A Study of Out of School Children Problem in Rajasthan using K-means clustering with Genetic Algorithm

Astha Pareek
Research Scholar
Dept. of CS& IT
THE IIS University,
Rajasthan, India

Amita Sharma, PhD
Asst. Professor
Dept. of CS& IT
THE IIS University,
Rajasthan, India

Manish Gupta, PhD
Project Director (OSD)
(Former), Dept. of Science
&Tech., Govt. of Rajasthan,
Rajasthan, India

ABSTRACT
Clustering technique has been broadly used in numerous disciplines, such as science, statistic, software engineering and other social sciences in order to identify natural groups in large amounts of data. K-means is one of the most generally used partitioning clustering algorithms that tries to locate a user specific number of clusters (k), which are represented by their centroids, by minimizing the square error function. There are two straightforward approaches to cluster center initialization i.e. either to choose the initial values arbitrarily or else to choose the first k samples of the data points. Both approaches cause the algorithm to converge to sub optimal solutions. In contrast Genetic algorithm is one of the most frequently used transformative calculations which perform worldwide research to discover the result to a clustering issue. The algorithm normally begins with an arrangement of haphazardly developed individuals called the populace and design consecutive, new eras of the populace by genetic operations for example population selection, fitness function, crossover and mutation. This paper compares K-means and genetic algorithm based data clustering. A new algorithm is proposed known as genetic algorithm K-means (GAKM).Comparison was done of the basis of external, internal and time complexity.

Keywords
Clustering, K-means, Genetic Algorithm, Dropout, never enrollment.

1. INTRODUCTION
Data mining lately has stimulated wide spread interest in the information engineering, mostly because of the presence of a huge amount of accessible information and the vital need to change over these information into relevant data and knowledge (Li et. al., 2010).Knowledge discovery and data mining in databases are consider as an equivalent words but data mining is essentially a stage in the development of knowledge discovery. The fundamental usefulness of data mining methods is to implement different techniques and algorithms in order to find out extract patterns of gathered data. These appealing patterns are accessible to the user and can be stored as a new knowledge in knowledge base (Sachin R and Vijay, 2012). Data mining involves the use of refined data analysis tools to find out previously unfamiliar, suitable pattern and affiliation in large data set (Jain and Dubes, 1988). There are different areas like business, science, geography, education etc where data mining finds its application. Education is one of the essential areas of data mining research. The term education data mining covers all kinds of problems analysis or resolved using data mining techniques. Problems like analysis about student performance, organization performances etc are quite popular ones.

Clustering is one of the functions of data mining. It (Selim and Ismail, 1984) is valuable method for the analysis of information sharing and patterns in basic data. The objective of clustering is to find solid and inadequate areas in data sets. There are two main methods of clustering:partitional clustering and hierarchical clustering. K-mean algorithm is major categories of partitioning algorithm it finally minimizes the objective function value (Zhou et. al., 2009).It is an iterative procedure. As indicated by the initial cluster centre, it characterizes the objects of data set and recalculates the clustering centre and the information object order. The end of the cycle denotes that the clustering foundation capacity has merged. K-means clustering calculation is quick and simple to execute. It is suitable for different sorts of information, for example, picture, content etc. However the clustering results rely on the initial cluster centre. If the choice is not suitable, clustering results will be more unstable and the quantity of iterations will add, so it will enhance the time and space complexity.

On the other hand it also exists several flaws, for example K value must be given in prior so it is sensitive to human subjective element, not objective; Since the algorithm arbitrarily selects initial cluster centers, so it is simple to fall into most favorable solution; Also, algorithm requires often checking the data sets; resultant in the competence of productivity of the algorithm is not extremely effective and so forth. How to select the most appropriate cluster centers, as well as to keep the accuracy, and to maintain the stability of the cluster centers, which have important significance in practical applications. Initial centroid or seed selection is very critical problem in K-means as it restrict algorithm to produce solution with uniqueness. Thus, it needs to be resolved. Another vital issue is the value of K. The value of K decides the number of clusters to be constructed from the data sets. There is no generalized value of K, as it varies from data to data. Different improved version of K-means has been proposed but there is no standardized approach to overcome these limitations. Diverse algorithms are proposed using optimization technique in literature. Genetic Algorithm is best optimization technique which is simple to understand and in comparison takes lesser time to converge. Hence a genetic algorithm is augmented before K-means clustering to provide a better initials seed. The fitness function of genetic algorithm has been designed according to the problem. Genetic algorithm gives good initial seed selection.

