An Extended K-Means Clustering with Genetic Algorithm and Min-Max Approach

Asst.Prof.VibhutiP. Patel[1] Asst.Prof.Avani P. Patel[2] Asst.Prof.Kushal D. Patel[3] Asst.Prof.Mihir A. Mishra[4] Asst.Prof.Pragnesh A. Patel[5]

[1,2,3,4,5]Assistant Professor

[1,2,3,4,5]GIDC Degree Engineering College - Navsari

Abstract— Clustering is one of the major data mining task that is division of data object into similar group; each similar group is called cluster. Object in the cluster are similar to each other and dissimilar with different cluster. It can be implemented by number of approaches. K means is one of the popular techniques for the clustering. Major drawbacks of the K means clustering algorithm are cluster depends on initial centroid and predefined value of k (cluster number). From the previous research work it has been found that GA (Genetic Algorithm) can be used to solve clustering problem and Min-Max approach can be used to improve robustness. This work proposes an extended clustering algorithm that combines the GA and K-means with Min-Max approach.

Key words: Min-Max Approach, Genetic Algorithm

I. INTRODUCTION

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

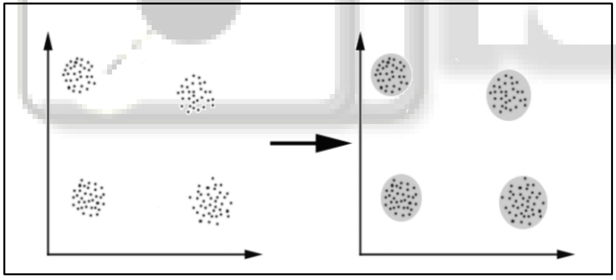


Fig. 1: Shows that how to Cluster the data[19]

This shows graphical example of clustering. In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are “close” according to a given distance.

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data, It can be shown that there is no absolute “best” criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection).So clustering having several methods like partition based method, hierarchical methods, distance based method, density based methods etc.

A. K-Means Clustering

K-means is one of simplest algorithm to understand. The main idea of K-means is summarized into below steps,

1) Randomly choose k objects from data set as the initial cluster centers.
2) Assigns each object to the cluster to which it is associated closely by considering the distance from the given centroid.
3) Compute the new position of each centroid by the mean value of object in cluster.
4) Repeat step 2 & 3 until the points stop moving, i.e. the mean squared error converges.

K-means algorithm is very sensitive in terms for selection of initial means. The K-means method is applied when the mean of a set of object is defined .Disadvantage of K-means is that, there is no specific answer for find the minimum number of clusters for any given data set.

B. Genetic Algorithm

Genetic algorithm (GA) has been proposed by many researchers to solve a global solution for clustering problem. Genetic algorithm is a computational abstraction of biological evolution that can be used to solve difficult optimization problems. It iteratively applies a series of genetic operators such as selection, crossover, and mutation to a group of chromosomes where each chromosome represents a solution to a problem.

The initial set of chromosomes is selected randomly from solution space. Genetic operators combine the genetic information of parent chromosomes to form a new generation of the population; this process is known as reproduction. Each chromosome has an associated fitness value, which quantifies its value as a solution to the problem. A chromosome representing a better solution will have a higher fitness value. The chromosomes computed to reproduce based on their fitness value, thus the chromosomes representing better solution have a higher chance of survival. After many generations, a chromosome, which has the maximal fitness value, is the best solution for the problem.

II. PROBLEM DEFINITION

K-means is used for data clustering but it has some drawbacks as predefine cluster number and repetition process for finding the centroid so with the Genetic algorithm it can improve result so here in this proposed system combination of GA and K-means with Min-Max approach that is improving the result of Clustering.

III. PROPOSED K-MEANS ALGORITHM WITH GA AND MIN-MAX APPROACH

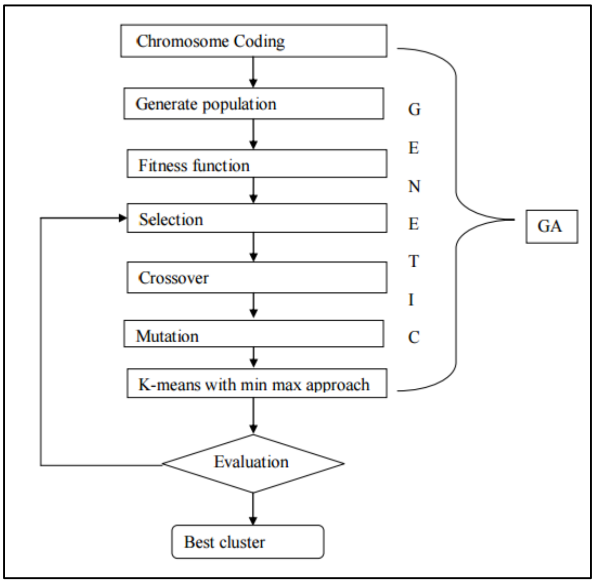


Fig. 2: shows the architecture of proposed system.

