Centroid Ratio for a Pairwise Random Swap Clustering Algorithm

Mrs. S.SUGANYA M.E,MISTE.,
ASSISANT PROFESSOR CSE DEPARTMENT
R.V.S SCHOOL OF ENGINEERING AN D TECHNOLOGY
DINDIGUL,TAMIL NADU

MS R.SOUNDHARYA M.E
Dept Of Software Enginnering
R.V.S SCHOOL OF ENGINEERING AND TECHNOLOGY
DINDIGUL ,TAMIL NADU

Ms.J.Laveena Grasy M.E
Dept Of Software Enginnering
R.V.S SCHOOL OF ENGINEERING AND TECHNOLOGY
DINDIGUL ,TAMIL NADU

Abstract: On dealing with data means clustering plays a vital role, in which clustering algorithm and cluster validity are the most common linked approach used for cluster analysis. To improvise the performance, propose the validity index scheme in clustering algorithm and cluster validity. In general the concept of centroid ratio is for comparing the two obtained cluster result. In the proposed system the centroid ratio in Pair wise Random Swap clustering algorithm, in order to avoid the local optimum problem faced by k-means algorithm. The methodology behind Centroid ratio is the swapping must be done between the convergence and perturbation in finding the nearest optimum by means of k-means. The centroid ratio is highly effective in MSE (Mean Square Error) which is fast and simple to process. The proposed system is applied with several database and compare the result with those traditional methods like Random Swap, Deterministic Random Swap, Repeated k-means or k-means++. As a result the proposed system is successfully attained the expected result and in order to improve its efficiency is productively applied in document clustering as well as color image quantization.

Keyword: Data Clustering, K-means, random hdeterministic swap and clustering evaluation

I Introduction

In the globalization world clustering is the important technical keyword which withholds various terms. Commonly clustering is nothing but a grouping of sets which similar to each other or which are linked with some factors. The clustering terms is mainly used in IR systems, information retrieval is difficult to retrieve certain thing from a large dataset to minimize the process the common terms are clusters and then processed. gives maximum result in the terms of cost compensation and time consumption during searching.

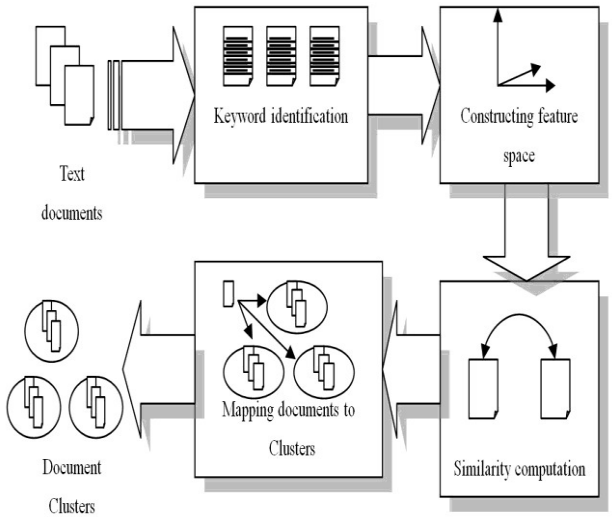

Fig 1: Sample document clustering method

The Sample document clustering method shows that how the document clustering is happening in the real time cases. Based on various clustering methodologies were developed on that prototype based clustering is a significant one. Prototype clustering is developing sequence of method in which the best one fits even with the unknown structure. To understand the methodology centroid is representing in the K- means cluster [1] and it

gives better result in memory capacity as well as computational cost in various causes. The term which makes its ineffective will be sensitive to its initializations. The ineffective problem is addressed by running k-means multiple times with randomly selected parameters [2].

In order to attain a best a result Swap-based clustering algorithm [3] is introduced with the intention of finding optimal path using k-means. In continuation to this swap-based clustering is implemented and obtained good quality results and multiple methods are developed to make this research in a effective manner. The simulated annealing [4] and genetic algorithms [5] is based on stochastic global optimization but cannot be in dealing with the problem of time complexity. A global k-means algorithm (GKM) [6] by means it able to cluster one center at a time using deterministic global search procedure but is not suitable for large data sets. In the next stage the k-means is developed into k++ by improving the accuracy and speed using choosing initial values [7]. But the data characteristic affect the process. In [8] that deals with high-dimensionality, data size, noise, outliers, types of attributes and data sets, with its scales attributes. The dealing of high dimensional space along with traditional indexing is not successful with clustering approaches [9].

