



## MODIFIED CENTROID RATIO BASED CLUSTERING TECHNIQUE USING GENETIC ALGORITHM

Neha Pathak<sup>1</sup>, Imran Hashim<sup>2</sup>, Anand Muni Mishra<sup>3</sup>

<sup>1</sup>Student (M.Tech. CSE), <sup>2</sup>HoD, <sup>3</sup>Research Guide

<sup>1,2,3</sup>Department of Computer Science & Engineering, Jawaharlal Nehru College of Technology (JNCT),  
Ratahara, Rewa, Madhya Pradesh 486001, INDIA

**Abstract:** Clustering is the mechanism for data mining. In this paper we are working on the modified centroid ratio based clustering technique using genetic algorithm for large database. In this paper we will work with MATLAB software to perform the evaluation of clustering techniques for different datasets.

**Keywords:** clustering, techniques, genetic algorithms, data sets, centroid, MATLAB software.

### I. Introduction

Clustering is very powerful tool for the mining of data. The diversity of data increases the applicability and uses of clustering technique. Presently various clustering technique are used on the basis of uses and formation of pattern. The simplest cluster process is partition clustering technique. The partition based clustering technique provides very efficient and simple cluster algorithm is called K-means. In the series of k-means algorithm various algorithms are proposed such as swap based clustering technique, centroid based clustering technique and many more clustering algorithms. On swap-based clustering, the centroid is perturbed by a certain strategy in order not to get stuck in local minima. The swap is acceptable if it improves the quality of clustering. This trial-and-error approach is easy to implement and highly effectiveness in practice. The Random Swap algorithm (RS), known originally Randomized Local Search, is based on randomization: a randomly selected centroid is swapped to another randomly selected location. After a local partition is carried out and the clustering fine tuning by two k-means-iterations. Two clustering's  $\{X, P_1, C_1\}$  and  $\{X, P_2, C_2\}$  from k-means, where the centroids and the partitions are strongly correlated with each other. The partition shows little difference (left) at the boundary between the clusters, while the centroids also focusing little difference on the location. In case of misplaced centroids (right), the partitions are very different. The evaluation of the clustering can be performed either on the partition P or the centroids C.

### II. Genetic Algorithms

Genetic algorithms are search algorithms, which is based on the method of natural selection and natural genetics. They combine the survival of the fittest between the string structures with a structured yet randomized information exchange to form a search algorithm with a sense of innovation of human research. These algorithms are run with a set of random solution called initial population. Each element of this population is called a chromosome. Each chromosome of this problem which consists in the string genes. The number of genes and their values in each chromosome depends on the specification of the population. In the algorithm the number of genes of each chromosome is equal to the number of pixel intensity value and genes of values demonstrates the filter denoising priority is associated with the process, where the higher priority means the noise is executed earlier. Set of chromosomes in each GA iteration is called a generation, which is evaluated by their fitness functions. The new generation means the offspring's generation is created by the application of certain operators on the current generation. These are called crossover which selects two chromosomes of the current population, combines and generates a new child (offspring), and the mutation which changes randomly chromosome genes values and creates new offspring. Then the best progeny are selected by evolutionary selection operator based on their fitness values. The GA four stages as indicated below algorithms:

Step 1: Read data (from matrix) and R values from matrix and get  $N_p$ ,  $N_g$ ,  $X_r$  and  $M_r$  from the structure

$N_p \rightarrow$  (initial population size),

$N_g \rightarrow$  (the number of generations),

$X_r \rightarrow$  (crossover probability),

$M_r \rightarrow$  (mutation probability)

Step 2: Calculate the bottom-level and the top-level of each matrix in the data;

Generate initial population ( $P_i$ );

$P_{current} \leftarrow P_i$ ;

Data from at  $\leftarrow$  Decoding heuristic ( $L$ , Data);

```

Bestcenter ← evaluate (data);
Step 3: while stop criterion not satisfied, do begin
Pdata ← {};
3-1: repeat for (Np/2) times
Father ← select (Pdata, sum_of_fitness);
Mother ← select (Pdata, sum_of_fitness);
Pdata ← Pdata U crossover (father, mother, child1, child2, Xr);
End repeat;
3-2: for each chromosomes ∈ Pdata do begin
Mutate (chromosomes, Mr);
End for
3-3:
Pnew ← Pnew U {four best chromosomes of data}
Pnon-data ← data;
Matrix ← decoding heuristic (data);
Best matrix ← evaluate (matrix);
End while
Step 3: Repeat the best center of data.

