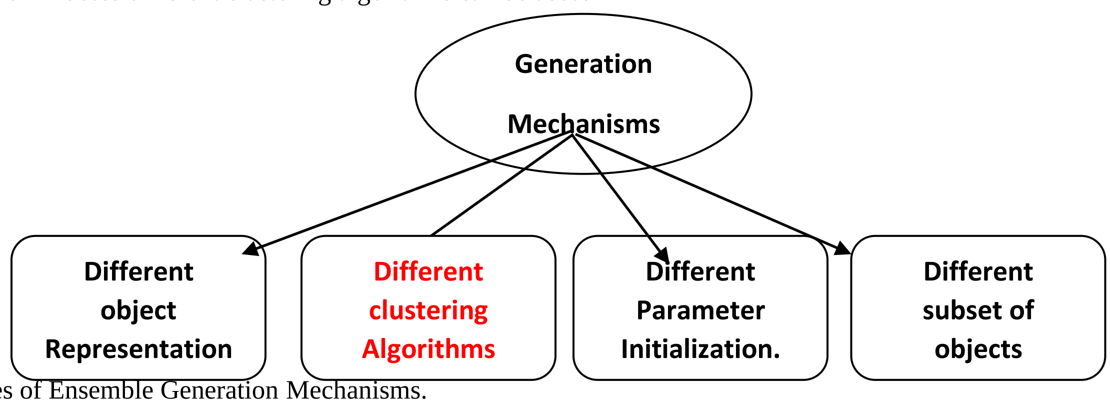


Fig 2.Types of Ensemble Generation Mechanisms.

In the analysis of k-means suffers with initialization sensitivity. Consensus clustering aims to combine several existing basic partition into mergedone. Consensus clusteringProvide robust and high quality performance. Greedy optimization of k-means consensus clustering is used to resolve the sensitivity of the k-means initialization. Greedywith KCC combined to achieve quality clustering. Now GKCC and spectral together with the objective function and itsstandard deviations.

II.Related studies

2.1 Clustering consensus

Cluster consensus function means combining the results of partition's from various cluster algorithm into single clustering. The basic partitions can have different numbers of clusters. Clustering consensus is essentially a fusion issue,

It can be split roughly into two categories:

2.1.1 Utility Function

2.1.2 Co-Association Matrix

2.1.1 Utility Function

The first category designs a utility function that measures the similarity between basic partitions and the final partition, and solves a combinatorial optimization problem by maximizing the utility function.

2.1.2 Co-Association Matrix

The second category employs a co-association matrix to calculate the number of times a pair of instances co-occurring in the same cluster, and then runs a graph partition method for the final consensus result.

