Journal of Theoretical and Applied Information Technology
15th August 2016. Vol.90. No.1
© 2005 - 2016 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A GENETIC ALGORITHM FOR OPTIMIZED INITIAL CENTERS K-MEANS CLUSTERING IN SMEs

BAIN KHUSUL KHOTIMAH[1], FIRLI IRHAMNI[2], AND TRI SUNDARWATI[3]

[1]Departement of Informatic Engineering, Faculty of Engineering, University of Trunojoyo Madura, Indonesia
[1]Departement of Informatics Management, Faculty of Engineering, University of Trunojoyo Madura, Indonesia
[1]Departement of Informatic Engineering, Faculty of Engineering, University of Trunojoyo Madura, Indonesia
E-mail: bain@trunojoyo.ac.id, firli45@yahoo.com, trisundarwati@yahoo.com

ABSTRACT

This research applies Genetic Algorithm to find the initial cluster centers and the centers of this cluster will be used as an input for the K-Means method. This method yield a more optimal performance compared to the conventional K-Means method since the centers point is optimized with Genetic Algorithm. Unfortunately, k-means is extremely sensitive to the initial choice of centers, and a poor choice of centers may lead to a local optimum that is quite inferior to the global optimum. The proposed research initial clustering and assigns at least one seed point to each cluster. During the second step, the seed-points are adjusted to minimize the cost-function. The algorithm automatically penalizes any possible winning chances for all rival seed-points in subsequent iterations base on cost-function to reaches global minimum. This method will perform optimization centers point K-means clustering using Genetic Algorithms. The Genetic Algorithms optimize centers point of the cluster more faster performance. The simulated experiments described in this paper confirm good performance have used distance measure. So, the analyses performance cluster used SSE (Sum of Squared Error). The minimum Cluster SSE of combination Genetic Algorithm with K-Means Clustering (GA+KMeans) the smallest in compared with the K-Means algorithm of Clustering SMEs.

Keywords: clustering, k-means, genetic algorithm, initial centers, SMEs

1. INTRODUCTION

SMEs that have a high potential in absorbing labor still has wide range of limitations that can not be overcome by optimally. The main problems often faced by SMEs is the difficulty of getting access to capital, limited resources human power, less prepared in the ability of management, limited the ability to read the information access market opportunities. The increased in SMEs, both in the fields of marketing, technology and capital needs to be done immediately. Government facilitation remains indispensable and in high intensity. Based on previous studies of clusters of small and medium-sized enterprises (SMEs) using K-Means is using non-quantitative form of financial numbers factors, but analysis the overall performance of non-financial information involving both qualitative and quantitative which may not be listed in the financial statements (Dehuri S., et al., 2006). Then, the research for Classification SMEs has used a method of clustering based on the initial capital, the average profit, average income, and the average production capacity, Forecasting activities is very used widely in industry fields, manufacturing processes, for instance supply chain, control of each SME in the City (Damaskopoulos T., et al., 2008).

Clustering is one method of data mining without direction (unsupervised) used in the process of grouping data. K-Means is one method of non-hierarchical clustering of data that seek to partition the existing data in the form of one or more clusters (Hruschka, E. R. and Ebecken, N. F. F. 2003). The method K-Means is process of clustering have been weakness caused by the determination of the initial cluster centers. K-Means algorithm searches local optimal solution to initial solution to refine the partition result (Wu K and Yang M., 2001). The Local approximation based heuristic method was used for K-Means clustering and proved it through an empirical study. So, They good initial clustering centroids can be used any of the other techniques. K-Means Clustering improves the clustering centroids to find centers the optimal clustering (Kumar et al., 2010). The study optimize of K-

Means algorithm in determining the initial centers cluster, where results Research shows that the K-Means algorithm has no weakness only has a dependency on the initial data, but also convergence fast and the quality of clustering (Min Feng, Zhenyan-wang 2011).

Experiments show that the algorithm has a cluster quality and good performance K-Means Clustering requires optimization of the central point of K-Means using Genetic Algorithms to obtain an effective and accurate cluster. Genetic Algorithm will determinate of the central point of the cluster that faster performance, which is called the fast genetic (Bashar, et al., 2009).

