ISSN 2320 2602

Volume 1, No.1, November – December 2012
International Journal of Advances in Computer Science and Technology
Available Online at http://warse.org/pdfs/ijacst05112012.pdf

A Proficient Accomplishment of Datamania of Genetic Algorithm by applying K-means Clustering

S.Mythili
Research Scholar, Department of Computer Applications
Hindusthan College of Arts & Science, Coimbatore, India
poovanthikaa@gmail.com
A.V. Senthil Kumar
Department of Computer Applications
,Hindusthan College of Arts & Science, Coimbatore, India

Abstract: The existing clustering algorithm has a sequential execution of the data. The speed of the execution is very less and more time is taken for the execution of a single data.To overcome this Parallel Implementation of Genetic Algorithm using K-Means Clustering (PIGAKM) is proposed but it has some more problems to retrieve the output without the error parameter. A new algorithm "A Proficient accomplishment of Datamania of Genetic Algorithm by applying K-means clustering" (PADGAKM) is proposed to overcome the problems in PIGAKM. PADGAKM is inspired by using KM clustering over GA. This process indicates that, while using KM algorithm, it covers the local minima and it initialization is normally done randomly, by KM and GA. It always converge the global optimum eventually and groups all the data by KM. To speed up GA process, the evaluation is done parallely by grouping the similar set of data. To show the performance and efficiency of these algorithms, the comparative study of this algorithm has been done.

Keywords: Clustering, Genetic algorithm, K-means Clustering, Crossover, Mutation, PIGAKM.

INTRODUCTION

Data mining [1] is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Clustering [3]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 the other cluster.

K-means clustering [6] is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. The problem is computationally difficult, however there are efficient algorithms that are commonly employed and converge fast to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data, however *k*-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

Genetic algorithm, [18]a population of strings (called chromosomes or the genotype of the genome), which encode candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem, evolves toward better solutions. Traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible. The evolution usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated, multiple individuals are stochastically selected from the current population (based on their fitness), and modified (recombined and possibly randomly mutated) to form a new population.

REVIEW LITERATURE

Clustering

Clustering [5]is grouping input data sets into subsets, called 'clusters' within which the elements are somewhat similar. In general, clustering is an unsupervised learning task as very little or no prior knowledge is given except the input data sets. The tasks have been used in many fields

and therefore various clustering algorithms have been developed.

Clustering task is, however, computationally expensive as many of the algorithms require iterative or recursive procedures and most of real-life data is high dimensional. Therefore, the parallelization of clustering algorithms is inevitable, and various parallel clustering algorithms have been implemented and applied to many applications.

A review a variety of clustering algorithms and their parallel versions as well. Although the parallel clustering algorithms have been used for many applications, the clustering tasks are applied as pre-processing steps for parallelization of other algorithms too. Therefore, the applications of parallel clustering algorithms and the clustering algorithms for parallel computations.

The goals of clustering

The goal of clustering [3] 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, we could be interested in finding representatives for homogeneous groups, in finding "natural clusters" and describe their unknown properties, in finding useful and suitable groupings in finding unusual data objects.

Stages in clustering

Clustering project can be runned by 3 stages.

Stage 1: Forming the Cluster

Identifies the members to analyse the results of diagnostic survey and prepare reports for next meeting.

Stage 2: Finding the Focus

Review the results of survey. It identify ways to deal offline with things that aren't a priority for everyone.

Stage 3: Developing the work plan

Review finding from Resource/Feasibility Investigation. It allocates work and resources and agree timeline.

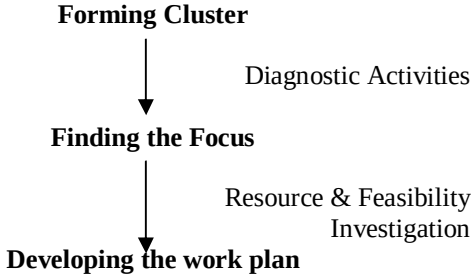

Fig 1: Stages in clustering

Euclidean distance

The **Euclidean distance** or **Euclidean metric** is the "ordinary" distance between two points that one would measure with a ruler.The **Euclidean distance** between points **p** and **q** is the length of the line segment connecting them (**pq**).