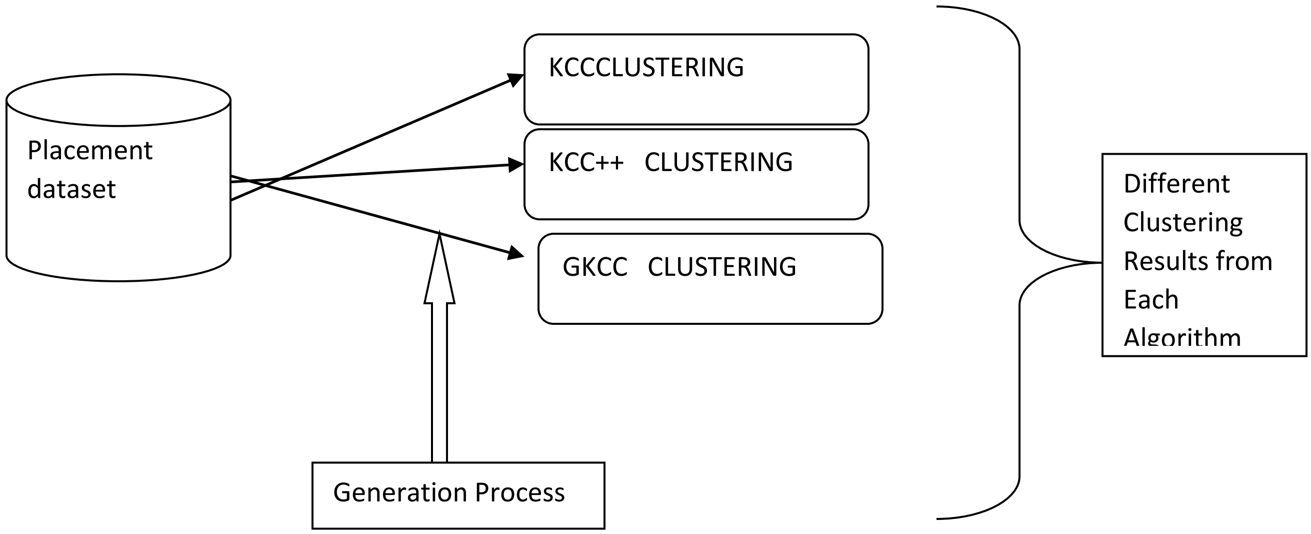


Fig 3.Ensemble Generation Process using with different Algorithms

III.Preliminary knowledge and problem Definition

3.1 K-Means Clustering Consensus
K-means is sensitive to initialization, on both complete and incomplete simple partitioning, Utility features that function for KCC. Experimental findings on different real- world data sets Show that KCC is highly efficient and in terms of clustering efficiency, comparable to state-of-the-art methods,in addition, KCC exhibits high robustness with significant missing values for incomplete simple partitioning. Clustering of consensus (CC) is basically a problem of Combinatorial optimization.

It is possible to loosely divide the current literature into two categories: CC with implied objectives (CCIO) and CC with specific objectives: (CCEO). In CCIO, methods do not set global objective functions. Instead the representative approaches contains graph-based algorithms, co-association matrix-based methods, relabeling and voting methods, geneticalgorithm and some heuristics are specifically implemented to find suitable solutions.

Methods in CCEO have specific global objective functions for the clustering ofconsensus.Among the older ones, Qu adratic Shared Knowledge dependent objective functionand using Kmeans clustering to find the solution is the Medi an Partition issue based on Mirkin distance.

This smart idea might be copied back to Mirkin's work on the Utility Function category, EM algorithm non negative matrix factorization KCC utility functions that establish the general KCC framework is other solutions for various objective functions. KCC is very stable, even with very few highquality basic partitions or extremely incomplete basic partitions.

3.2 Genetic k-means Clustering
The genetic k-means algorithm (GKA), which crosses the genetic algorithm combinedalong with k-means algorithm, is a new clustering tool. This fusion approach provides to accomplish the robustness and highefficiency. The result of, GKA will still meetfaster thanany other genetic algorithms.

3.4 Consensus Clustering of Greedy K-Means (GKCC)
Greedy optimization of K-means-based Consensus Clustering (GKCC) in an expanded partition function space based on greedy center allocation. In a unified system, we strive to overcome the sensitivity of K-means initialization and basic partition generation. A highly efficient version of K-means, inspired by greedy K- means, initializes the K centers with the previous K-1 centers

and greedy searches the remaining one using greedy K-means for initialization of K-means and generation of simple partitions.

Greedy K-means, however, generates n partitions with a certain number of clusters, and for next-step optimization only one partition is chosen. When n is very high, the time complexity becomes costly.Therefore, as the basic partitions for later consensus fusion, the intermediate partitions created by greedy Kmeans are further utilized. A 59- sampling method is used to speed up the speed in order to stop brute- force global search to solve the high time complexity.

For consensus fusion, it is possible to use these intermediate partitions The whole process is comparable to greedy K-means. In each step, the centroids are used to greedily scan for one additional centre in the previous phase, and then K- means are carried out to change the current centroids.

Therefore in order to create subsequent basic partitions, the original data and basic partitions are merged as new data . We therefore provide a new basic partition generation strategy that strongly Couples the subsequent fusion and creates anend ensemble clustering operation. GKCC incrementally adds fresh centers and overcomes the initialization sensitivity of Kmeans..

GKCC's advantages consist of three phases.

1. For a stable and high-quality clustering, it blends greedy K-means and KCC.
2. To create subsequent basic partitions, the original data and basic partitions make the consensus cluster a one-step operation.
3. GKCC overcomes the sensitivity problem of initialization of K- means and delivers a robust output of high quality.

.

3.5Validity measurements

The cluster performance of various cluster strategies calculated in terms of outside measurements, Rn and NMI. Normalized Mutual info (NMI) and normalized Rand index Rn, were wont to value the cluster performance.

IV. Problem definition

GKCC is projected to undo the sensitivity of K-means low-level formatting, high time complexity of greedy K-means. Ensemble clustering fuses various partitions into a single partition.Now we used various clustering algorithm has applied for same dataset for finding ensemble clustering. We going to obtain different clustering results using from three clustering algorithm. Later we combine these results into single cluster by utility function of one of the ensemble algorithms.

V.Experimental Results.