Combination method of GA and K-Means is also used by Kim et al, 2008 for grouping customers in making recommender online shopping system on the market. Similarity value in each segment is represented by a cluster describe similarities customer behavior patterns. With this cluster analysis, similarity of customer behavior patterns can be explored through the clusters formed (Chiang M M, and B Mirkin. 2010). The more accurate the cluster is formed it will be increasingly obvious similarity patterns of customer behavior. So, the development genetic algorithm recovery iteration cluster that has faster performance and produce better clusters (Lu Y. et al., 2004). The Cluster analysis divided the data into a number of groups that give meaning and useful, based on the information were contained within it that defines objects and their relationships (Tan, Steinbach, and Kumar, 2006). A GA has also been applied the K-Means Clustering to improve the features of the results of the K-Means clustering algorithm known as the new proposed were Initializing Modified Genetic Algorithm K-Means (MGAIK) inspired by an initialization method (Chittu.V, and Sumathi N., 2011).

This research develops optimization of K-Means with Genetic Algorithm capable of producing grouping with the level of variation in the cluster are better in comparison K-Means algorithm is simple. The combination Genetic Algorithm (GA) and K-Means Clustering (GA+KMeans) works on the coding of parameter set rather on the parameters in case study SMEs themselves. With a total generated within the cluster is worth less compared with using the K-means simple. So this method is used for determination of the central point on the optimization of K+Means clustering. With the data obtained from in 2010-2012 with the label parameters in accordance with the productivity of SMEs. The benefits that can be obtained from this study is the incorporation K-Means clustering methods and algorithms Genetics, could be used to forming clusters that have similar characteristics more compact and tight, so that later can be used by business people to create a marketing strategy that is focused on clusters formed, by looking at the characteristics that exist in each cluster.

These systems are known as hybrid power systems. To have automatic reactive load voltage control SVC device have been considered. The multi-layer feed-forward ANN toolbox of MATLAB 6.5 with the error back-propagation training method is employed.

Developed between 1975 and 1977 by J. A. Hartigan and M. A. Wong, K-Means clustering is one of modeling methods (Mitra and Acharya, 2004). K-Means clustering were a set of *n* observations in d-dimensional space (integer) have is given determine a set of *c* points to minimize mean squared distance betwen nearest centers. The problem can be set up as an integer programming problem but because solving integer programs with a large number of variables is time consuming, clusters are often computed using a fast, heuristic method that generally produces well. K-Means algorithm is one such method where clustering requires less efforts. The beginning number of cluster *c* is determined and the centers of these clusters. Any random objects as the initial centroids can be taken or the first *k* objects in sequence can also serve as the initial centroids observations x_n where each observation is a *d*-dimensional real vector, then K-Means algorithm clustering aims to partition the *n* observations into *c* sets ($c < n$) minimize a measure of dispersion within the clusters (Kanungo et al., 2002). K-Means clustering algorithm is an iterative algorithm with minimize the sum of squared errors between vector objects with the centers of the cluster closest for The standard K-Means algorithm minimizes the within-cluster sum of squares distance according to the equation (1) given below:

$$f_i = arg_z min\left(\sum_{i=1}^{c} \sum_{x^i \in z^j} \left\| x^i - \mu^j \right\|^2\right) \qquad (1)$$

μ_j is the centers of the cluster (mean vector) in the cluster to j. The process starts by randomly selecting k pieces of data as the cluster centers stage awal. Pada initial, K-Means algorithm randomly select k pieces of data as the centroid. Then, the distance between the data and the centroid is computed using Euclidian distance. Data is placed in the nearest cluster, calculated from the midpoint cluster. Centroid will be determined when all the data has been placed in The closest cluster. The process of determining the centroid and the placement of the data in the cluster is repeated until the value of convergent centroid. There are two issues in creating a K-Means clustering algorithm: optimal number of cluster and repair centers. In many cases, number of cluster is given then the important part is where to put cluster centers so that scattered points can be grouped properly. Centers of cluster can be obtained by first assigning any random point and then optimizing the mean distance as given in equation (1). The process is repeated until all the centers positions are optimized shows that the K-Means algorithm has the disadvantage of not only has a dependency on the initial data until convergence.

