International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-1 Issue-2, June 2012

Genetic K-Means Algorithm – Implementation and Analysis

Sonia Sharma, ShikhaRai

Abstract: ***K-means algorithm is most widely used algorithm for unsupervised clustering problem. Though it is accepted but it has some problems which make it unreliable. Initialization of the random cluster centres, number of clusters and terminating condition play a major role in quality of clustering achieved. This paper empirically analyses a derived form [Krishna &Narasimha, 1999] of K-means using Genetic algorithm approach. The new algorithm prevents algorithm to converge towards local minima by considering a rich population of potential solutions. A tool that implements this algorithm is presented in the paper. The time complexity and execution expectation is also tested over an exhaustive set of data of different dimensions.***

Keywords; K-Means Clustering, Genetic Algorithm, Local Minima, Optimization.

I INTRODUCTION

Genetic algorithm is search heuristic usually applied in Optimization problems. The power of Genetic algorithm lies in its ability to perform parallel search in complex spaces. The behavior of genetic algorithms is highly influenced from natural evolution studied in biological sciences. In complex scenarios where the research problem under study involves a multi-dimensional search space and deterministic algorithms fail to meet time constraints, stochastic techniques like genetic algorithms are used. A robust approach can be applied in various searching problems in numerous domains like Pattern Recognition, Machine learning, VLSI etc. Data clustering is an unsupervised technique for discovering close groups (clusters) within data. Clustering has been a technique of a substantial interest in knowledge discovery in engineering and scientific domains like medicine, computer vision, remote sensing, psychology etc. The grouping is done in a manner such that the data patterns in a group are more similar to each other as compared to data pattern of another group. It uses a dissimilarity measure to achieve clustering that is defined based on the objective and organization of data. -means algorithm is a widely used clustering technique. It is one of the most established algorithms that have been used extensively in a variety of applications. The expected number of clusters (say, K) to be found is provided by the user. K-means uses a dissimilarity metric that guides the process towards optimal configuration.

Revised Manuscript Received on 30 June 2012.
* Correspondence Author
Mrs Sonia Sharma*, Department of Computer Science JMI, Radau, Yammuna Nagar (Haryana), India.
Miss ShikhaRai, Department of Computer Science JMI, Radau, Yammuna Nagar (Haryana), India.

II GENETIC K-MEANS

A. *GENETIC ALGORITHM*

Genetic Algorithm (GA) works on the coding of parameter set rather on the parameters themselves. These encoded parameters are known as chromosomes or solutions. The objective function value at a solution is the objective function value at a corresponding parameter.GA's solve optimization problems using a population of a fixed number, called the population size, of solutions. A solution consists of a string of symbols, typically binary symbols. GA's evolve over generations. During each generation, they produce a new population from the current population by applying genetic operators viz., natural selection, crossover, and mutation. Each solution in the population is associated with a figure of merit (fitness value) depending on the value of the function to be optimized. The selection operator selects a solution from the current population for the next population with probability proportional to its fitness value. Crossover operates on two solution strings and results in another two stings. Typical crossover operator exchanges the segments of selected stingsacross a crossover point with a probability. The mutation operator toggles each position in a string with a probability, called the mutation probability.

B. *K-Means Clustering*

K–Means (KMA) clustering is a method of cluster analysis which aims at portioning of n observations into k clusters. Each of the observation belongs to a cluster with the minimum distance between cluster centre and the observation point. It is done iteratively so that the observation point is at least distance from the centre of cluster. The mean distance between the cluster centre and observation is minimized during this iteration process.

The main problem with the KMA is that it does not guarantee to converge to global minima. It does not guarantee to achieve the best optimal solution available. Since stochastic optimization approaches are good at avoiding convergence to a local optima, these approaches could be used to find a globally optimal solution. For the purpose of finding the global minima we are using Genetic Algorithm (GA) which helps in finding the global minima.

C. *GENETIC K- MEANS CONCEPTS*

K-means algorithm is the most popularly used algorithm to find a partition that minimizes total within cluster variation measure.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

There are many variations of the KMA .We briefly explain below one of its simple variant that will be use in the development of GKA. KMA is an iterative algorithm. It start with a random configuration of cluster centers. In every iteration, each pattern is assigned to the cluster whose center is the closest center to the pattern among all the cluster centers. The cluster centers in the next iteration are the centroids of the patterns belonging to the corresponding clusters. The algorithm is terminated when there is no reassignment of any pattern from one cluster to another or the variation measure ceases to decrease significantly after an iteration. A major problem with this algorithm is that it is sensitive to the selection of initial partition and may converge to a local minimum of variation if the initial partition is not properly chosen.

