algorithm, this is considered as insignificant and also can be solved by the execution of the algorithm for many numbers of times. In those situations, where k-means can be used as the integral part of higher level applications, this null cluster problem can produce anomalous behavior of that system and it may lead to the significant performance degradation. Hence we propose modified K-means Algorithm.

3. Modified K-Means Algorithm

The execution steps of the m_k-means algorithm to form clusters are essentially similar to those of the original k-means algorithm. The processor maintains the cluster structures in its own local memory and iterates through the steps of the m_k-means algorithm to evaluate a final set of cluster centers Z. The execution steps to be followed are summarized below.

Input: a set D of d-dimensional data and an integer K.
Output: K clusters
Algorithm:
begin
randomly pick
K points ∈D to be initial means;
while measure M is not stable **do**
begin
compute distance $d_{ii} = ||x_i - z_k||$ for each
k, j where 1≤j≤K and 1≤j≤N, and
determine members of new K subsets based
upon minimum distance to z_k for ≤j≤K;
compute new center z_k for ≤j≤K using k-means;
compute C
end
end

$$J(K,m) = \sum_{k=1}^{K}\sum_{i=1}^{N}(u_{ki})^m d^2(x_i, c_k)$$

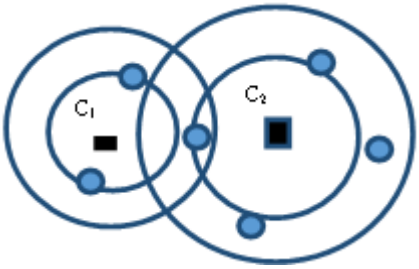

Figure 7: Modified K-means cluster

Clustering in gene expression data sets is a challenging problem. Different algorithms for clustering of genes have been proposed. However due to the large number of genes only a few algorithms can be applied for the clustering of samples. k-means algorithm and its different variations are among those algorithms. But these algorithms in general can converge only to local minima and these local minima are significantly different from global solutions as the number of clusters increases

4. Results

Experimental Work

Experimental work was designed to compare the performance of proposed K-mean algorithm. Number of data elements selected was 1000. And for the sake of experiment, 8 numbers of clusters (k) were entered at run time. The process was repeated 10 times for different data sets generated by MATLAB. The proposed K-mean algorithm is efficient because of less number of iterations and improved cluster quality, as well as reduced elapsed time. In Figure 2, Basic and proposed K-mean clustering algorithms are compared in terms of different data sets. For each run different data sets are generated by MATLAB and entered, to observe the number of iterations. In Figure 3, Basic and proposed K-mean clustering algorithms are compared in terms of same data set. For each run same data set is entered, to observe that at each time numbers of iterations are different in basic K-mean clustering algorithm. The numbers of iterations are fixed in proposed K-mean clustering algorithm because initial centroid's are not selected randomly. Basic K-mean clustering algorithm gives different clusters, as well as clusters size differs in different runs. Table 1 shows different results for same data set as well as elapsed time

Table 1: For different data set

	normal k-means	Modified k-means
1	31	24
2	32	24
3	29	17
4	34	20
5	34	19
6	28	25
7	28	25
8	29	25
9	30	27
10	59	49

We can represent the above table in graphical interface bar chart as in figure 7.

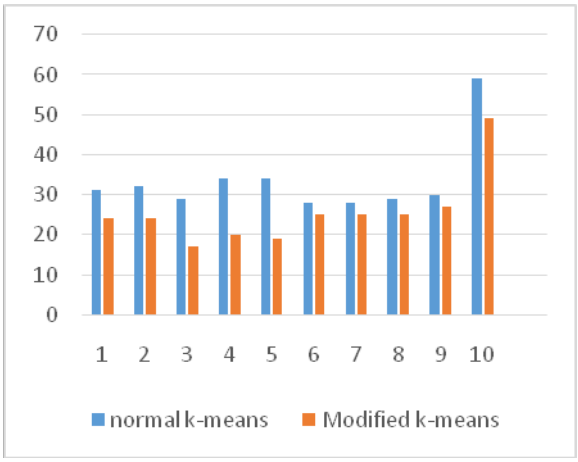

Figure 8: For different data set

Proposed K-mean clustering algorithm gives same clusters, as well as clusters size is same in different runs. Table 2 shows same number of iterations and cluster size.

In this as the size of data becomes high the value of the iterations becomes much higher and the time complexity will be high. Hence by considering it for the genetic data as the total number of genes will be in the order of merely thousands we can go through the modified k-means approach which will produce the more efficient results in less number of iterations. As the same mean will be there for cluster there won't be change in any of the iteration to other.