Governments need data from SMEs based groups which developed many clusters of characteristics in common on SMEs. This SME development programs by providing training in quality and quantity. Handicraft of Batik Bangkalan has five industrial segments, namely Batik industry of large, medium, small, micro and centers, each of which has disadvantages. SME development program is a recommendation for priority delivery of training in quality and quantity. This time conditions in production are used in homes and limited marketing. Then, the Department of Industry & Commerce is difficultly monitor the quality of products produced, so the need for a centralized container to facilitate such activities. This study, we propose an algorithm to determine the central election cluster best centers cluster and assign the cluster number (K) to produce optimum quality cluster. This study applied to SMEs to generate mappings based on common characteristics of SMEs based on the parameters - parameters on SMEs. Features data consists of numeric data and categorical data were contained features of SMEs, namely: total assets, number of employees, skills, use of IT, the average number of customers, the average number of products sold, the amount of production, the number of production defects etc. SMEs that applied in this study is Madura Batik SMEs especially in Bangkalan, which have characteristics and have the potential to develop its marketing value ranging local market share to overseas. Each group of SMEs, can be taken into consideration and direction of development of the SME sector more intensive industries. The purpose of this study is to provide SMEs optimal grouping results by using a genetic algorithm and k-Means clustering. The method gives results grouping subtle or not a lot of shifting the centers of the cluster, so that if any new data is entered SMEs will be able to login in a particular group with matching criteria and did not change the results of the previous grouping.

2. GENETIC ALGORITHM

Genetic algorithm is a search algorithm prior of mechanism of natural selection the same as precise algorithms used in solving optimization problems complex. Genetic algorithms are widely used in business applications, technical, sceduling flow job and other. The algorithm starts for solutions a population taken andis used to form a new population. The new population will be better formed were chosen to form the new solusisolusi selected according to their fitness (Dehuri S., Ghosh A., and Mall R., 2006).

The issues related to global and local minimum based on clustering is important and critical. The objective function of the K-Means algorithm is not convex and hence it may contain many local minima. The process of random searching and information sharing make these algorithms best tool for finding global solutions (Min F. and Wang Z., 2011). The algorithms one of such algorithm i.e. Genetic Algorithm (GA) for data clustering have advantages of finding global optimal solution. In this section we aim to propose a hybrid sequential clustering algorithm based on combining the K-Means algorithms and GA algorithm. The motivation for this idea is the fact that GA algorithm, at the beginning stage of algorithm starts the clustering process due to its fast convergence speed and then the result of GA algorithm is tuned by the K-Means near optimal solutions. Flow chart of proposed algorithm is shown in Figure 1.

The purpose of which is expected after the author doing research is combined K-Means clustering with Genetic Algorithm (GA) to optimize the cluster centers, particularly in the data set with the data type mixture of numerical and categorical, which is expected to generate grouping (clusters) better data. A recommender system using GA+KMeans clustering in an online shopping market explained that the performance of clustering using K-Means clustering method has a high degree of dependence on the determination of the point of initial cluster centers randomly generated, so that often cause results clustering solution that is trapped in a local minimum. Genetic algorithms are used to improve the determination of the central point early in the process based K-Means clustering. Stages algorithm GA+KMeans perform preprocessing of data is the implementation of a Genetic algorithm to obtain the optimal initial centroid, among others:

1) Representation of chromosomes
   In this study using integer encoding. Genes represented in the form of a string of bits. The binary length represents the length of chromosomes. The length of chromosomes is much more the number of attributes and number of clusters.
2) Initialize population
   In the process of generating the initial population, the centers of the cluster K is encoded on each chromosome and randomly generated based on data that have been done before the data preprocessing.
3) Calculation Evaluation fitness function
   At this stage of the evaluation is to calculate the value of the objective function, which is part of the judge or measure each chromosome of the criteria resolution.
4) Calculate the distance to the centroid of each original data is represented in the chromosome. The formula used in calculated objective function value that is using euclidean distance with the formula:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2} \quad (2)$$

The combination of K-Means algorithm and GA will generate the better result compared to the result of individual algorithm. This algorithm will remove the drawbacks of combination both K-Means Algorithm and GA Algorithm for producing the best optimized result. GA algorithm is a probabilistic approach to find the optimal solution and hence in every run it generates a new optimal solution to find global optimal point. It is normally suggested to take 50 runs of the algorithm and find the mean value of it for further for running processing. Although GA is a good clustering method, it does not perform well when the dataset is large or complex. K-Means is added in sequence to the GA used at the initial stage to help discovering the vicinity of the optimal solution by a global search. The result from GA is used as the initial seed of the K-Means algorithm, which is applied for refining and generating the final result.