Literature states that problem associated with higher education have been analyzed using mining approach. But problems related to school education is not well explored. Education has always been important but in today's history research suggest that elementary education is very important in the development of children. Universal enrollment and

retention are important aspects in elementary education. Globally, every country suffers with these categories of problems. In India, there are proper school and good government policy laded down by government to enrich the elementary education system. But the outcomes are not at all satisfactory. Rajasthan is the largest statue of India where a survey was conducted by sarva shikha to study the scenario of elementary education. The widespread elementary education can be accomplished in two ways i.e. universal maintenance and universal enrollment. Universal enrollment of elementary education is associated to out-of-school children and universal maintenance of elementary education is a linked to in school children. So the children of age group between 6-14 yrs can be categorized into 2 categories i.e. those who are presently enrolled in schools and those who are presently out-of-schools. Thus, the aim of Sarva Shiksha Abhiyan is to move toward out of school children and enrolled them in schools as well as to retain the enrolled children in the schools. Out of school children is quite evident problem around the globe. Indian schools suffer with the same problem. Various reports and surveys were conducted to identify the reasons behind the concern problem. But due to diversity in the nature and environment, there are no common reasons across India. Every state has their diverse reasons of dropout and out of school. CTS survey was conducted at elementary education in Rajasthan and about 12 lakh of data was collected of children falling in the age group of 6-14 yrs. The CTS survey report is the primary source to frame dataset for the study of the problem. The data comprises of child name, district code, gender, disability factor, class number status of children, and reason for dropout etc.

This is an interdisciplinary research where computerized algorithms were used to study the education problem. The main contributions of the research are:

- An appropriate system was designed to study the out of school children problem using clustering technique.
- An improvement was introduced for handling the initial seed selection problem in K-means clustering using genetic algorithm with respect to education data set.
- A suitable K value was observed for the given data set.
- This research has drawn a path for further research in the concern area.

A system has been proposed for school educational data study. A second contribution is simple but efficient approach to improve the initial seed selection process in K-means clustering considering cluster quality. The new proposed algorithm introduced Genetic algorithm to optimize initial seeds in K-means clustering (GA with K-means). The system also designed a feature to analyze and validate the clustering. Clustering validation process uses two tpes: internal and external metrics. These metrics draws a clear picture of cluster formation. In the designed system internal metrics: silhouette coefficient and average silhouette width and external metrics: purity and entropy have been coded. The final contribution of this paper is to list suggestions for improvement of school education in Rajasthan.The entire paper is composed in this way. The paper discusses research method in section 2, data sources and description in section 3 and basic K-means algorithm and new algorithm in section 4.5. In section 6, the experimentation, results and analysis are illustrated. Section 7 frames the conclusions.

In this research, the algorithms (basic K- means and GA based K-means) were applied on CTS (child tracking survey) data for analysis.

2. RESEARCH METHDOLOGY

Sources of Data

Secondary data are utilized for 33 areas of Rajasthan for the year 2010.A variety of financial variables which are mostly responsible for dropouts and never enrollment. Information were drawn from family unit survey report directed in the year 2010 by Rajasthan Elementary Education Council, Jaipur to develop a child tracking system(CTS) for every child having the age between 6-14 years. The information for dropout rates and enrollment and numerous other financial variables crosswise over Rajasthan are considered for the year 2010 in light of the accompanying reasons: i. Since household census survey in Rajasthan was conducted in 2010, it was not conducted in subsequent years. ii. District-wise enrollment rate and dropout rates as published in CTS household census survey report are used because district-wise dropout rates are not available in DISE data and the children population of age of 6-14 year is required which is published in household census survey report.

3. K- MEANS CLUSTERING ALGORITHM

It (Ray and Turi, 1999) is a partition-based cluster analysis method. As indicated by the essential K-mean clustering algorithm; clusters are entirely reliant on the collection of initial clusters centroid .K data element are chosen as initial centers , then distance of all data fundamentals are figured by Euclidean distance method. The procedure is preceded until no more changes happen in clusters. The following figure shows steps of the basic K-mean clustering algorithm (Wang, J. and X. Su, 2011).

Steps:

- Randomly select k data objects from data set D as initial centers.
- Repeat;
- Calculate the distance between each data object di (1 <= i<=n) and all k clusters C j(1 <= j<=k) and assign data object di to the nearest cluster.
- For each cluster j (1 <= j<=k), recalculate the cluster center.
- Until no change in the center of clusters.

Time complexity of K-mean Clustering is represented by O(nkt). Where n is the number of objects, k is the number of clusters and t is the number of iterations (Dong, J. and M. Qi, 2009).