The spherical k-means [10] is the algorithm processed by cosine distance which is considered for high dimensional data's. Moreover the clustering validity is categorized into two types such as external and internal validity [11][12] The method of used method is Rand index and the Jaccard coefficient methods. But it is also fails to attain higher accuracy and time consumption problems. In [13], a cluster-level measure discussed about combining two clustering methods based on the global structure of prototypes and centroid. In [14][15][16][17] k-means is employed in different strategies such as remove one cluster or merging the two existing clusters into an agglomerative clustering. The evaluation measures by contingency matrix is studied in the literature [18][19][20]. But it is not that is much of successive in finding the nearest pair in the clustering methods.

II RELATED WORKS

Related work section have discussed with best known existing work such as k-ymeans and swap based clustering's. K-means is the best known learning algorithm in solving the various clustering problems.K-means mainly works with the idea of center k centroid there will be various result will attain from various location. To cluster in best way centroid is the effective method. It uses a simple cost function for evaluating the quality of clustering and optimization problem in which the most common function used is mean squared error (MSE).

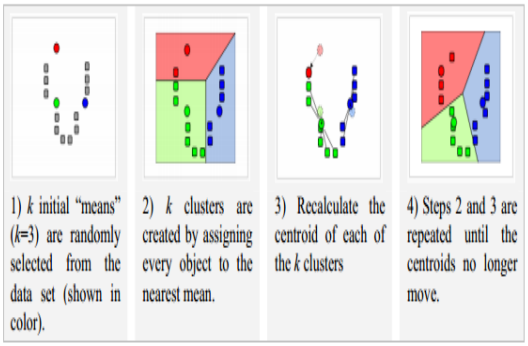

Fig 2: Demonstration of k-means algorithm

It cluster the similar objects based on minimum Euclidean distance to each clusteriods. But the major drawback in using the k-means algorithm is in dealing with different solution the final result obtained is not stable and those strategies were discussed on [2p][7][13].

The next important algorithm is swap based clustering it is a local heuristic search method in finding the optimal centroids. The k-means is not effective in dealing with the local optimum problems as well as allocation of centorids. Earlier to this the trial-and-error method is simple to apply and very efficient while handling. Here the working methodology is a sequence of centroid swaps between existing centroids and a set of candidate centroid by means of fine tunning in the locations. In this order Random Swap algorithm (RS), originally called Randomized Local Search [3]. In simple manner the swapping practice, is done by randomly replacing selecting centriod with selected data objects. The swap is simple to implement but it was not that much of effective to the real time dataset.

```
Input: X, M
Output: C, P, MSE
C ← InitializeCentroids(X) ;
P ← OptimalPartition(X,C) ;
for T times do
    C^new ← RandomSwap(C);
    P^new ← LocalRepartition(P,C^new) ;
    KmeansIteration(P^new,C^new) ;
    if f(P^new,C^new) < f(P,C) then
        (P,C) ← P^new,C^new ;
    end
end
MSE = f (see Eq. 1) ;
return C, P, MSE ;
```

Fig 3: A Pseudo code for Random swap

III OVERVIEW

The proposed system is based on pair wise random swap algorithm with the centroid ratio. The ratio indices are based on three factors such as data set, the point level partitions, and centroids. In this stage the pairing problem is solved bipartite graph in the minimum pairing is between the two centroids of nearby distance.