3. METHODOLOGY

This study was designed to SMEs parameter data grouping consisting of Total Labor, Investment Value, Value Production, Raw material type, value Raw Materials, Marketing. So, the GA+KMeans clustering methods performed to yield stable clusters using cluster centers point optimization with Genetic Algorithms.

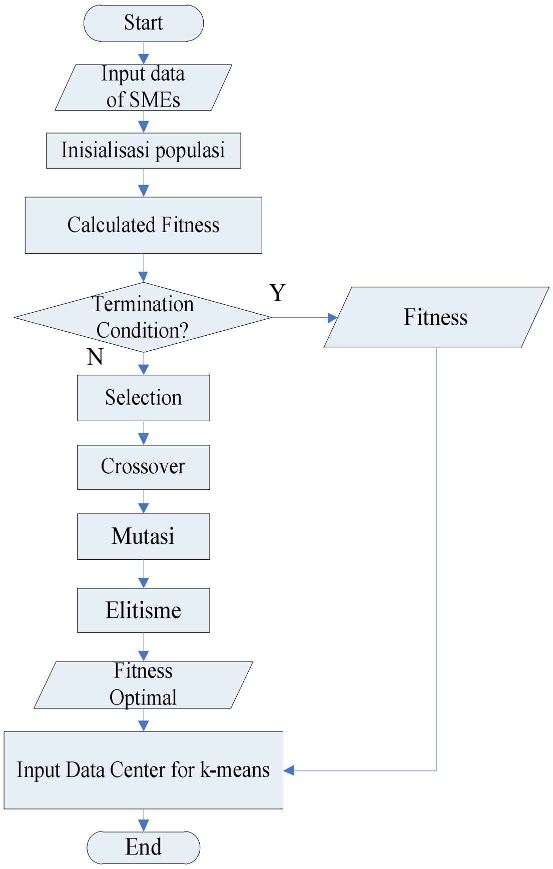


Figure-1. Flow Chart Of GA For Centers Data

Then the results of the K-means clustering that has been in the GA optimization can form clusters that will be used to analyze the characteristics of the perpetrators of SMEs. Data will be processed to be in the cluster using the GA+KMeans and compared with the usual method of K-means. To measure the performance of each method used sum of square error (SSE). With this analysis is expected to help trade services selection policy in conducting training. SMEs need to be given special attention and assistance as appropriate. This study is expected to obtain a more optimal cluster centers that will assist in the formation of clusters which is more accurate and to achieve a better level of computing with the type multivariate data.

3.1 COST FUNCTION CRITERION

Clustering results measurement methods are often referred to as Clustering criterion. Clustering criterion is what will be the fitness value of each chromosomes that have been evaluated. The higher the fitness value, the more evaluated the good chromosome. Clustering criterion used is a cost function, or cost spent to put the object in the corresponding cluster. The general solution is used to determine whether Nice or not a partition is to use Clustering Criterion. Cost function that is widely used is the search for matrix dispersion in the cluster (within the cluster dispersion matrix). One of the ways to define the cost function is the following formula.

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} y_{il} d(X_{I}, Q_{l}) \quad (3)$$

Here, $Q_{1=} [q_{11}, q_{12}, \ldots\ldots, q_{1m}]$ is a vector representative or often referred to as a prototype for the cluster, and Y_{il} is an element of the matrix Ynxl partition. d is similarity measure, which generally uses Euclidean distance.

3.2 DATA IN EUCLIDEAN SPACE

In analyzing the validity of the Cluster by using SSE (Sum of Squared Error) is the total kuadarat errors that occur when n data $x_1,\ldots, x_i$ are grouped into clusters with each cluster centers is *C*. The value of SSE depends on the number of clusters and how data is grouped into the Cluster-Cluster. The smaller the value of SSE, the better its results Clustering using SSE formula is as follows:

$$SSE = \sum_{l=1}^{K} \sum_{x \in C_i} dist\,(c_i, x)^2 \quad (4)$$

Another method for evaluating using both inter and intra cluster scatter is the validity index method (Kim et al., 2008). Centroid will be established if all the data has been placed in the nearby clusters. The process of determining the centroid and the placement of the data in the cluster centroid is repeated until the value of convergent (centroid of all the clusters are not changed again). The Measurement Effective clustering is maximizes intra-cluster similarities and minimizes inter-cluster similarities. Then, we define inter-cluster and intra-cluster similarities and the similarity between a data pair establish. The different ways recently effort in using indexes relating within and between cluster distances. While the value of average distance between *i* and all other entities of the cluster. Euclidean Distance *y(i)* is the minimum of the average distances between *i* and all the entities in each other cluster. The resulting clusters will determine mean intra-cluster distance (MICD) defined for the k th cluster as:

$$Vc^2 = \frac{1}{n_c - 1} \sum_{i=1}^{n} (y_i - \bar{y}_c)^2 \quad (5)$$

So, performance for inter-cluster minimum distance (ICMD) defined as:

$$V_w = \frac{1}{N - c} \sum_{i=1}^{c} (n_i - 1).Vi^2 \quad (6)$$

In order to create cluster validity index, the behavior of these two measures around the real number of clusters (K∗) should be used variant amount Cluster automatically. Whereas in analyzing whether or not a cluster formation process may be obtained from inter and intra huungan formed Cluster (Cluster density). Statement to measure the relationship between the density of both inter and intra cluster can be determined by evaluating the validity Cluster.

4. RESULTS AND DISCUSSIONS

The test data that is widely used the clustering problems for the assessment of the credit worthiness called dataset SMEs with 20 target variable of which 13 variables including type of category and the remaining 7 variable of type arithmetic. From the results of the GA, the initial population of 15 chromosomes, trials with parameter 0.50 crossover, 0.25 of mutasi and 50 of generations. Firstly the working of the proposed scheme and refinement in the cluster centers is Euclidean Distance illustrated. Secondly to evaluate the performance of the proposed clustering algorithm, few experiments have been conducted on two artificial generated data set problems and another two with standard data mining benchmark problems. The test results of minimize Euclidean Distance finally for Optimization with GA+KMeans, and K-Means are shown in Table 1. Comparison some experiments conducted to see the difference in the total cost for each method.

Table 1 shows obtain the total cost (GA+KMeans) is smaller more than conventional K-Means. This result was influenced by the determination of the total population and maximum generation. In this case, this research used total population and younger generations respectively.

The production capacity of each SME has different units, for it is necessary to standardize the data variable that will be used with the process of normalization. The original data to calculate the distance to the centroid for Similarity measure used in K-means method is the minimum distance data with the centroid. In this study to find fitness, the formula is used to enter variable Euclidean Distance of K-Means clustering. The result of Table 1 and table 2 are calculated Genetic algorithm methode to the maximum generation, the total population and the number of clusters is used to calculate the distance eucledian distance used to obtain a stable cluster and the end result.

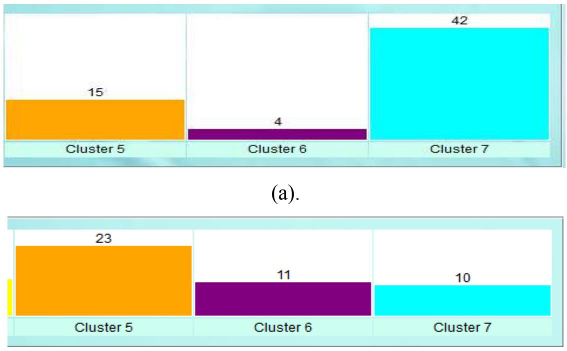
Table-1. Comparison Of K-Means And GA+Kmeans Algorithm With Total Cost Fitness

Total cost K-Means+GA			K-Means
Total Iteration	Max Generation	Total Cost	Total Cost
80	5	2621.53	2121.52
80	13	2345.34	2045.31
80	23	2412.35	2012.32
80	28	2105.32	2005.12

Table-2. Eucledian Distance

ED (C2)	ED (C3)	ED (C4)	Arg min
0.508	0.600	0.401	0.401
0.906	0.953	0.849	0.849
0.282	0.421	0.234	0.234
0.677	0.669	0.637	0.637
0.387	0.005	0.201	0.005
1.869	2.123	2.074	1.869
0.343	0.248	0.162	0.162
0.264	0.403	0.446	0.264
0.542	0.474	0.444	0.444