4. LIMITATIONS OF K-MEAN ALGORITHM

The benefit of K-means clustering algorithm is that it is well-organized and it can be appropriate for high dimensional data. But it suffers with several limitations which are as follows:

- Sensitive to the selection of initial cluster center (local minima problem). It also restrict algorithm to produce solution with uniqueness.
- In K-means clustering algorithm the value of K is very important. There are no appropriate facts for the outcome

of the value of K (number of cluster to create) and susceptible to initial value, for distant initial value, there may be distant clusters developed.

- This type of clustering algorithm has a robust compassion to the noise data objects. If there is a certain amount of noise data in dataset, it will influence the concluding clustering results, leading to its fault.
- K-means clustering algorithm for the analysis of clusters of random shape is most complex.
- This algorithm has the limitation on amount of data. In the iterative procedure, each time there is need to modify the cluster to which data objects belongs.

The contemporary research aims to overcome the problem of initial centroids selection using genetic algorithm. In the next section proposed algorithm has been discussed.

5. GENETIC ALGORITHMS

Genetic Algorithm is a programming technique who forms its basis from the biological evolution (Kleinberg et. al., 2002). It has been increasingly adopted as a critical thinking methodology. It is simple to implement and is successful in providing good solution in practical problems. It works tremendously wonderful in any problem where domain knowledge is limited. Genetic algorithm utilizes the standards of determination and development to create a new solution for a given problem. In clustering algorithm GA has been applied in two ways: as a combination or before the clustering algorithm. In both the approaches K-mean constructed significantly high quality clusters. The literature review revealed that, the major focus of GA based algorithm was to generate high quality clusters in optimized time. There are number of algorithms to handle initial selection problem but no standard algorithm is available. In the current scenario, the out of school children problem is studied using basic K-means clustering. According to the problem, genetic algorithm was designed with K-means to deliver satisfactory answers of the problem. A new fitness function has been introduced in the proposed algorithm to produce centroids for initial startup.

6. PROPOSED METHODOLOGY

Jenn-Long Liu, Yu-Tzu Hsu and Chih-Lung Hung (2012) proposed Genetic algorithm k-means a combination method which link a genetic algorithm and K-means clustering. The role of GAKM is to decide the optimal weights of the attributes and cluster centers of clusters that are desirable to categorize the data set.

This paper presents genetic algorithm K-means clustering (GAKM) to solve the initial centroid problem. This algorithm was able to produce high quality cluster in less time for CTS data set. This factor was tested using cluster quality measure like silhouette index, purity and entropy. Thus the approach in developing new algorithm was problem specific and a criterion of selecting of initial centroid was influence from the nature of domain. Fitness function of GA was designed to cover maximum data in problem space. Distance between selected initial centroids was calculated and centroid combination with highest value was declared as the fittest combination. This combination was provided to K-means for cluster formation. Steps of the algorithm are given below:

1. Select random k clusters, n chromosomes with n population.
2. Repeat until condition satisfied.
   - 2.1 Evaluating the fitness functions of initial chromosomes
   - 2.2 Based on fitness value rank selection was done
   - 2.3 Create the new generation of chromosomes using crossover and mutation and go to step 2.1.

   [Condition: Same set of population was not obtained or maximum limit of iterations achieved]
3. Supply centroids generated by Genetic Algorithm process to clustering algorithm.
4. Generate final cluster.

The basic steps of genetic algorithm include initialization of population, selection, crossover, fitness calculation and mutation. The detail related to Genetic Algorithm has been described as follows:

Table 1: Genetic Algorithm K-means

S.No.	Parameter	Size/Value
1	Population Size	1210917
2	Maximum Generation	14
3	Probability of Crossover	25%
4	Probability of mutation	25%
5	Selection Strategy	Ranking
6	Termination Criteria	Same Generation

7. RESULTS AND DISCUSSION

The experimental results comparing the K-means and genetic algorithm, GAKM algorithm are provided on CTS data sets. In this research, data was collected from CTS survey conducted by RCEE Government of Rajasthan and about 12 lakhs of data has been reported in the CTS survey.

In CTS study of Rajasthan student of the age group 6-14 years was divided into two categories School Dropout and Never Enrolled. This survey gives the complete database of students within school and out of school. The analysis of the algorithm and the outcome of algorithm of different cases were defined to study the dropout problem. The experiment was done on the case given below:

CASE: The most significant reasons for class dropout in a particular district.

This case was selected to investigate drop out problem in Rajasthan. In our research data sets are having two parameters: reason for out of school and drop out. Reason for out of school parameter ranges from 1-14 while drop out parameter ranges from 0-8.These values were chosen of the basis of two criteria: (a) the variation in centroid selection with respect to the value of K (b) average silhouette width of clustering.