```

### III. Proposed Technique

We have working on the modified centroid ratio based clustering technique using genetic algorithm. In this paper we will work with MATLAB software for performance analysis of proposed method along with the algorithm. In centroid ratio based cluster the design of the internal indices is based on three elements: the data set, the point level partitions, and centroids. Mean square error (MSE) is a traditional criterion for evaluation of clustering, which is calculated from these three elements. External indices, however, use only partitions by comparing the given clustering against the ground truth. The ground truth is usually built by using human assessors or the output of another clustering algorithm. External indices count the pairs of points of agreement or disagreement of the two partitions. A criterion such as MSE uses quantities and features inherent in the dataset, which gives a global level of evaluation. As it relates to points and clusters, the time complexity is at least  $O(MN)$ . The partition-based criteria are based on point by point analysis of two partitions, which generally gives a time complexity of  $O(N^2)$ . The time complexity of point-pair measures can be reduced to  $O(N + M^2)$  by a contingency matrix. As a necessary structure of the clustering, the centroid reveals the allocation of the clusters. Two clustering's  $\{X, P1, C1\}$  and  $\{X, P2, C2\}$  from k-means, where the centroids and the partitions are strongly correlated with each other. The partition shows little difference (left) at the boundary between the clusters, while the centroids also focusing little difference on the location. In case misplaced centroids (right), the partitions are very different. The evaluation of the clustering can be performed either on the partition P or the centroids C.

### IV. Proposed Algorithm

In this section discuss the modified algorithm of centroid ratio for clustering technique using genetic algorithm. The Centroid Ratio of cluster used Random swap in terms of data iteration and reduction of iteration in processing of cluster. In the process of modification set the auto level center selection using genetic algorithm. The genetic algorithm processes the data in fashion in random manner. The auto swapping process of clustering technique assigned the variable of center point. Here steps are given below:

1. Auto = (X,C) ← empty //initialize data and randomly center point
2. C\_list ← K-means (Ci\_list,  $K_{\text{auto}}$ )
3. Input C\_list X, the clustering number pn, population scale XN, probability auto P stop conditions cS ;
4. Code the data in real number and initialize population S(i), i = 0 at random;
5. Evaluate the fitness of all individual in the present time D(s);
6. CR clustering requires optimization of cluster center, which way data thrashing of waiting cluster. Therefore the fitness function of algorithm is determined by f(x).
7. 
$$G(s) = \frac{N(s)}{D(s)} = \frac{\sum_{i=0}^{n-1} A_i s^i}{\sum_{i=0}^n a_i s^i}$$
 Arbitrate the termination situation. If the termination situation is satisfied, then turn to step 9, if not, turn to step 10;
8. Crack to find and compute the optimal clustering centers.
9. find final population of GA

```

10. Take the CR optimization on population P (i) and produce the next generation A (i + 1). Then turn to
    step
11. for h ∈ A(i+1) do
12. h.nn ← CR (A(i+1)- {h})
13. h.sc ← Compute-SC (h, h.nn)
14. AUTO ← AUTO ∪ {h}
15. AUTO ← AUTO ∪ {h.nn}
16. if h.sc < thsc then
17. E ← E ∪ {(h, h.nn)}
18. End if
19. end for
20. count ← Matrix
    a. for each pair of components (g1, g2) ∈ G do
21.  $\mu_1 \leftarrow \text{mean-dist}(g1)$ ,  $\mu_2 \leftarrow \text{mean-dist}(g2)$ 
22. if  $\frac{\mu_1 + \mu_2}{2 * \text{centroid\_dist}(g1, g2)} > 1$  then g1 ← Merge (g1, g2)
23. end for
24. Ntype ← empty
25. for x ∈ N list do
26. h ← PseudopointOf(x) // find the corresponding pseudo point
27. MCR
28. end for

```

## V. Results

For the experimental process used MATLAB software. MATLAB is well known software for the analysis of algorithm. For the validation of clustering technique used seven dataset such as E-coil glass iris and bricks dataset such as S1, S2, S3, and S4. Our experimental result shows that the modified centroid based clustering algorithm is better than swap based clustering technique and centroid based clustering technique. If more clusters are present results to better quality. The execution time taken by algorithm of proposed method gives better performance.