Fig.2. The Show is a graph comparing the number of clusters on SMEs that the results are not comparable. Tests conducted with 100 iterations between K-Means usual method with GA+KMeans. Results chart performance with the GA+KMeans more balanced than in the K-means ordinary. In these algorithms, every run generates a new solution so the values reported are averaged of 100 iteration.. Clustering method genetic algorithm and K-Means clustering can be used to categorize similar small industries in one or more regions, evidenced by mistake SSE ≤ 0.1. The division of a dynamic group of SMEs conducted to examine the performance of the method, the results of the testing methods used to place the results of similar groups based on the accuracy of the most minimal. Groupped used as a reference to a policy in the development of SMEs to see the resemblance between the clusters. Results will be treated in the same cluster with sesame same cluster, so it will be easier for the government to provide assistance to increase SME.



(a).

(b).

Figure-2. Results SSE Of Chart Performance 100 Iteration: A. Group Cluster K-Means; B. Group Cluster GA+Kmeans

Table 3. the same as Fig. 2 present the comparison of algorithms considering *intra- cluster* and *inter-cluster* distances. Clusterisasi comparative test GA+Kmeans and K-means better than K-Means. This is because the centroid value that is used to process the K-Means is not random but in doing optimization with Genetic Algorithms. The results obtained mean clustering algorithms (K-Means, GA+KMeans) by using the sum of squares error (SSE), Variant cluster Within and Between.

Table-3. The Comparison Performance V Of Mean Intra-Cluster Distance (MICD) And Inter-Cluster Minimum Distance (ICMD)

Cluster	Method	SSE	MICD	ICMD
2	K-Means	0.4925	0	0
3		0.5595	0.0896	0.1826
4		0.5664	0.0343	0.1243
5		0.5713	0.0462	0.1673
2	GA+ KMeans	0	0	0
3		0.0670	0.0667	0.2667

4		0.0851	0.0276	0.1976
5		0.1074	0.0376	0.2014

The third test is done with 50 iterations shows the cluster results of both methods. Table 4. Shows the value of SSE for 50 iterations. Values are the smallest SSE In all clusters found on the GA+KMeans method. It can be concluded that the method has a good performance with test parameters and p.mutasi GA with p.crossover higher then it can affect the outcome of the process of grouping

Tabel-4. The Result Intra-Cluster Distance Within 50 Iterasi

Value of Within Cluster Variation K-Means	**Value of Within Cluster Variation GA+KMeans**
0.384 ±0.032	0.184 ±0.032
0.444 ±0.032	0.284 ±0.032
0.684 ±0.032	0.372 ±0.032
0.784 ±0.032	0.364 ±0.032

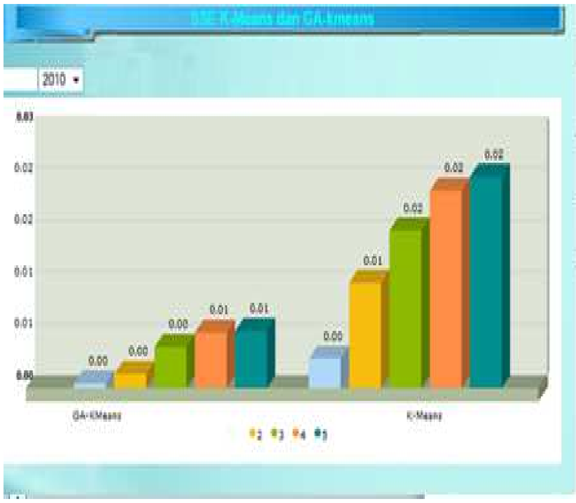


Figure-3. Shift Of Graph GA+Kmeans

Table -5. Comparison SSE Of K-Means And GA+Kmeans

SSE K-Means	SSE GA + KMeans
0.014692	0.00207
0.013801	0.00304
0.011104	0.00735
0.014172	0.00911

Figure 3. Graph above is a graph comparing the performance of SSE using five iterations between K-means usual method with GA-KMeans. Charts for all the smaller cluster GA+KMeans of the K-Means ordinary. So, it can be concluded GA-KMeans method better performance than the K-Means conventional, since in cluster 6 and 7 total distance SSE both methods are very much.