If $\mathbf{p} = (p_1, p_2,..., p_n)$ and $\mathbf{q} = (q_1, q_2,..., q_n)$ are two points in Euclidean *n*-space, then the distance from **p** to **q**, or from **q** to **p** is given by:

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

The distance between points **p** and **q** may have a direction (e.g. from **p** to **q**), so it may be represented by another vector, given by

$$\mathbf{q} - \mathbf{p} = (q_1 - p_1, q_2 - p_2, \cdots, q_n - p_n)$$

One dimension

In one dimension, the distance between two points on the real line is the absolute value of their numerical difference. Thus if *x* and *y* are two points on the real line, then the distance between them is given by:

$$\sqrt{(x - y)^2} = |x - y|.$$

In one dimension, there is a single homogeneous, translation-invariant metric (in other words, a distance that is induced by a norm), up to a scale factor of length, which is the Euclidean distance. In higher dimensions there are other possible norms.

Two dimensions

In the Euclidean plane, if $\mathbf{p} = (p_1, p_2)$ and $\mathbf{q} = (q_1, q_2)$ then the distance is given by

$$\mathbf{d}(\mathbf{p},\mathbf{q}) = \sqrt{(p_1 - q_1)^2 - (p_2 - q_2)^2}.$$

Alternatively, it follows from the coordinates of the point **p** are (r_1, θ_1) and those of **q** are (r_2, θ_2), then the distance between the points is

$$\sqrt{r_1^2 + r_2^2 - 2r_1 r_2 \cos(\theta_1 - \theta_2)}.$$

Three dimensions

In three-dimensional Euclidean space, the distance is

$$d = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}.$$

N dimensions

In general, for an n-dimensional space, the distance is

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_i - q_i)^2 - ... + (p_n - q_n)^2}.$$

Squared Euclidean Distance

The standard Euclidean distance can be squared in order to place progressively greater weight on objects that are further apart. In this case, the equation becomes

Squared Euclidean Distance is not a metric as it does not satisfy the triangle inequality, however it is frequently used in optimization problems in which distances only have to be compared.

K-means Clustering

K-means clustering [2] is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. The problem is computationally difficult, however there are efficient algorithms that are commonly employed and converge fast to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data, however *k*-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

Algorithm

Input:

D={d1,d2,…dn} // sets of n data items.

K // number of desired clusters

Output:

A set of k clusters.

Steps:

1. Choose k data items from D as initial centroids;
2. Repeat the process of selecting the items

Assign the each item di to the cluster which has the nearest and the suitable centroids;

Calculate the new mean value for each cluster;

Repeat the process until the criteria is satisfied.

Genetic algorithm

In a genetic algorithm,[21] a population of strings (called chromosomes or the genotype of the genome), which encode candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem, evolves toward better solutions. Traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible. The evolution usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated, multiple individuals are stochastically selected from the current population (based on their fitness), and modified (recombined and possibly randomly mutated) to form a new population. [19]The new population is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population. If the algorithm has terminated due to a maximum number of generations, a satisfactory solution may or may not have been reached.

Working of Genetic Algorithm

A. Initialization

Initial many individual solutions are (usually) randomly generated to form an initial population. The population size depends on the nature of the problem, but typically contains several hundreds or thousands of possible solutions. Traditionally, the population is generated randomly, allowing the entire range of possible solutions (the search space). Occasionally, the solutions may be "seeded" in areas where optimal solutions are likely to be found.