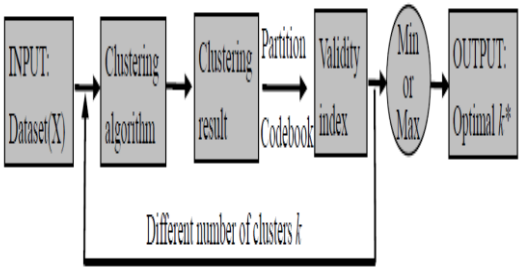

Fig 4: A view of proposed system

The fig4 shows the proposed system and how is it working in which the dataset in taken is input which is passed in to the proposed algorithm which was discussed below. Based on the result the cluster validity is analyzed by means of three important factors such as external index, internal index and relative index. The external index calculates the clustering similarity of ground truth. In which the internal index measure the decency of clustering without the knowledge of external information and in relative index is measured by different parameters value resulting by same algorithm.

The pair wise random swap algorithm: The core this algorithm is as discussed by means pair wise algorithm with centroid ratio by overcoming the local optimum problem faced in k-means. As shown in fig4 it takes the data set along with number of clusters as an input. It generates two set of centroids and the pair wise ratio is calculated for centroids then performing swap algorithm. The algorithm ends only when the two clustering are matched. Our proposed algorithm is like deterministic swap clustering (DR) in which centroid ratio and the allocated position are random before the centroids are to be swapped.

IV SYSTEM DESIGN

In this section the dataset is designed as an initial work, the dataset set maybe of synthetic, real and

documental data. After designed then process starts which under goes preprocessing by means of java codes in an efficient frameworks. There stop word and stemming is done, the stop word is the removal of conjunction words and stemming is the removal prefix-suffix. The resultant data is cluster by means of clustering algorithm. This is then taken as input and centroid ratio is finding between them for nearest pair by calculating clustering index. Here the proposed algorithm is effectively implemented to get exact result. The pair wise algorithm analyzes the local optimum as an end result. Then by fine-tuning the process is handled to improvise its performance.

Expected Output: The prepared dataset is undergoes on pair wise random swap algorithm and the results are collected for next process. The validity of the proposed centroid ratio is calculated by menas of similarity among the rankings on a number of clustering results. The efficiency of the proposed algorithm is depends on two factors such as how many swaps are needed and how much time consumed by each swaps based on the calculation only the result efficiency is predicted. Generally a large number of iteration is needed to produce a quality clustering result but our algorithm takes fewer iterations to reach the same clustering quality. Meanwhile by fast variant at the early iteration because most of the centroids are still active and the processing time is also reduced when compared to the earlier algorithms. To achieve the prominent level of our proposed system this algorithm is applied to color image quantization and the fig 5 represents result of that methodology.

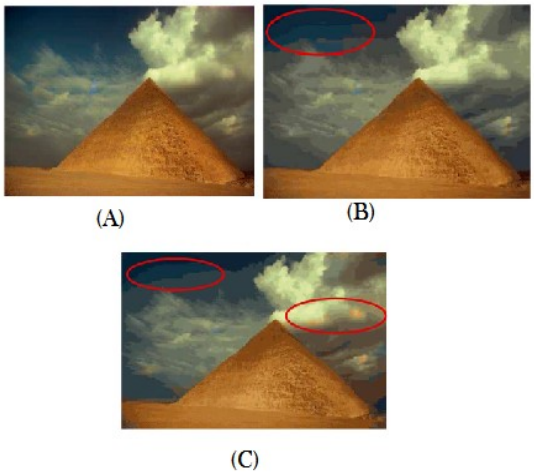

Fig 5: Sample quantization result image

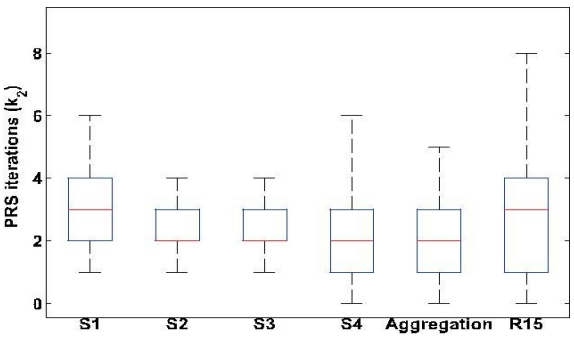

Fig. 6: Boxplot of the required number of PRS iterations.