A. Chromosome Coding

A chromosome represents a possible solution for clustering. Many ways to code chromosome.

B. Generate Population

GA proceed to generate population of solutions randomly and, then improve it repetitive application of Selection, crossover, mutation and K-means with min-max approach. Suppose for example "ABCBCAB" is 7 data points in cluster than First and sixth points in Cluster A, second, fourth and seventh in cluster B and third and fifth in cluster C.

C. Fitness Function

Main objective of K-means is minimizing the sum of squared distances. In this proposed system uses the Euclidian distance between the data as fitness function.

$$\min \sum_k \sum_{i \in C} \| X_i - C_k \|^2$$

Where C_k iscenter. In this proposed system uses the Euclidian distance between the data as fitness function.

$$d(X,Y) = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \cdots (X_n - Y_n)^2}$$

D. Genetic Operators

1) Selection:

This process choose good individual from the population. The selection is based on fitness function and better solution give higher chance to survival.

2) Crossover:

It combines two or more parent chromosome to generate one more offspring.

For example:
Parent 1: 1101100100110110
Parent 2: 1101111000011110
Generate offspring:
Offspring 1:11011 11000011110
Offspring 2: 11011 00100110110

In this new algorithm two point crossover is used.

3) Mutation:

It change randomly one point from new offspring .It tkes a solution out of a local optimum and moves it towards global optimum by replacing one of the data point.

4) K-means with Min-Max Approach

After these all steps new data points offspring will generate on this k-means will apply with Min-Max approach. Iteration step will apply here and chromosome assign to the closest centroid. In this proposed system instead of traditional k-means iteration step, use Min-Max approach to find closest centroid. Min-Max approach: Our method starts from a randomly picked set of centers and set the minimum and maximum boundary so according to that it is try to find closest centroid. In code of Min-Max K-means: - private static double min value - private static double max value At the display time cluster value depends on thus boundary value and according to that it will display the results. Proposed System

IV. ALGORITHM

1) Do chromosome coding from data points
2) Randomly Generate initial population.
3) Evaluate the fitness of individual of initial population.
4) Produce new population by selection, crossover and mutation operator and evaluate fitness of individuals of new population.
5) Instead of iteration step of traditional k-means proposed system uses Min-Max approach to find the closest center of individual.
6) Output: It shows clustering result.

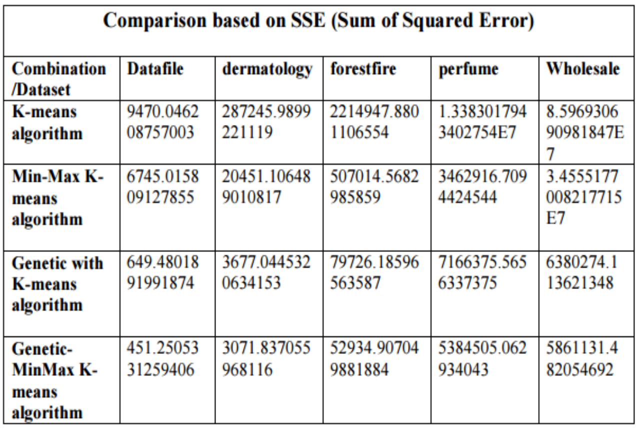
V. DATASETS

1) Dermatology: It is composed of 366 patient records that suffer from six different types of the Eryhemato-squamous disease. Each patient is described by both clinical and histopathological features.
2) Perfume data set: It is composed of 560 perfume records. The data was obtained from 20 different perfumes by using a handheld odor meter (OMX-GR sensor).
3) Forest fire dataset: This dataset is available on UCI Machine learning repository which contains 517 instances and 13 attributes. There is no missing value in data set.
4) Wholesale Customers: The data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units on diverse product categories.

Datasets	Instances	Attribute	Attribute Characteristic
Dermatology	366	33	Categorical, Integer
Perfume	560	6	Integer
Forest Fire	517	13	Real
Wholesale Customers	440	8	Integer
Datafile	75	5	Integer

Table 1: Result

VI. Result



Comparison based on SSE (Sum of Squared Error)					
Combination /Dataset	**Datafile**	**dermatology**	**forestfire**	**perfume**	**Wholesale**
K-means algorithm	9470.046208757003	287245.9899221119	2214947.8801106554	1.3383017943402754E7	8.596930690981847E7
Min-Max K-means algorithm	6745.015809127855	20451.106489010817	507014.5682985859	3462916.7094424544	3.4555177008217715E7
Genetic with K-means algorithm	649.4801891991874	3677.0445320634153	79726.18596563587	7166375.5656337375	6380274.113621348
Genetic-MinMax K-means algorithm	451.2505331259406	3071.837055968116	52934.907049881884	5384505.062934043	5861131.482054692

Table 2: Result

VII. Conclusion

The Clustering speed of traditional K-Means method is fast and it can deal with big dataset. This work solves the problems with K-Means clustering. It improves the result with Min-Max Approach and GA solves local optima problem. First data is given to the Genetic algorithm that passes from Genetic Function then it is given to K-Means with Min-Max approach. Compared to traditional k-means, this new algorithm gives lesser sum of squared error.