B. Selection

During each successive generation, a proportion of the existing population is selected to breed a new generation. Individual solutions are selected through a fitness-based process, where fitter solutions (as measured by a fitness function) are typically more likely to be selected. Certain selection methods rate the fitness of each solution and preferentially select the best solutions. Other methods rate only a random sample of the population, as the former process may be very time-consuming.

C. Reproduction or Crossover

The next step is to generate a second generation population of solutions from those selected through genetic operators: crossover (also called recombination).

For each new solution to be produced, a pair of "parent" solutions is selected for breeding from the pool selected previously. By producing a "child" solution using the above methods of crossover and mutation, a new solution is created which typically shares many of the characteristics of its "parents". New parents are selected for each new child, and the process continues until a new population of solutions of appropriate size is generated. Although reproduction methods that are based on the use of two parents are more "biology inspired", some research suggests that more than two "parents" generate higher quality chromosomes.

These processes ultimately result in the next generation population of chromosomes that is different from the initial generation. Generally the average fitness will have increased by this procedure for the population, since only the best organisms from the first generation are selected for breeding, along with a small proportion of less fit solutions, for reasons already mentioned above.

Although Crossover and Mutation are known as the main genetic operators, it is possible to use other operators such as regrouping, colonization-extinction, or migration in genetic algorithms.

D. Termination or Mutation

This generational process is repeated until a termination condition has been reached. Common terminating conditions are:

- A solution is found that satisfies minimum criteria

- Fixed number of generations reached
- Allocated budget (computation time/money) reached
- The highest ranking solution's fitness is reaching or has reached a plateau such that successive iterations no longer produce better results
- Manual inspection
- Combinations of the above

PROPOSED METHOD

In an existing Genetic Algorithm, there is only one string in each execution and all the genetic operations are done on it. There is a problem in existing system between the execution and the solution quality and the results given. In the existing system the evaluation is taken place many times and the results obtained will not be correct. To get the solution of the problem, it takes lot of time to complete. The evaluation of the process in done parallel by grouping the similar items together. The similar items are executed parallel. By this parallel execution two data can be run at a time. So the time consuming to take the execution will be reduced.

Genetic algorithm works with the individual string which is executed individual by the several processors."A Proficient Accomplishment of Datamania of Genetic Algorithm by applying K-means Clustering"(PADGAKM).this technique specifies the grouping of the similar string and executing the similar string simultaneously.

This method presents a "Proficient Accomplishment of Datamania of Genetic Algorithm by applying K-Means Clustering (PADGAKM) which uses multiple substrings within single dynamic parameters. Simple genetic algorithm involves only one initial string with is fixed genetically operational parameters selected in advance and it requires more time for distance calculations and crossovers in each generation than K-means needs in one iteration. The technique is proposed in the paper Parallel Implementation of Genetic algorithm using K-means clustering has some drawback that it is executing the strings parallelly according to the order it is given.

In this technique the substrings are grouped into a string and the genetic parameter is created for that string. Then the strings are executed parallelly. The groupings of the strings are done by applying K-Means clustering. The crossover and the mutation probability specify the easy way to group the similar substrings that are matched with the substrings one another.

The substrings which are grouped together will have the same genetic parameter, which helps to evaluate the process quickly as well as error free. For the single execution group of the similar substrings are executed.

The working of this proposed system can be viewed easily by the flow chart represented below. The initial string is divided into number of sub strings, the substrings are verified in the interrelation process. The similar strings are grouped and then the string is taken to the evaluate the fitness of the grouped string. The condition is checked, and then the strings are migrated. Genetic parameter of the string is specified and the solution is given.