Table 2: Two significant reasons for class dropout in particular District

K-means Clustering							Genetic Algorithm K-means					
District	**Clusters**	**Cluster Centroid**	**Pop**	**ASV**	**Purity**	**Entropy**	**Clusters**	**Cluster Centroid**	**Pop**	**ASV**	**Purity**	**Entropy**
1	5	13,3	9359	0.68	0.79	0.10	5	5,2	11340	0.99	0.93	0.17
	9	5,0	7342	0.74	0.73	0.42	9	14,0	4745	0.90	0.70	0.38
	12	14,0	5142	0.75	0.73	0.24	12	14,0	4534	0.81	0.70	0.22
2	5	12,0	10254	0.79	0.88	0.60	5	12,0	9602	0.95	0.95	0.36
	9	12,6	6654	0.78	0.83	0.61	9	12,0	9661	0.93	0.96	0.29
	12	12,0	9697	0.78	0.75	0.40	12	12,0	9661	0.80	0.82	0.41
3	5	5,0	23106	0.76	0.76	0.38	5	6,0	15687	0.96	0.83	0.45
	9	6,0	14850	0.71	0.89	0.42	9	14,1	18900	0.88	0.74	0.35
	12	13,0	14217	0.73	0.70	0.12	12	5,0	16186	0.80	0.78	0.49
4	5	13,2	14115	0.68	0.72	0.12	5	12,0	14193	0.98	0.89	0.25
	9	5,1	12362	0.74	0.88	0.09	9	14,6	8508	0.91	0.72	0.30
	12	13,0	7626	0.77	0.80	0.18	12	10,0	7760	0.96	0.77	0.08
5	5	13,0	5205	0.74	0.70	0.17	5	13,0	5139	0.96	0.76	0.49
	9	13,0	4744	0.79	0.95	0.13	9	14,0	4333	0.95	0.87	0.00
	12	14,0	4706	0.72	0.84	0.10	12	14,0	4333	0.79	0.83	0.33

Table 2 shows comparative analysis of K-means clustering and genetic algorithm K-means.Firstly in K-means clustering there are 2 cluster centroids for 5 districts which are having the highest population for every district and for every k value chosen for analysis i.e. 5, 9, 12. There are four fields in the table shown: population size, average silhouette value of cluster, purity of the cluster and entropy of the cluster. District 1 when k was 5, class dropouts was 3 due to reason 13. The average silhouette value of cluster was 0.68, purity was 0.79 and entropy was 0.10. District 1 with value of k was 9 class dropout was 0 due to reason 5. The average silhouette value of cluster was 0.74, purity was 0.73 and entropy was 0.42. District 1 shows class dropout at 0 due to reason14 when k was 12. The average silhouette value of cluster was 0.75, purity was 0737and entropy was 0.24.The first significant reason of school dropout was 14 for district 1, as this reason had high average silhouette value and purity with low entropy. For purity measure, if the cluster purity value is closer to 1 then cluster has high purity and closer to 0 it has low purity. The entropy is inverse of purity close to 0, it is refers good cluster and closer to 1 refers to poor cluster.

In the contrast to genetic algorithm K-means (GAKM), district 1 when k was 5, class dropouts was 2 due to reason 5. The average silhouette value of cluster was 0.9909, purity was 0.9323 and entropy was 0.1737. District 1 with value of k was 9 class dropout was 0 due to reason 14. The average silhouette value of cluster was 0.9095, purity was 0.7080 and entropy was 0.3829. District 1 shows class dropout at 0 due to reason14 when k was 12. The average silhouette value of cluster was 0.8121, purity was 0.7022 and entropy was 0.2422. These results were also compared with the factual data set prepared using excel. The reasons produced by GAKM are more accurate and closer to actual data.

8. COMPARISON OF BASIC K CLUSTERING WITH GAKM CLUSTERING

The comparison of basic K-means clustering and genetic algorithm K-means (GAKM) was done on three parameters i.e. internal index, external index and complexity. External indexes require a prior data for the purposes of evaluating the results of a clustering algorithm, whereas internal indexes do not. There are three comparison measures were used:

- Internal Index: Average Silhouette Value of clusters and Average silhouette width of clustering algorithm.
- External Index: Purity and Entropy of clustering
- Complexity : Time Complexity

Internal indices are used to measure the goodness of a clustering structure without external information. In the figure three line charts has been drawn for k value 5,9,12 for all districts using table 3.The chart revealed that clusters formed by basic K clustering and GAKM were initially closed but with the variation in k value GAKM algorithm generated better clusters.