5.1 Datasets

Reg.No	Stud.Name	Dept.Name	Aptitude.Mark	English.Mark	Programming. Mark	Code.Mark
150035	Sasidharan Sampath	IT	8	2	7.7	-2
150051	Harshada Prabhakar	CSE	5.3	3.7	0	3.7
150075	Arokya Rohit Symon	ECE	5.1	3.7	7.3	0.7
150093	Varun Viyas	ECE	-2	2.3	5	6
150113	Sowndarya K	CSE	3.7	2.7	4.7	6.7
150201	Mahes Waran .N	ECE	0.7	1	4.7	10.7
150369	Peddireddy Alekya	ECE	-2	1.3	0	-1.3
150402	Aravind S	EEE	3.7	1	3.7	1.3
150403	A.Asiya Shafreen	ECE	0.7	2	6	6.7

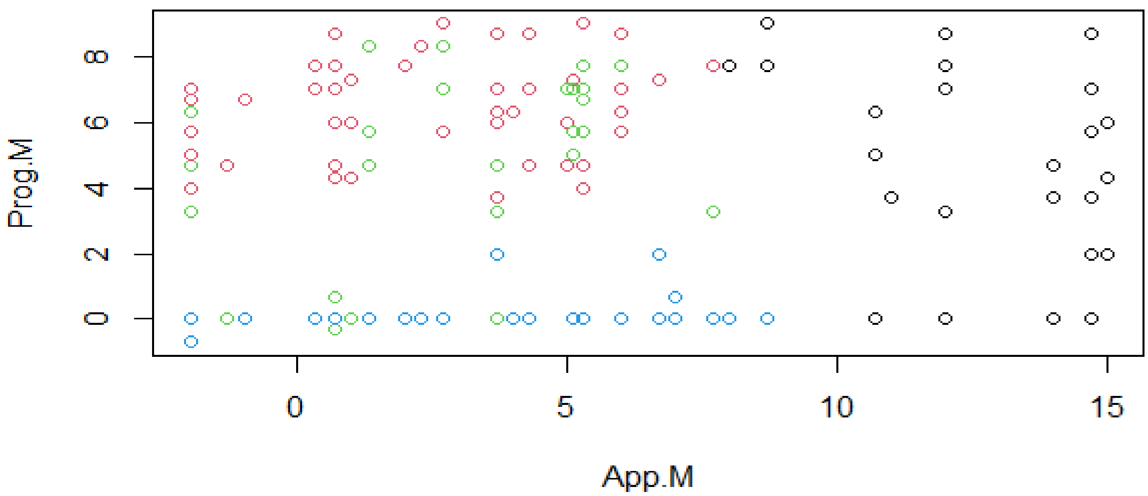
150414	Girija Yesvanthaiyah	ECE	0.7	3.7	0	12
150417	Mondi Nagendra Yadav	IT	8.7	3.7	7.7	6
150432	Jaseema Yasmin	CSE	0.3	2.3	7.7	12
150436	Maria James Anto	ECE	6	2.7	7.7	10.7
150441	Shankar Ut	ECE	5.3	7	0	14.7
150444	Priyasarasu Asokar	EEE	5.3	6	7.7	14.7
150465	Vignesh M	EEE	5.3	6	0	2.7
150474	Dangudubiyam Sri Praveen Sai	EEE	5	6	4.7	15
150476	Kaushik M G	ECE	4	2	0	-2
150481	Arun Coumar	ECE	2.7	6	9	3.7
150491	Govardhanan Muthaiyan	EEE	4.3	1	7	0.7
150513	Ramachandiran K	EEE	1.3	8	5.7	6

5.2 Experimental Result

```
head(TCS_19)
# A tibble: 6 x 4
  App.M Eng.M Prog.M Code.M
<dbl><dbl><dbl><dbl>
1   7.7   5      3.3   12
2   7     5      0.7    6
3   1     3.7    6     12
4  -2     4.7    4.7   10.7
5   3.7   2      6     14.7
6   0.7   2      6     14.7
> nrow(TCS_19)
[1] 199
> n<-TCS_19
> res<-kmeans(n,4)
> res$size
[1] 38 52 48 61
> res$cluster
  [1] 3 4 3 3 3 3 1 1 2 4 2 3 3 3 4 2 3 3 1 3 3 3 3 4 3 4 2 2 3 4 1 2 1 1 1 1 3 1 3
 [40] 2 3 4 4 1 3 3 4 1 4 1 1 1 1 3 1 4 4 2 4 2 1 2 3 4 1 4 4 4 3 3 4 3 1 1 3 2 3 2
 [79] 4 2 2 2 2 2 4 2 3 4 3 4 4 2 4 4 2 4 3 3 1 1 4 2 3 3 3 3 2 4 2 4 4 4 4 3 2 4 2
[118] 4 4 4 4 2 1 4 1 1 1 1 4 1 2 2 2 2 4 2 4 3 4 4 2 3 2 4 1 1 1 4 3 3 4 3 2 2 2 4
[157] 2 4 2 4 4 2 2 2 2 4 2 2 3 2 1 1 4 4 4 4 2 4 2 3 2 4 1 2 2 4 2 3 3 2 3 4 3 4 1
[196] 4 1 1 1
```