Step 1: Code the problem with the parameters as a form of string. Then use binary code method to transfer the parameters from the problem space into coding space

Step 2: Defines the individual string function: The string is compared with another string and it will be sorted by reading the fitness of each substring, the individuals are sorted by objective values , pi denotes the order of individual i , denoted by the string (i) is given by :

Step 3 : Parameter design : The maximum value of the total number generations is denoted by T; the size of the each substring is denoted by N ; the number of the substring id denoted by M ; the rate of migration id represented by R ; the probability of selection is denoted by S ; and the probabilities of crossover and mutation, donated by C and M respectively. Before the execution of the process, there is no guidance used to determine the values of M and N. The probability of selecting the i individual depends on the rate si, which is proportional to its degree of fitness, that is

$$\mathbf{S_i} = \mathbf{fitness\ (i)} / \sum \mathbf{fitness\ (i)}$$

Where t is the number of the current generation, aj and bj are the initial values mj and cj of for the j-th substring respectively, Tj is a scaling constant number that is larger than or equal to T.

Step 4: Create initial string randomly.

$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}.$$

Here d represents the distance of each sub strings.

Step 5: Decode string and evaluate individual substring.

Step 6: Transfer information between substring and exchange their individuals.

Step 7: Calculate crossover probability and mutation probability of the each string and adjust the functions of each string in the datasets.

Step 8: Genetic algorithm is performed for grouping the similar sub strings ny using including selection, crossover and mutation.

Step9: Process will complete when the termination condition is satisfied.