Table 3: Average Silhouette Value (K=5,9,12)

Districts	K= 5 Average Silhouette Width		K=9 Average Silhouette Width		K=12 Average Silhouette Width	
	Basic K-means clustering algorithm	GAKM algorithm	Basic K-means clustering algorithm	GAKM algorithm	Basic K-means clustering algorithm	GAKM algorithm
1	0.67	0.72	0.69	0.70	0.55	0.63
2	0.66	0.78	0.57	0.68	0.51	0.64
3	0.67	0.75	0.62	0.67	0.53	0.68
4	0.68	0.75	0.69	0.70	0.57	0.64
5	0.68	0.76	0.64	0.68	0.56	0.60

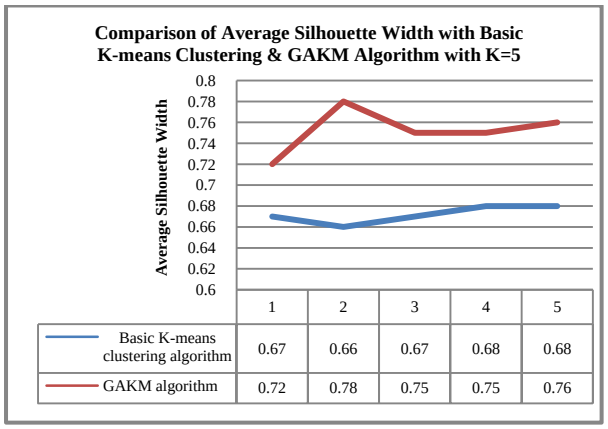

Fig. 1: Comparison of Average Silhouette Width with Basic K-means Clustering & GAKM Algorithm with K=5

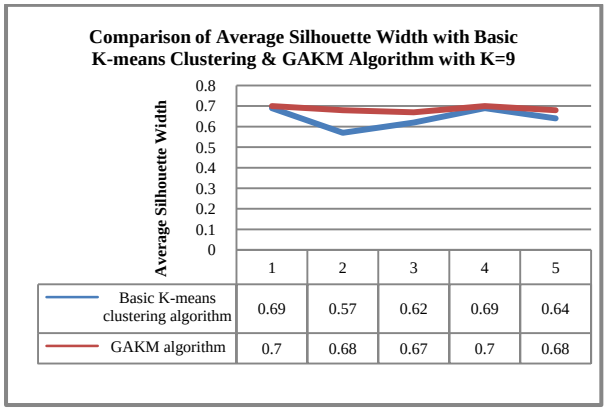

Fig. 2: Comparison of Average Silhouette Width with Basic K-means Clustering & GAKM Algorithm with K=9

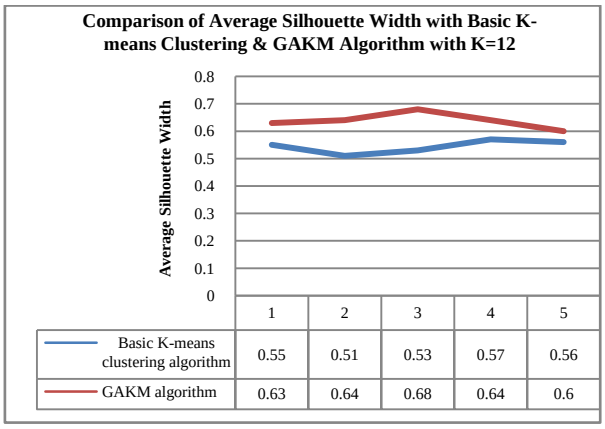

Fig. 3: Comparison of Average Silhouette Width with Basic K-means Clustering & GAKM Algorithm with K=9

Purity is a one of most important validation measure to resolve the cluster quality. The entropy and purity are broadly used measures. Entropy uses external information class labels in this case. The purity of the clusters is calculated on the position to the class labels or ground fact is called as entropy. The lesser entropy means better clustering. The entropy increases when ground truth of objects in the cluster further expands. The larger entropy means that the clustering is not of high-quality. The amount of disorder is created by using entropy.

Table 4: Comparison of entropy analysis between K-means and GA (K=5, 9, 12)

District No	K= 5 (Entropy)		K=9 (Entropy)		K=12 (Entropy)	
	Basic K-means clustering algorithm	GAKM algorithm	Basic K-means algorithm	GAKM Algorithm	Basic K-means algorithm	GAKM Algorithm
1	0.55	0.35	0.79	0.57	0.72	0.46
2	0.65	0.45	0.80	0.53	0.65	0.52
3	0.49	0.33	0.70	0.67	0.53	0.43
4	0.53	0.47	0.41	0.41	0.49	0.43
5	0.58	0.49	0.35	0.32	0.48	0.44