V CONCLUSION

The new prototype based clustering algorithm that effectively compares two clustering and finds which one is unstable with improper located centroids. This ratio is generated with indices and MSE values to get the proper value. Then implements the pair wise swapping algorithm based on the similarity. The result comparison shows that proposed system is more effective than the results obtained by Random Swap, Deterministic Random Swap, Repeated k-means, and k-means++. The proposed algorithm is applicable for document clustering as well as color image quantization which proves its efficiency in proper manner.

REFERENCE

1) A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognit. Lett., vol. 31, no. 8, pp. 651–666, 2010.
2) A. Jain, M. Murty, and P. Flynn, "Data clustering: A review," ACM CSUR, vol. 31, no. 3, pp. 264–323, 1999.
3) P. Fränti and J. Kivijärvi, "Randomized local search algorithm for the clustering problem," Pattern Anal. Applicat., vol. 3, no. 4, pp. 358–369, 2000.
4) G. Babu and M. Murty, "Simulated annealing for selecting optimal initial seeds in the k-means algorithm," Indian J. Pure Appl. Math., vol. 25, no. 1–2, pp. 85–94, 1994.
5) K. Krishna and M. Murty, "Genetic k-means algorithm," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 29, no. 3, pp. 433–439, Jun. 1999.
6) A. Likas, N. Vlassis, and J. Verbeek, "The global k-means clustering algorithm," Pattern Recognit., vol. 36, no. 2, pp. 451–461, 2003.
7) D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in Proc. 18th Annu. ACM-SIAM SODA, Philadelphia, PA, USA, 2007, pp. 1027–1035.
8) H. Xiong, J. Wu, and J. Chen, "K-means clustering versus validation measures: A data-distribution perspective," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 39, no. 2, pp. 318–331, Apr. 2009.
9) C. Aggarwal, A. Hinneburg, and D. Keim, "On the surprising behavior of distance metrics in high dimensional space," in Proc. 8th ICDT, vol. 1973. London, U.K., 2001, pp. 420–434.
10) I. Dhillon, Y. Guan, and J. Kogan, "Iterative clustering of high dimensional text data augmented by local search," in Proc. IEEE ICDM, Washington, DC, USA, 2002, pp. 131–138.
11) J. Wu, H. Xiong, and J. Chen, "Adapting the right measures for k-means clustering," in Proc. 15th ACM SIGKDD Int. Conf. KDD, Paris, France, 2009, pp. 877–886.
12) Y. Liu, Z. Li, H. Xiong, X. Gao, and J.Wu, "Understanding of internal clustering validation measures," in Proc. 10th ICDM, Sydney, NSW, Australia, 2010, pp. 911–916.
13) S. Khan and A. Ahmad, "Cluster center initialization algorithm for k-means clustering," Pattern Recognit. Lett., vol. 25, no. 11, pp. 1293–1302, 2004.
14) P. Fränti, O. Virmajoki, and V. Hautamäki, "Probabilistic clustering by random swap algorithm," in Proc. 19th ICPR, Tampa, FL, USA, 2008, pp. 1–4.
15) P. Fränti and O. Virmajoki, "Iterative shrinking method for clustering problems," Pattern Recognit., vol. 39, no. 5, pp. 761–775, 2006.
16) H. Frigui and R. Krishnapuram, "Clustering by competitive agglomeration," Pattern Recognit., vol. 30, no. 7, pp. 1109–1119, 1997.
17) P. Fränti and O. Virmajoki, "On the efficiency of swap-based clustering," in Proc. 9th ICANNGA, Kuopio, Finland, 2009, pp. 303–312.
18) E. Dimitriadou, S. Dolnicar, and A. Weingassel, "An examination of indexes for determining the number of clusters in binary data sets," Psychometrika, vol. 67, no. 1, pp. 137–160, 2002.
19) C. Michele and M. Antonio, "A fuzzy extension of some classical concordance measures and an efficient algorithm for their computation," in Proc. 12th Int. Conf. KES, Zagreb, Croatia, 2008, pp. 755–763.
20) M. Meila, "Comparing clusterings—An information based distance," J. Multivar. Anal., vol. 98, no. 5, pp. 873–895, 2007.