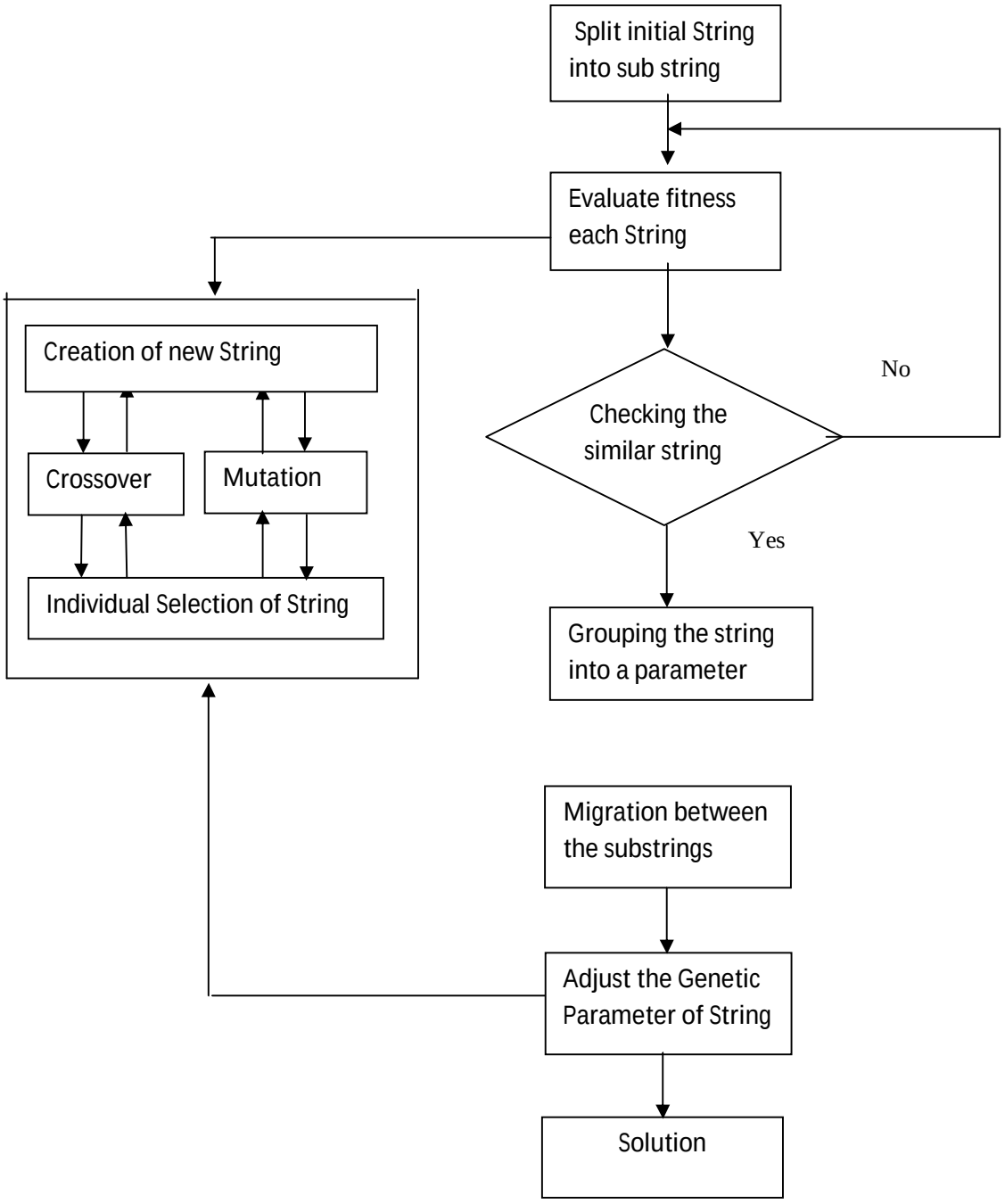

Fig 2: Structure of Proficient Accomplishment of Datamania with Genetic Algorithm by applying k-Means Clustering

EXPERIMENTAL RESULTS

Datasets

The datasets are used in these experiments are Iris and Iymphoma. Moreover, data files used in these experiments are chosen among a huge variety given by MAT LAB.

Dataset1 is made based on a mathematical model to form their clusters with small amount of points interleaving. Dataset1 consists of several points with the 0.2 points interval. The first interval may starts with (0.125, 0.25) and the points may move with the intervals (0.625,0.25),(0.375,0.75),(0.875,0.75) and so on. The first two points have the horizontal interleaving on the boundary. In addition, points 2 and 3 have the same boundary. This datasets are clustering into 4 different clusters.

The Iris dataset used as the Dataset2. It is called as Anderson's Iris data because the Edgar Anderson was the person collecting the data to quantify the geographic variation of Iris Flowers in the Graspe Peninsula. The Iris flowers have 3 specifications.

1. Iris Setasa
2. Iris Virginica
3. Iris Versicolor.

k-means clustering result for the Iris flower data set and actual species

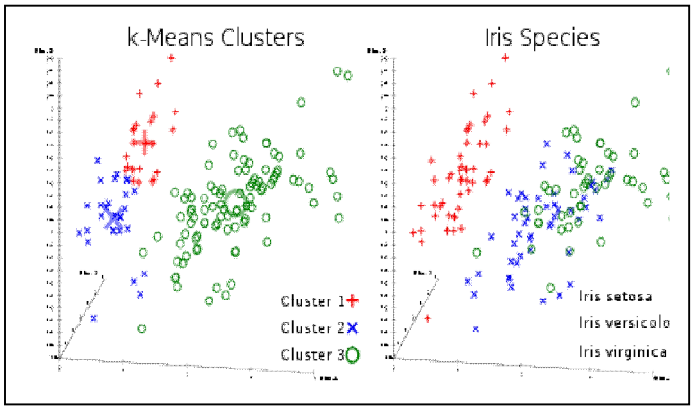

Fig 3: Clustering of data **Fig 3**: Clustering of Iris Dataset

4 features were measured from every sample; they are the length and the width of the petals in centimetres. Based on these 4 features the four values of the data are accepted. It is used as a typical test for many classification techniques. In the proposed method the Dataset2 is the Iris dataset.

The Lymphoma is used as the Dataset3. This dataset has 4 continuous features which is to be taken into consideration. These data strings are taken as the training sets and it is grouped into 3 clusters.

The different types of genes are clustered into 3 groups and the remaining are kept as the samples. In the grouped lymphoma the datasets are tested and the average rate and average time taken to execute the process can be calculated.