VIII. Future work

K-Means clustering with Min-Max approach gives robust result for clustering problem. Traditional K-Means combined with Genetic algorithm gives global solution, in future speed of genetic algorithm can be improved. Min-Max approach of K-Means can be combined with kernel-based clustering so that non-linear cluster can be detected in the data. This new approach can be applicable to categorical dataset.

Acknowledgment

I would like to take this opportunity to express our sincere thanks to several people without whose help and encouragement, it would be unfeasible for me to carry out this work. I would like to express my sincere thanks to Prof.Avani Patel, Prof.Mihir Mishra, Prof.Kushal Patel, Prof.Pragnesh Patel for their guidance, encouragement and support which were always the inspiration for me during the work. Warmly Thanks to them for constant inspiration and encouragement. They all are always been present whenever i needed the support. I would like to thank my parents for their valuable support and encouragement.

References

[1] K. A. Abdul Nazeer, M. P. Sebastian” Improving the Accuracy and Efficiency of the k-means Clustering Algorithm” Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K.

[2] Dharmendra K Roy and Lokesh K Sharma”GENETIC K-MEANS CLUSTERING ALGORITHM FOR MIXED NUMERIC AND CATEGORICAL DATA SETS”InternationalJourna of Artificial Inteligence&Application,vol.1,No.2,April 2010

[3] Omnia Ossama, Hoda M.O. Mokhtar *, Mohamed E. El-Sharkawi” An extended kmeans technique for clustering moving objects” Egyptian Informatics Journal (2011) 12, 45–51

[4] Youguo Li, Haiyan Wu” A Clustering Method Based on K-Means Algorithm” 2012 International Conference on Solid State Devices and Materials Science

[5] BinLu, FangyuanJu, “An optimized genetic K-means clustering algorithm” 2012 International Conference on Computer Science and Information Processing (CSIP)

[6] Xiaoqing Lin, Wei Zheng, Dongyang Jiang,” Research on Selection of Initial Center Points Based on Improved K-means Algorithm” 2012 2nd International Conference on Computer Science and Network Technology

[7] Piyaphol Phoungphol, Inthlr A Srivrunyoo,” Boostinggenetic Clustering ALGORITHM” Proceedings of the 2012 International Conference on Machine Learning and Cybernetics, Xian, 15-17 July, 2012

[8] Yang Yong, ”The Research of Imbalanced dataset of sample sampling method based on k-means cluaster and genetic algorithm”2012 International Conference On Future Electrical Power and Energy Systems

[9] G.Kiran Kumar, T. Bala Chary, “A New and Efficient K-Means Clustering Bibliography 56 DDU/M.Tech/ CE/13MTPBS002 Algorithm”, International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 11, November 2013

[10] Chintan Shah1 and Anjali Jivani2,” Comparison of Data Mining Clustering Algorithms” 2013 Nirma University International Conference on Engineering (NUiCONE)

[11] Rudolf Scitovski⇑ , Kristian Sabo, Analysis of the k-means algorithm in the case of data points occurring on the border of two or more clusters, Department of Mathematics, University of Osijek, TrgLj. Gaja2013

[12] Anupama Chadha, Suresh Kumar” An Improved K-Means Clustering Algorithm: A Step Forward for Removal of Dependency on K”, 2014 International Conference on Reliability, Optimization and Information Technology -ICROIT 2014, India, Feb 6-8 2014

[13] GrigoriosTzortzis n, AristidisLikas,” The MinMax k-Means clusteringalgorithm”, Department of Computer Science & Engineering, University of Ioannina,2014

[14] Data Mining-Clustering,JERZY STEFANOWSKI Institute of Computing Sciences Poznan University of Technology,Poznan, Poland

[15] Introduction to Clustering Methods, http://www.jmp.com/

[16] An introduction to data clustering,Jean-Karim Hériché EMBL 18 May 2012

[17] The k-means algorithm (Notes from: Tan, Steinbach, Kumar + Ghosh)

[18] https://en.wikipedia.org/wiki/Cluster_analysis

[19] https://en.wikipedia.org/wiki/Data_mining