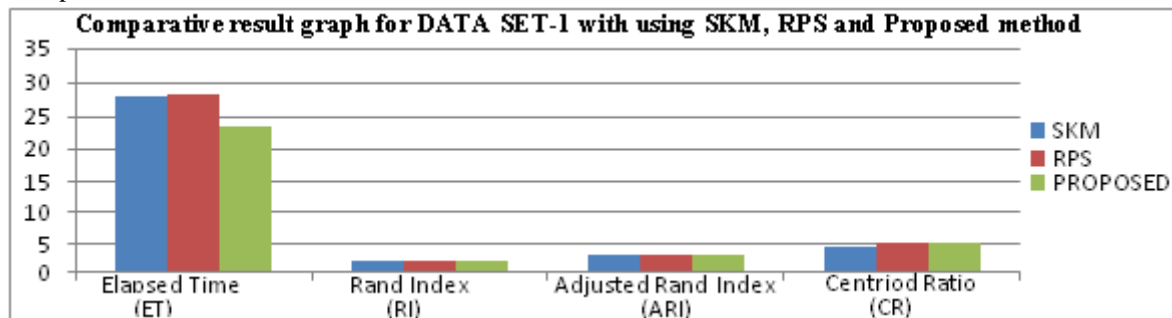


Figure 1.1: Shows that the Comparative result graph for data set-1 data set with using SKM, RPS and Proposed method, to find the value of Elapsed Time, Rand Index, Adjusted Rand Index and Centriod Ratio. Finally we find that our proposed method gives better result than other centroid ratio method.

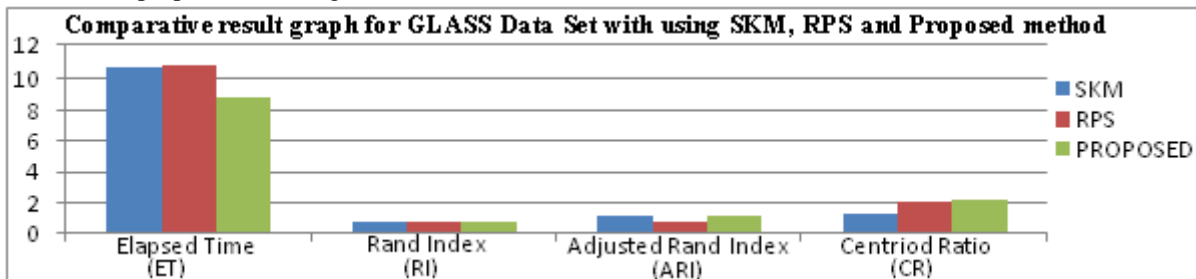


Figure 1.2: Shows that the Comparative result graph for Glass Data Set with using SKM, RPS and Proposed method, to find the value of Elapsed Time, Rand Index, Adjusted Rand Index and Centriod Ratio. Finally we find that our proposed method gives better result than other centroid ratio method.

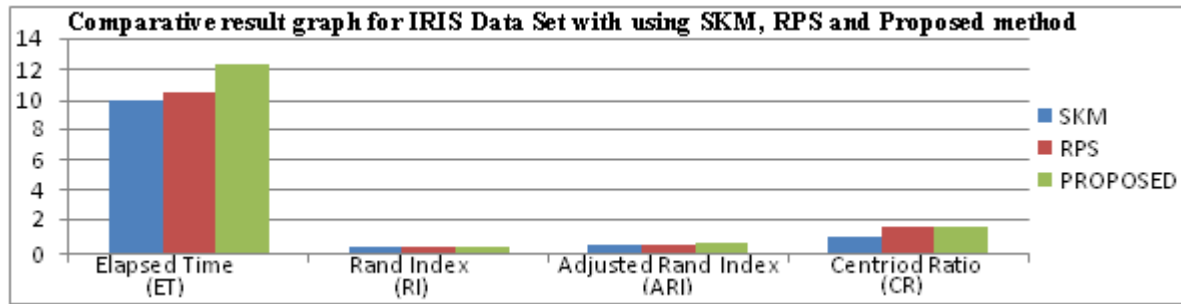


Figure 1.3: Shows that the Comparative result graph for Iris Data Set with using SKM, RPS and Proposed method, to find the value of Elapsed Time, Rand Index, Adjusted Rand Index and Centriod Ratio. Finally we find that our proposed method gives better result than other centroid ratio method.