Table 2: For Same Dataset

Gene	normal k-means	Modified k-means
1	31	24
2	32	24
3	29	17
4	34	20
5	34	19
6	28	25
7	28	25
8	29	25
9	30	27
10	59	49

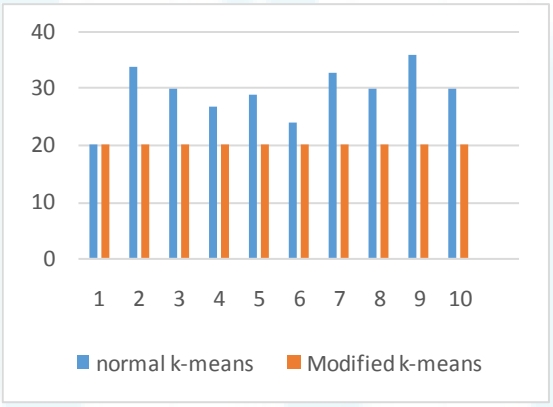

Figure 9: For Same Data Set

As a result, many categorical data clustering algorithms have been introduced in recent years, with applications to interesting domains such as protein interaction data. The initial method was developed in by making use of Gower's similarity coefficient. Following that, the k-modes algorithm in extended the Conventional k-means with a simple matching dissimilarity measure and a frequency-based method to update centroids (i.e., clusters' representative).

For the more values in the case of gene data the normal k-means and modified k-means algorithm will show the following results in the case of yeast gene and mitochondria.

Here we can neatly observe that the number of iterations will be reduced as compared to that of normal k-means algorithm to modified k-means algorithm. Hence even though it have more executing steps due to this the execution time becomes low in the case of the modified approach.

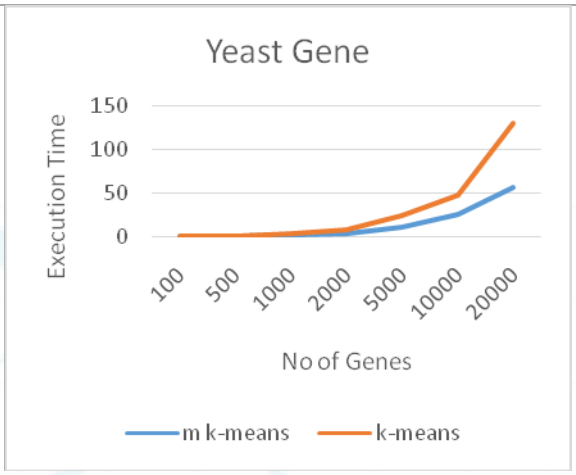

Figure 10: Representation of Graph between Execution Time and Genes in Yeast Gene

5. Conclusion

This paper presents a novel, highly effective link-based cluster ensemble approach to categorical data clustering. The empirical study, with different ensemble types, validity measures, and data sets, suggests that the proposed link-based method usually achieves superior clustering results compared to those of the traditional categorical data algorithms and benchmark cluster ensemble techniques. The prominent future work includes an extensive study regarding the behavior of other link-based similarity measures within this problem context. Also, the new method will be applied to specific domains, including tourism and medical data set.

References

[1] P.J. Rousseeuw and L. Kaufman, Finding Groups in Data: Introduction to Cluster Analysis. Wiley Publishers, 1990.
[2] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On Combining Classifiers," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 3, pp. 226-239, Mar. 1998.
[3] J.C. Gower, "A General Coefficient of Similarity and Some of Its Properties," Biometrics, vol. 27, pp. 857-871, 1971.
[4] D. Cristofor and D. Simovici, "Finding Median Partitions Using Information-Theoretical-Based Genetic Algorithms," J. Universal Computer Science, vol. 8, no. 2, pp. 153-172, 2002.
[5] A. Strehl and J. Ghosh, "Cluster Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions," J. Machine Learning Research, vol. 3, pp. 583-617, 2002.
[6] S. Guha, R. Rastogi, and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes," Information Systems, vol. 25, no. 5, pp. 345-366, 2000.
[7] G. Karypis and V. Kumar, "Multilevel K-Way Partitioning Scheme for Irregular Graphs," J. Parallel Distributed Computing, vol. 48, no. 1, pp. 96-129, 1998.
[8] C. Domeniconi and M. Al-Razgan, "Weighted Cluster Ensembles: Methods and Analysis," ACM Trans. Knowledge Discovery from Data, vol. 2, no. 4, pp. 1-40, 2009.

[9] X.Z. Fern and C.E. Brodley, "Solving Cluster Ensemble Problems by Bipartite Graph Partitioning," Proc. Int'l Conf. Machine Learning (ICML), pp. 36-43, 2004.
[10] D. Gibson, J. Kleinberg, and P. Raghavan, "Clustering Categorical Data: An Approach Based on Dynamical Systems," VLDB J.,vol. 8, nos. 3-4, pp. 222-236, 2000.
[11] A.L.N. Fred and A.K. Jain, "Combining Multiple Clusterings Using Evidence Accumulation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 835-850, June 2005.
[12] M.J. Zaki and M. Peters, "Clicks: Mining Subspace Clusters in Categorical Data via Kpartite Maximal Cliques," Proc. Int'l Conf.Data Eng. (ICDE), pp. 355-356, 2005.
[13] D. Liben-Nowell and J. Kleinberg, "The Link-Prediction Problem for Social Networks," J. Am. Soc. for Information Science and Technology, vol. 58, no. 7, pp. 1019-1031, 2007.
[14] A.K. Jain and R.C. Dubes, Algorithms for Clustering. Prentice-Hall, 1998.
[15] T. Boongoen, Q. Shen, and C. Price, "Disclosing False Identity through Hybrid Link Analysis," artificial Intelligence and Law, vol. 18, no. 1, pp. 77-102, 2010.