The result that has been obtained by developing PIGAKM has checked by giving the artificial datasets. The datasets is based on the mathematical model form their clusters with small amount of points interleaving. The dataset1 is the artificial dataset; Iris and Lymphoma are the two real-life data sets.

Table1 shows the results that have been obtained using these three datasets. Moreover Average Error and Average Time are listed to view the execution between them. In the table KM shows the K-means Clustering , KM (5) shows the 5 iterations of the datasets , KM(10) shows the 8 iterations of the datasets ,for the each iteration the average time will be reducing, so that the execution will be done quickly.

In this table GAKM shows the grouping of the datasets which are similar so that the processing time can be done quickly. The average rate and average time is represented as ARAT , the average time is reduced by grouping the similar datas and executing it parallelly.

PIGAKM shows the previously solved problem and its solution. To get the accurate result and the best solution is given in PADGAKM with the most efficient manner.From this table it is clearly view that the proposed system PADGAKM shows very less error rate compared with the existing system PIGAKM.

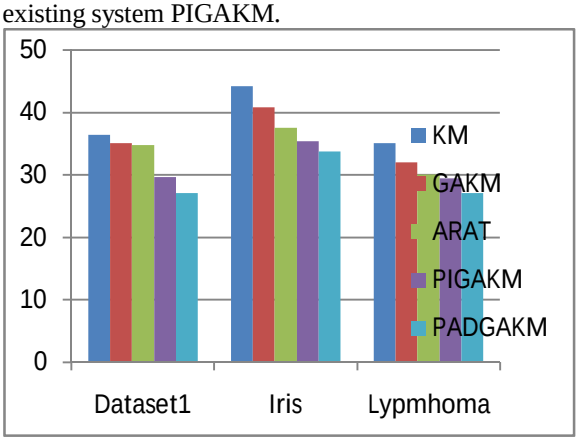

Fig 4: Average Error Parameter

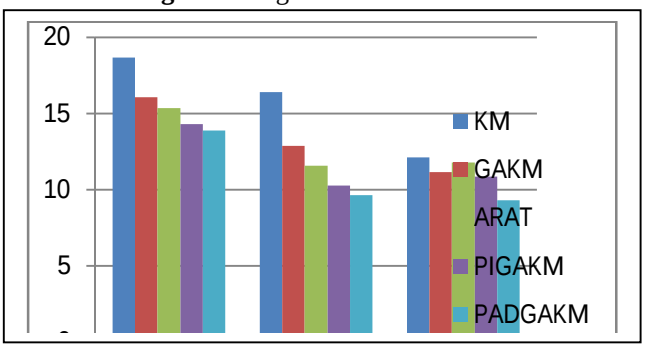

Fig 5 : Average Time Parameter

Table 1: Performance Table

Dataset	Meter Values	KM	KM(5)	KM(10)	GAKM	ARAT	PIGAKM	PADGAKM
Dataset	Error	36.34	37.66	35.53	35.01	34.72	29.59	27.09
	Time	18.64	18.24	16.90	16.05	15.35	14.28	13.86
Iris	Error	44.16	45.08	42.35	40.78	37.54	35.38	33.74
	Time	16.38	15.98	14.76	12.84	11.56	10.26	09.61
Lymphoma	Error	35.03	36.76	33.85	31.92	30.06	29.43	27.06
	Time	12.08	11.26	10.82	11.13	11.78	10.84	09.27

Fig 4 & Fig 5 represents the graphical chart for the Table 1. Here the KM shows the k-means Clustering values , GAKM shows the data items which are group with the help of genetic algorithm and the K-Means clustering. ARAT shows average time rate and average error rate for the values given in KM. PIGAKM shows the result that has been done in the first process. Finally PADGAKM shows the results obtained in the proposed work.