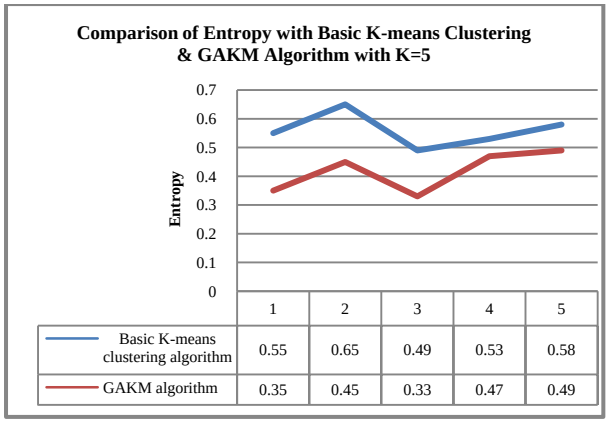

Fig. 4: Comparison of Entropy with Basic K-means Clustering & GAKM Algorithm with K=5

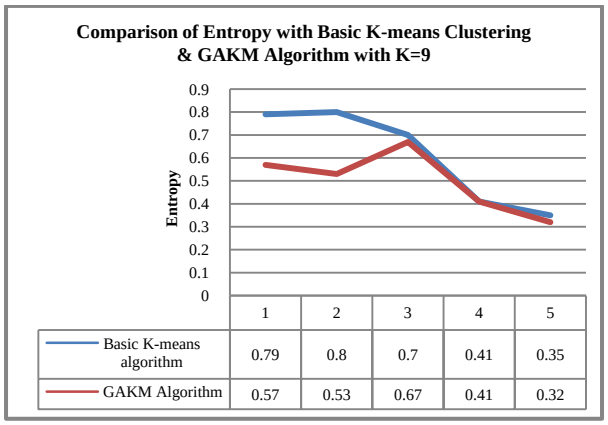

Fig. 5: Comparison of Entropy with Basic K-means Clustering & GAKM Algorithm with K=9

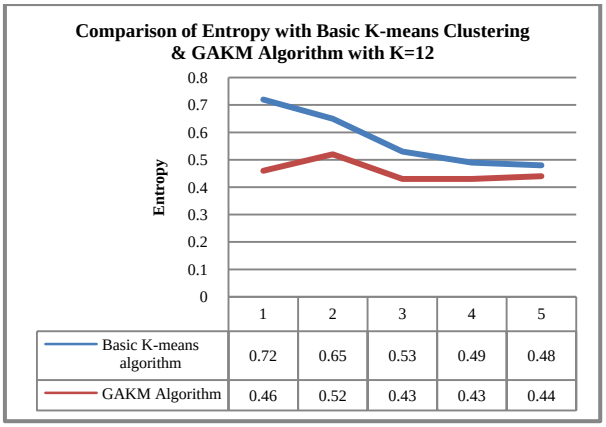

Fig. 6: Comparison of Entropy with Basic K-means Clustering & GAKM Algorithm with K=9

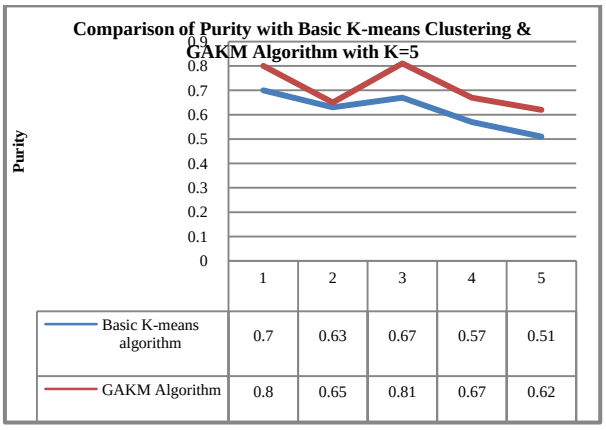

Fig. 7: Comparison of Purity with Basic K-means Clustering & GAKM Algorithm with K=5

Table 5: Purity (K=5, 9, 12)

	K= 5 (Purity)		K=9 (Purity)		K=12 (Purity)	
District No	**Basic K-means algorithm**	**GAKM Algorithm**	**Basic K-means algorithm**	**GAKM Algorithm**	**Basic K-means algorithm**	**GAKM Algorithm**
1	0.70	0.80	0.56	0.61	0.63	0.70
2	0.63	0.65	0.68	0.81	0.64	0.75
3	0.67	0.81	0.85	0.88	0.63	0.59
4	**0.57**	**0.67**	0.67	0.65	0.63	0.66
5	0.51	0.62	0.78	0.79	0.74	0.77