Within cluster sum of squares by cluster:
[1] 15.15100 39.82097 23.87947
(between_SS / total_SS = **88.4 %)**

```
> plot(TCS_19[c("App.M","Prog.M")],col=res$cluster)
```




```
4 clusters created.
> table(TCS_19$App.M,res$cluster)

 1  2  3  4
  -2    0  8  7  5
  -1.3  0  1  1  0
  -1    0  1  0  2
  0.3   0  2  1  1
  0.7   0  7  9  5
  1     0  3  2  0
  1.3   0  0  4  1
  2     0  1  0  1
  2.3   0  1  0  1
  2.7   0  2  4  5
3.7   0  7  7  6
  4     0  1  0  2
  4.3   0  3  0  2
  5     0  2  2  0
  5.1   0  1  3  1
  5.3   0  4  5  6
  6     0  5  1  8
  6.7   0  1  0  3
  7     0  0  0  5
  7.7   0  1  1  3
  8     1  1  1  2
  8.7   2  0  0  2
10.7  6  0  0  0
  11    1  0  0  0
  12    8  0  0  0
  14    5  0  0  0
  14.7 12  0  0  0
  15    3  0  0  0
```

```
> z<-TCS_19[,-c(4)]
```

```
> m<-apply(z,2,mean)
> s<-apply(z,2,sd)
> z<-scale(z,m,s)
> View(z)
```

VI.CONCLUSION AND FUTURE ENHANCEMENT

In this paper, K-means algorithm applied in placement dataset and partitioned into clusters which defines evaluation of level of performance in various placement soft skill such as Aptitude, English, Programming Logic and Coding skills of the cluster performance is calculated by R-Tool and normalized by mean/standard deviation formula. Outside measurements called Rn and NMI were calculated for placement dataset.vaious kcc++, genetic k-means, Greedy optimized k-means algorithm also applied for same dataset. Later the calculated cluster performance can be compared and find which algorithm provides quality cluster may be foretold in future work.

REFERENCES:

1. Fang-Xiang Wu,"Genetic weighted k-means algorithm for clustering large-scale gene expression data" BMC Bioinformatics 2008, 9(Suppl 6):S12, DOIhttps://doi.org/10.1186/1471-2105-9-S6-S12
2. Diyar Qader Zeebaree, Habibollah Haron, Adnan Mohsin Abdulazeez and Subhi R. M" Combination of K-means clustering with Genetic Algorithm: A review". Zeebaree International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 24 (2017) pp. 14238-14245 © Research India Publications. http://www.ripublication.com
3. Dataram Soni Madhulatha D.C. Wyld et al" Comparison between K-Means and K-Medoids Clustering Algorithms ". (Eds.): ACITY 2011, CCIS 198, pp. 472–481, 2011. © Springer-Verlag Berlin Heidelberg 2011
4. 4." A Survey on Different Clustering Algorithms in Data Mining Technique" P. IndiraPriya, Dr. D.K.Ghosh2International Journal of Modern Engineering Research (IJMER) Vol.3, Issue.1, Jan-Feb. 2013 pp-267-274 ISSN: 2249-6645.
5. Jyotismita Goswami "A Comparative Study on Clustering and Classification Algorithms" International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-3, June 2015 ISSN: 2395-3470.
6. 6.Pooja Kumari; Praphula Kumar Jain; Rajendra Pamula"An efficient use of ensemble methods to predict students academic performance2018" 4th International Conference on Recent Advances in Information Technology (RAIT)
7. M. Shovon and M. Haque An approach of Improving Student Academic Performance by using K-means clustering Algorithm and Decision tree vol. 3 pp. 8 2012.
8. L. Juhanak et al. "Using process mining to analyze students' quiz-taking behavior patterns in a learning management system" in Computers in Human Behavior 2017.
9. Alejandro Pena-Ayala "Educational data mining: A survey and a data mining-based analysis of recent works" in Expert Systems with Applications pp. 1432-1462 2014.
10. Hashmia Hamsa Simi Indiradevi and Jubilant J. Kizhakkethottam "Student academic performance Prediction Model Using Decision Tree and Fuzzy genetic algorithm" Procedia Technology vol. 25 pp. 326-332 2016.
11. Evandro B. Costa* Baldoino Fonseca Marcelo Almeida Santana Fabrsia Ferreira de Arajo and Joilson Rego "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses" Computers in Human Behavior vol. 73 pp. 247-256 2
12. A. F. ElGamal "An educational data mining model for predicting student performance in the programming course" International Journal of Computer Applications vol. 70 no. 17 2013.
13. Amirah Mohamed Shahiria Wahidah Husaina and Nuraini Abdul Rashida "A Review on Predicting Students Performance using Data Mining Techniques" The Third Information Systems International Conference Procedia Computer Science vol. 72 pp. 414-422 2015.
14. J. J. Wu, H. F. Liu, H. Xiong, J. Cao, and J. Chen, "Kmeans-based consensus clustering: A unified view", IEEE Trans. Knowl. Data Eng., vol. 27, no. 1, pp. 155–169, 2015.
15. J. J. Wu, H. F. Liu, H. Xiong, and J. Cao, "A theoretic framework of K-means-based consensus clustering", in Proc. 23 rd Int. Joint Conf. Artificial Intelligence, Beijing, China, 2013.