It has been seen that for first two problems GA for generate better solution than K-Means, GA+KMeans but for the other two the other algorithms are better while the proposed algorithm generates better solution among all of them. It is also seen that the deviation in results by proposed Hybrid GA+KMeans clustering algorithm is much less than its counter parts and hence proves its stability. It is because initial clustering made by GA is further tuned with K-Means algorithm which has capability of obtaining better local optimal solution. Hence, the proposed solution always generates better solution than its counter algorithms.

5. CONCLUSION

This paper investigated the application of the GA in sequence with K-Means to clustering problem. After designing and manufacturing applications GA+Kmeans Method for Clustering of Small and Medium Enterprises (SMEs) can the analysis of the maximum generation that is proportional to the fitness value which is derived, because the greater the number the greater the generation of observation room will be great anyway mutation and crossover are also directly proportional to the solutions generated. GA+Kmeans algorithm capable of generating grouping levels in the cluster variasi better in K-means ordinary.

Proposed of Genetic Algorithm has been better convergence to lower quantization SSE (Sum of Squared Error), larger inter-cluster distances and smaller intra-cluster distances performance more better than K-Means Conventional. The variation in GA+KMeans the solutions obtained for different cases is also reported minimum SSE in the proposed research. It can be concluded that the drawback of finding optimal solution to find centers by K-Means can be minimized by using GA over it, so the variations in GA algorithm and its hybridization with K-Means algorithm is proposed for future research.

REFRENCES:

[1] Bashar Al-Shboul, and Sung-Hyon Myaeng, "Initializing K-Means using Genetic Algorithms", *Engineering and Technology International Journal of*

Computer, Electrical, Automation, Control and Information Engineering, World Academy of Science, Vol.3, No.6, 2009, pp. 1481-1485.

[2] V.N. Chittu and N. Sumathi, "A Modified Genetic Algorithm Initializing K-Means Clustering", *Global Journal of Computer Science and Technology*, Vol.11, No.1, 2011, pp. 54-62.

[3] A Kumar, Y Sabharwal and S. Sen, "Linear time approximation scheme for clustering problems in any dimensions", *Journal of association for computing machinery* (ACM), Vol. 57, No. 5, 2010, pp. 1-32.

[4] T. Kanungo, D.M Mount., N Netanyahu., C Piatko., R Silverman. and A.Y Wu. "An efficient K-Means clustering algorithm, Analysis and implementation", *IEEE Trans. Patterns Analysis and Machine Intelligence*, Vol. 24, No.7, 2002, pp. 881-892.

[5] K.J. Kim and HA Ahn, "Recommender system using GA K-means clustering in an online shopping market", *Expert Systems with Applications*, Vol.34, No.2, 2008, pp.1200-1209.

[6] K. Wu. and M Yang, "Alternative c-means clustering algorithms", *Pattern Recognition*, Vol. 35, 2002, pp. 2267 – 2278.

[7] F. Min and Z. Wang, "A Genetic K-means Clustering Algorithm Based on the Optimized Initial Centers", *Computer and Information Science*, Vol. 4, No.3, 2011, pp. 88-94.

[8] S. Mitra and T. Acharya, "Data Mining", *Wiley Publications*, 2004.

[9] T Damaskopoulos., R Gatautis. and E Vitkauskaite, "Extended and Dynamic Clustering of SMEs", *Economics of Engineering Decisions*, Vol.56, 2008, pp. 11-21

[10] Chiang M. M. and Mirkin B., "Intelligent Choice of the Number of Clusters in K-Means Clustering", *Journal of classification,* An Experimental Study with Different Cluster Spreads, Vol. 27, 2010, pp. 3-40.

[11] P. N. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining", United States: Pearson Education Inc., 2010.

[12] S. Dehuri, A. Ghosh, and R. Mall, "Genetic algorithms for multi-criterion classification and clustering in data mining", *International Journal of Computing & Information Sciences*, Vol.4. No. 3, 2006, pp. 143-154.

[13] Y Lu, S Lu, F Fotouhi, Y Deng, S Brown, "Incremental genetic K-means algorithm and its application in gene expression data analysis", BMC *Bioinformatics*, Vol. 5, 2004, pp. 1-10.

[14] E. R. Hruschka and N. F. F.. Ebecken, "A genetic algorithm for cluster analysis", *Intelligent Data Analysis*. Vol.7, No. 1, 2003, pp. 15-25.