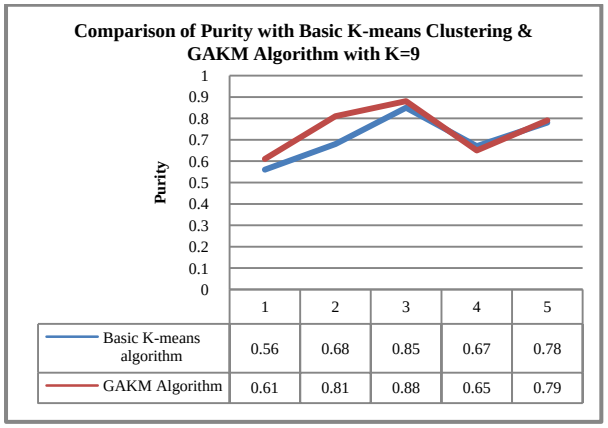

Fig. 8: Comparison of Purity with Basic K-means Clustering & GAKM Algorithm with K=9

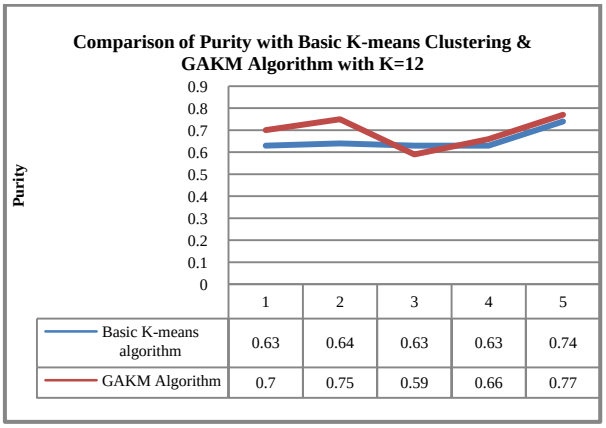

Fig. 9: Comparison of Purity with Basic K-means Clustering & GAKM Algorithm with K=9

Time Complexity

Time complexity is a important measure in comparing performance of any algorithm. Table 6 shows the time complexity of basic K clustering and GAKM clustering algorithm. Figure 10, 11 and 12 shows the graphical representation of the table 6 .

Table 6: Time Complexity

District No	K= 5 (Time Complexity in Sec)		K=9 (Time Complexity in Sec)		K=12 (Time Complexity in Sec)	
	Basic K-means algorithm	GAKM Algorithm	Basic K-means algorithm	GAKM Algorithm	Basic K-means algorithm	GAKM Algorithm
1	120	70	125	114	141	102
2	119	95	121	105	118	118
3	156	69	225	95	200	101
4	207	102	139	111	152	115
5	96	42	112	55	129	69

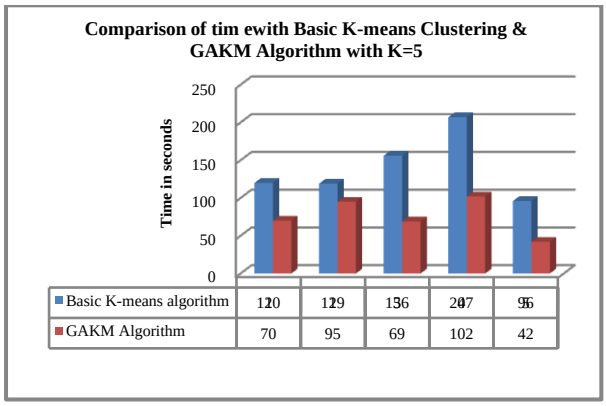

Fig. 10: Comparison of time with Basic K-means Clustering & GAKM Algorithm with K=5

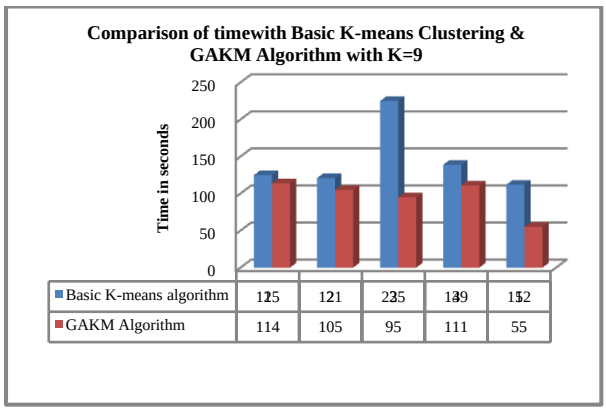

Fig. 11: Comparison of time with Basic K-means Clustering & GAKM Algorithm with K=9