The genetic operators that are used in GKA are the selection, the distance based mutation and the K-means operator. In this section we explain GKA by specifying the coding and initialization schemes and, the genetic operators.

1) *Coding*: The natural way of coding such w into string is to consider a chromosome of length n and allow each allele in chromosome to take values from {1,2,…..,k}.
2) *Initialization*: Way of selecting initial population is random. Each allele in the population can be initialized to cluster number selected from uniform distribution over the set {1,2,…,k}.
3) *Selection*: Selection operator randomly select a chromosome from the previous population.
4) *Mutation*: The Mutation changes an allele value depending on the distances of the cluster centroids from the corresponding data point. It may be recalled that each allele corresponds to a data point and its value represents the cluster to which the data point belongs operator is defined such that the probability of changing an allele value to a cluster number is more if the corresponding cluster center is closer to the data point
5) *KMO*:A algorithm with the above selection and mutation operators may take more time to converge, since the initial assignments are arbitrary and the subsequent changes of the assignments are probabilistic. Moreover, the mutation probability is forced to assume a low value because high values of Pm lead to oscillating behavior of the algorithm. To improve this situation, a one-step K-means algorithm, named K-means operator (KMO), is introduced.The following two steps constitute KMO on string which yields best solution :

a) calculate cluster centers for the given matrix W;
b) reassign each data point to the cluster with the nearest cluster center and thus form best solution.

In this approach our main aim is to minimize the Total Within Cluster variation i.e TWCV. This is taken into consideration by first finding the pattern belonging to the cluster and then secondly the pattern being used. Then the TWCV being denoted as S(W) can be shown as:

$$S(W) = \sum_{k=1}^{K} S^{(k)} = \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik} \sum_{j=1}^{d} (x_{ij} - c_{kj})^2$$

$$S^{(k)}(W) = \sum_{i=1}^{n} w_{ik} \sum_{j=1}^{d} (x_{ij} - c_{kj})^2$$

III IMPLEMENTATION

The implementation is done C++ and made using GCC compiler. It is tested over Windows XP, Windows 7 and Ubuntu Linux. The object oriented design is followed.

A. *Class Design*

The purpose and the function being implemented by the class is shown below:

i) GA.h- Genes Initialization, Population Fitness Calculation, Rank Population, Create Next Generation, Mutate Population, KMO Population.
ii) Genome.h- Calculate Gene Fitness, TCWV Calculation, Mutate Gene, KMO Gene.
iii) GAData.h- Load Data Points.
iv) Test.cpp- Test class (main function).

B. *Results & Analysis*

The code is tested for execution on a Intel(R) Core(TM) I3 CPU M 380 @ 2.53, 3.00GB Installed memory, 64 –bit OS system.

The results are shown in the Appendix at the end of this paper. We tested the toolbox over different scenarios and presented them graphically at the end.

In first analysis, we run the code over 2D and 3D data for 500 points, with initial population of 500 chromosomes. We ran it over 100 generations and measured the execution times. As we expected the relation comes out to be linear with number of points. The graphical representation is presented in Figure 1 in Appendix.

Number of Points	Execution times for 2D data (ms)	Execution times for 2D data(ms)
50	1670	1934
100	2402	2574
150	2995	3182
200	3728	4024
250	4321	4633
300	4945	5444
350	5600	6037
400	6225	6989
450	6754	7316
500	7410	7972

i

In second analysis, we were interested in assuring the nature of relation between execution times with number of generations to be run. As expected it varied linearly with generations. The graph is shown as Figure 2 in Appendix.

Number of Generations	Execution time for 2D data (ms)
100	1997
200	3354
300	4665
400	6053
500	7378

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

A demonstration is also presented in the Figure 3. It shows the points on the left hand side which after a few generations are grouped as clusters. These clusters are shown in different colours. The right sections shows how the Standard Error (TWCV), reduces as the generation progresses. There could be some minor perturbations as mutation comes into action.

IV CONCLUSION

The toolbox provided is a first step towards solving some clustering problems in domain of Image processing. To apply clustering on image features(vectors of size upto 128 dimensions) the clustering algorithm needs to be robust, scalable and stable. This toolbox intends to help us not only in engineering problems in Image Processing, but also in Financial data, data mining applications, etc. We intend to make some possible adjustments in the design of code to make it faster and more appropriate.