16. H. F. Liu, T. L. Liu, J. J. Wu, D. C. Tao, and Y. Fu, “Spectral ensemble clustering”, in Proc. 21 th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 2015
17. H. F. Liu, J. J. Wu, T. L. Liu, D. C. Tao, and Y. Fu, “Spectral ensemble clustering via weighted K-means: Theoretical and practical evidence”, IEEE Trans. Knowl. Data Eng., vol. 29, no. 5, pp. 1129–1143, 2017.
18. C. Domeniconi and M. Al-Razgan, “Weighted cluster ensembles: Methods and analysis”, ACM Trans. Knowl. Discov. Data, vol. 2, no. 4, p. 17, 2009
19. H. S. Yoon, S. Y. Ahn, S. H. Lee, S. B. Cho, and J. Kim, Heterogeneous clustering ensemble method for combining different cluster results, in Proc. 2006 Int. Conf. Data Mining for Biomedical Applications, 2006, pp. 82–92.
20. S. Vega-Pons and J. Ruiz-Shulcloper, “A survey of clustering ensemble algorithms”, Int. J. Patt. Recogn. Artif. Intell., vol. 25, no. 3, pp. 337–372, 2011.
21. Junjie Wu1, Hongfu Liu1, Hui Xiong2, Jie Cao3 “A Theoretic Framework of K-Means-Based Consensus Clustering” Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence
22. Junjie Wu, Hongfu Liu, Hui Xiong, Jie Cao“K-Means-Based Consensus Clustering: A Unified View” january 2015,IEEE Transactions On Knowledge And Data Engineering 27(1):155-169 ,DOI: 10.1109/TKDE.2014.2316512
23. Ms. Archana Singh1, Prof.Dr.V.H.Patil2“Two Layer k-means based Consensus Clustering for Rural Health Information System” International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 07 | July -2017
24. Mr. Anand Khandare1, Dr. A.S. Alvi2 “Efficient Clustering Algorithm with Improved Clusters Quality” IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 6, Ver. V (Nov.-Dec. 2016), PP 15-19
25. Diyar Qader Zeebaree, Habibollah Haron, Adnan Mohsin Abdulazeez and Subhi R. M. Zeebaree “Combination of K-means clustering with Genetic Algorithm: A review” International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 24 (2017) pp. 14238-14245
26. K.krishna” Genetic K-Means Algorithm” February 1999 IEEE Transactions On Cybernetics 29(3):433-9DOI: 10.1109/3477.764879
27. MdZahidulIslamaVladimirEstivill-CastrobMd,AnisurRahmanaTerryBossomaiera.”Combining K-Means and a genetic algorithm through a novel arrangement of genetic operators for high quality clustering”Expert Systems with ApplicationsVolume 91, January 2018,
28. Ashok Kumar D, Usha T. A, Sivaranjani C.” Combination of K-means clustering with Genetic Algorithm: A review” International Journal of Engineering Research and Development e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com Volume 12, Issue 11 (November 2016), PP.01-09.