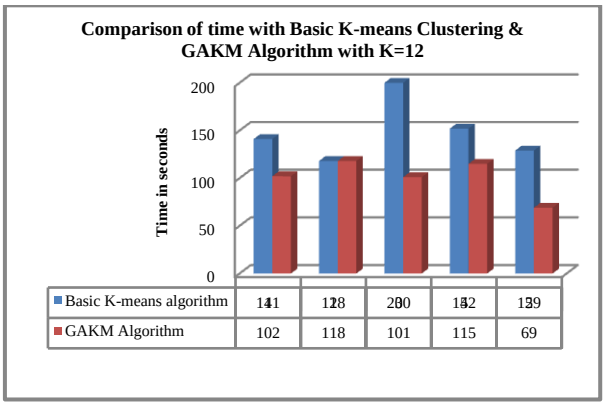

Fig. 12: Comparison of time with Basic K-means Clustering & GAKM Algorithm with K=12

9. CONCLUSION

The proposed GAKM (Genetic algorithm K-means) makes the K-means clustering algorithm gets rid to the dependence on the initial centroid point and reduces the time complexity. This algorithm is applied on CTS (child tracking survey) data set for testing and result proves the validity of the algorithm. It can be concluded by the results that engagement of children in agriculture work, engagement of children in grazing cattle, poor economic condition of parents and marriage of children in early age have influence over enrollment and dropout rates in Rajasthan. It indicates that Child Labour Act., Right to Education Act. and Child Marriage Act. are not being effectively implemented in Rajasthan state. If these three Acts are implemented effectively than school dropout and never enrollment due to engagement of children in agriculture work, engagement of children in grazing cattle, poor economic condition of parents and marriage of children in early age will certainly decreased to a great extent in Rajasthan. By controlling only these four variables, cent percent enrollment rate and zero dropout rate cannot be achieved because there are some educational variables which are also responsible for enrollment and dropout in elementary education in Rajasthan.

10. REFERENCES

[1] XIAO FENG LI., CHAN XIN., and LI LI YANG. 2010. Study of Data Ming Classification based on Genetic Algorithm", 20IO 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE), 978-1-4244-6542-2/$26.00 © 2010 IEEE

[2] BARAHATE SACHIN R and SHELAKE VIJAY M. 2012. A survey and future vision of data mining in educational field, 978-0-7695-4640-7/12,IEEE.

[3] A. K. Jain and R. C. DUBES. 1988. Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall..

[4] S. Z. SELIM and M. A. ISMAIL. 1984.K-means type algorithms: a generalized convergence theorem and characterization of local optimality, in IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 6, No. 1, 81--87.

[5] JIEMING ZHOU., J.G., and X.CHEN. 2009. An Enhancement of K-means Clustering Algorithm, in Business Intelligence and Financial Engineering, BIFE '09. International Conference on, Beijing.

[6] S. RAY., and R. H. TURI. 1999. Determination of number of clusters in k-means clustering and application in colour image segmentation, In Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques, 1999, 137-143.

[7] WANG, J., and X, SU. 2011. An improved K-means clustering algorithm, In 3rd International Conference on Communication Software and Networks (ICCSN), Xi'an.

[8] DONG, J., and M. QI. 2009. K-means Optimization clustering algorithm, in 3rd International Conference Algorithm for Solving Clustering Problem," in Second International Workshop on Knowledge Discovery and Data Mining (WKDD), Moscow.

[9] YOGITA CHAUHAN., VAIBHAV CHAURASIA., and CHETAN AGARWAL. 2014. A Survey Of K-Means And GA-KM The Hybrid Clustering Algorithm, International journal of scientific and technology research volume 3, issue 6, ISSN 2277-8616.

[10] JENN-LONG, LIU., YU-TZU HSU., and CHIH-LUNG. 2012.Development of Evolutionary Data Mining Algorithms and their Applications to Cardiac Disease Diagnosis", WCCI 2012 IEEE World Congress on Computational Intelligence.

[11] RAJASHREE DASH., and RASMITADAS. 2012. Comparative Analysis of K-means and Genetic Algorithm Based Data Clustering, International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 2, 257-265.

[12] MUKHERJEE D. 2011. Reducing out-of-school children in India, National University of Educational Planning and Administration, Delhi.

[13] BASUMATARY R. 2012. School dropout across Indian state and UTs: An econometric study, Int. Res. J. Social Sci., Vol. 1(4), 28-35.

[14] UNESCO www.unesco.org(2013)

[15] BOKOVA., and BUSH. 2012. Literacy is key to unlocking the cycle of poverty, at http://www.chron.com/ opinion/outlook/article /Literacy-is-key-to-unlocking-the -cycle-of-poverty-3848564. php (2012)