APPENDIX I - RESULTS

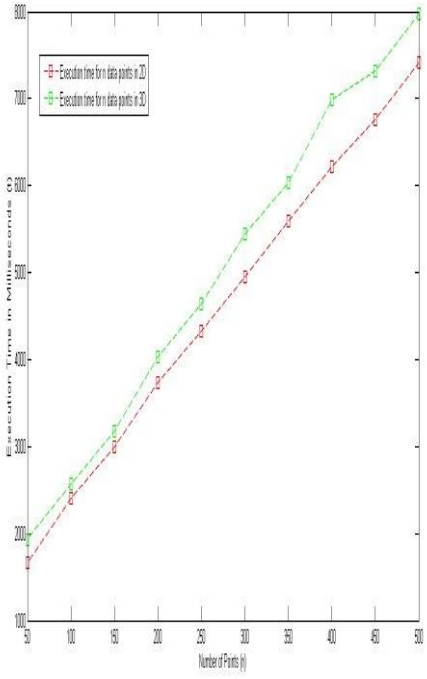

Figure 1: Analysis for 2D and 3D data over Population Size=500, Generations100, Mutation Rate=0.02

V ACKNOWLEDGMENT

The authors are quite thankful to the Computer science department of JMIT ,Radaur college for supporting our work.

REFERENCES

1. Krishna K, Murty M " Genetic K-means algorithm" ,IEEE Transactions on Systems, Man and Cybernetics ,Part B: Cybernetics 1999 , 29:433-439.
2. William A. Greene "Genetic Algorithms For Partitioning Sets" University of New Orleans New Orleans, LA 70148,2000
3. **[3]**Monica Chi," Evolutionary Hierarchical Clustering Technique",2001
4. N. Sujatha," Refinement Of Web Usage Data Using Clustering From K Means Using GenticAlgorithm",Research Scholar, Department of Computer Science Madurai Kamaraj University, Madurai,European Journal of Scientific Research , 2010
5. Hall, L.O. Ozyurt, I.B. Bezdek, J.C," Clustering With Genetically Optimized Approach",1999
6. R. J. W. Hodgson,"Genetic Algorithm Approach to Particle Identification by Light Scattering",2000

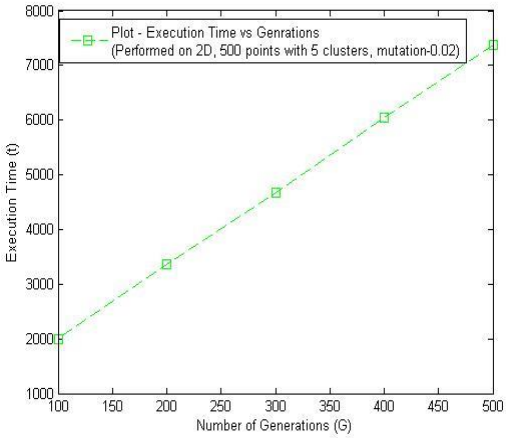

Figure 2: Execution Time vs. Number of Generations - Shows a linear behaviour

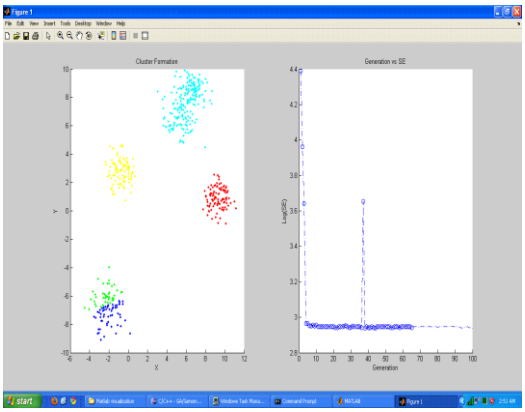

Figure 3: Execution Results for Population Size=500, Generations100, Mutation Rate=0.02. Left section shows points, color represents clusters. Right sections shows the graph between Standard Error and Generations.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

AUTHOR PROFILE

SONIA SHARMA received her B.Tech degree in Computer Engineering from Seth Jai Parkash Institute Of Engineering and Technology, Radaur, Haryana,India in 2005, and received her M.Tech. degree in Computer Science and Engineering from Seth Jai Parkash Institute Of Engineering and Technology, Radaur, Haryana, India in 2007. Her area of interest includecryptography and Genetic algorithm.

SHIKHA RAIreceived her B.Tech degree in Computer Engineering from N.C College of Engineering, India in 2006, and received her M.Tech. degree in Computer Science and Engineering from Seth Jai Parkash Institute Of Engineering and Technology, Radaur, Haryana, India in 2012. Her area of interest include Genetic Algorithm

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)