CONCULSION

This experimental evaluation scheme was created to provide a correct base of performance and also a comparison with other methods. From these experiments on the datasets, it is observed that proposed approach using the parallel implementation of genetic algorithm has provided the correct results in the terms of finding the good clustering configuration. Interdependence information within the clusters and discriminative information for clustering. The proposed system is developed to produce dynamic parameters that have been developed to produce the correct results. The system is developed to produce the string parameter dynamically not an individually. The proposed approach is helpful in selecting significant centers, from each cluster. At last, the experimental results of PADGAKM are better than the simple genetic algorithm that has been already used. The average time and error rate are very less compared to other methods.

REFERENCES

[1]. J.Han and Michelin , "Data mining concepts and techniques," morgan Kauffman, 2006.

[2]. Jiawei Han M .K, Data Mining Concepts and Techniques , Morgan Kaufman publishers, An Imprint of Elsevier, 2006.

[3].K.Krishna, and M.Murty,:Genetic k-means Algorithm,"IEEE Transactions on System, Vol.29, No.3,1999.

[4]. Y.Lu,S.Lu,F.FOTOUHI,Y.Deng, and S.Brown, "FCKA: A fast genetic K-means clustering Algorithm,"ACM Symposium on Applied Computing,2004.

[5]. U.Maulik,and S.Bandyopadhyay,"Genetic Algorithm-Based Clusrtering Technique" Pattern Recognition 33, 1999.

[6]. L.Hall,B.Ozyurt,andJ.Bezdek,"Clustering With A Genetically Optimized approach "IEEE Transactions on Evolutionary computation",vol.3,No.2,1999.

[7].Siarry, P.,A.Petrowski and M.Bessaou,"A multiple population genetic algorithm aimed at multimodal optimization", Advances in Engineering Software33(2002). [8].Rongjun Li, and Xianying Chang,"A Modified Genetic Algorithm with Multiple Subpopulations And Dynamic Parameters Applied in CVAR model",IEEE Transactions on Intelligent agents, Web Technologies and Internet Commerce, 2006.

[9].L.Hall, B.Ozyurt, and J.Bezdek,"Clustering With A Genetically Optimized approach ,"IEEE Transactions on Evolutionary computations, Vol 3, No. 2, 1999.

[10].P.Bradley, and U.Fayyad,"Refining Initial Points for K-means Clustering," In Proceeding of 15th International Conference on Machine Learning, 1998.

[11].K.A Abdul Nazeer , M.P Sebastian "Improving the Accurancy and Effiency of the K-means Clustering Algorithm " WCE 2009, London.

[12].Harikrishna Narasimhan , Purushothaman mraj " Contribution- Based Clustering algorithm for Content Based Image retrieval", 5th International Conference on Industrial and Information Systems, 2010, India.

[13]. Agglomerative Fuzzy K-Means Clustering Algorithm with Selection of Number of Clusters.

[14]. A self-stabilizing k-clustering algorithm for weighted graphs " Journals of Parallel and Distributed Computing, volume 70 Issue 11, November, 2010".

[15].Techniques of Cluster Algorithms in Data Mining "Johannes Grabmeier University of Applied Sciences, Deggendorf, Edlmaierstr Deggendorf, Germany.

[16].Refinement of K-Means clustering using Genetic algorithm " Journal of Computer Application , Volume IV Issue 2, 2011.
[17].A Comprehensive overview of basic Clustering Algorithm, Glenn Fung , June 22, 2001.
[18]. Genetic Algorithm-Based Clustering Technique, "by Ujjwal Maulik , Sanghamitra Bandyopadhyay , Sanghamitra B.
[19]. Genetic Algorithm Based Clustering: "A Survey Emerging Trends in Engineering and Technology, 2008" First International Conference on 16-18 July 2008.
[20]. Using Genetic Algorithms in Clustering Problems Marco Painho and Fernando Bação
Higher Institute of Statistics and Information Management, New University of Lisbon, Travessa.
[21].Genetic K-means Clusstering Algorithm for mixed Numeric categorical Datasets "International Journal of Artifical Intelligence & Applications , Volume 1, No.2, April 2010.