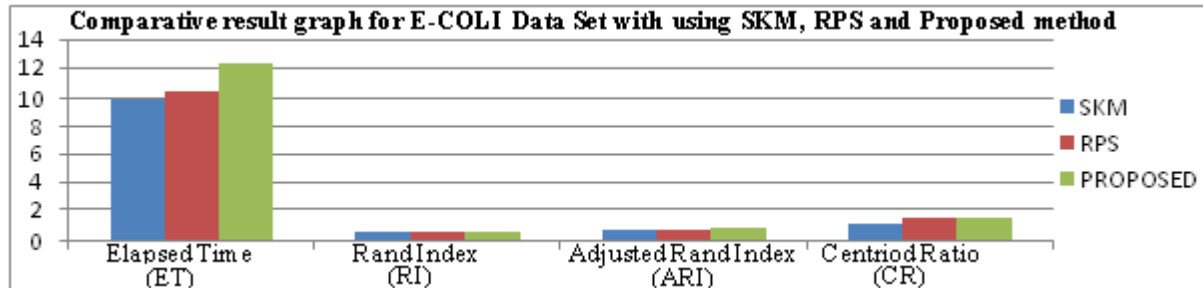


Figure 1.4: Shows that the Comparative result graph for E-Coli Data Set with using SKM, RPS and Proposed method, to find the value of Elapsed Time, Rand Index, Adjusted Rand Index and Centriod Ratio. Finally we find that our proposed method gives better result than other centroid ratio method.

DATA SET	METHOD	Elapsed Time (ET)	Rand Index (RI)	Adjusted Rand Index (ARI)	Centriod Ratio (CR)
DATA SET-1	SKM	28.37	1.30	2.17	4.71
	RPS	28.78	1.33	1.85	5.12
	PROPOSED	23.91	1.35	2.22	5.22
GLASS DATA SET	SKM	10.40	0.50	0.72	1.59
	RPS	10.47	0.53	0.54	2.00
	PROPOSED	9.32	0.55	0.76	2.10
IRIS DATA SET	SKM	9.95	0.40	0.50	0.95
	RPS	10.30	0.43	0.54	1.36
	PROPOSED	12.20	0.45	0.56	1.46
E-COLI DATA SET	SKM	12.88	0.80	1.25	2.81
	RPS	17.22	0.83	1.47	3.22
	PROPOSED	20.36	0.85	1.30	3.32

## VI. Conclusion

We have worked on the modified centroid ratio based clustering technique using Genetic Algorithm. In this paper we will work with MATLAB software to perform evaluation of modified centroid ratio based clustering by using above proposed algorithm we have achieved better results in modified centroid ratio based clustering technique than the swap based clustering technique and existing centroid based clustering technique.

## VII. References

- [1] Damien Hanyurwimfura, Liao Bo, Dennis Njagi, Jean Paul Dukuzumuremyi "A Centroid and Relationship based Clustering for Organizing Research Papers" International Journal of Multimedia and Ubiquitous Engineering, Vol-9, 2014. Pp 219-234
- [2] Guansong Pang , Shengyi Jiang "A generalized cluster centroid based classifier for text categorization" Information Processing and Management, Elsevier ltd. Vol- 49, 2013. Pp 576-586.
- [3] Joeri Hofmans, Etienne Mullet "Towards unveiling individual differences in different stages of information processing: a clustering-based approach" Springer 2011. Pp 1-12K. Elissa, "Title of paper if known," unpublished.

- [4] Kadim tasdemir, Erzsebet Merenyi "A Validity Index for Prototype-Based Clustering of Data Sets with Complex Cluster Structures" IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETIC, 2011. Pp 1-32.
- [5] Nenad Tomasev, Milos Radovanovic', Dunja Mladenic, Mirjana Ivanovic "The Role of Hubness in Clustering High-Dimensional Data" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, REVISED JANUARY 2013, Pp 1-12.
- [6] S. lee, J.i Song, Y. Kim "An Empirical Comparison of Four Text Mining Methods" Journal of Computer Information System, Vol. 51, 2010, Pp 1-10.
- [7] V. Gupta, G. S. Lehal "A Survey of Text Mining Techniques and Applications" Journal of emerging technologies in web intelligence, 2009 Pp 60-76.
- [8] K. Sarkar "Sentence Clustering-based Summarization of Multiple Text Documents" International Journal of Computing Science and Communication Technology, 2009 Pp 325-335.
- [9] L. Cerf, J. Besson, C. Robardet, J.-F. Boulicaut. Data peeler "Constraint-based closed pattern mining in n-ary relations" In SDM, 2008 Pp 1-12.
- [10] Z. Pei-ying, L. Cun-he "Automatic text summarization based on sentences clustering and extraction" In proceeding of International. 2nd Conference of Computer Science and Information Technology, ICCSIT, IEEE, 2009 Pp 167-170.