Parallel Implementation of Genetic Algorithm using K-Means Clustering

Dr. A.V. Senthil Kumar
Director, Department of Computer Applications ,Hindusthan College of Arts & Science, Coimbatore, India
Email: avsenthilkumar@yahoo.com

S.Mythili
Research Scholar, Department of Computer Applications ,Hindusthan College of Arts & Science, Coimbatore, India
Email: sathyapriya_aru2007@gmail.com

---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. A new algorithm Parallel Implementation of Genetic Algorithm using K-Means Clustering (PIGAKM) is proposed to overcome the existing algorithm. PIGAKM 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 by PIGAKM. To speed up GA process, the evalution is done parallely not individually. To show the performance and efficiency of this algorithms, the comparative study of this algorithm has been done.

Keywords : Clustering, Genetic algorithm, K-means, Mutation , Parallel.

Date of Submission: January 31, 2012 Date of Acceptance: March 12, 2012

I.INTRODUCTION

Data mining is a non-trival process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. Data mining is an extraction of the hidden predictive information from large databases[1]. Data mining scours the databases for hidden patterns finding predictive information that experts may miss, as it goes beyond their expectation [2].

Clustering is one of the techniques used for grouping the data's. a clustering is a collection of data points that are similar to one another within the same cluster[5]. Clustering is a method of unsupervised classification, where data points are grouped into cluster based on their similarity. The main goal of their algorithm is to produce the efficiency and effectiveness for the proposed result[11]. k-means algorithm is effective in producing clusters for many practical application. The computational complexity of the original k-means algorithm is very high , especially for large data sets. This algorithm results in different types of clusters depending on the random choice of initial centroids. Several attempts were made by researches for improving the performance of k-means clustering algorithm[3].

Genetic algorithms are adaptive search algorithm based on the evolutionary ideas of genetics. GA is started with a set of solutions represented by chromosomes called population [8]. GA simulates the survival of the fittest among individuals over consecutive generation for solving a problem. Each generation consists of a population of character strings may be identified by a complete encoding of the DNA structure. This techniques generates new individuals by switching subsequences of the strings. This process appears while changing the character's randomly[7].

II.REVIEW LITERATURE:

2.1 Clustering

Clustering means grouping of similar data into one cluster or splitting a large set of data into smaller data sets. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. The goal of a clustering algorithm is to maximize the intra-cluster similarity and minimize the inter-cluster similarity [6].Clustering algorithms can be broadly classified into 5 types.

1.Partitional clustering 2. Hierarchical clustering 3.Density based clustering 4.Grid based clustering 5.Model based clustering[11,12].

The clustering algorithm finds the centroid of a group of data. To determine cluster membership, the algorithm evaluate the distance between a point and the cluster centroids[6]. Clustering is also called as data segmentation because in some application large data sets are grouped together. Clustering is also a process of partitioning a given set of objects into the disjoint clusters. Clustering is used to group set of objects into classes such that similar objects are placed in the same cluster while dissimilar objects are in separate cluster [5].

Fig.1 represents the string patterns are given ,then the similar strings are selected. The selected string patterns are interrelated and its is grouped. Finally the grouped string will be represented by clusters.

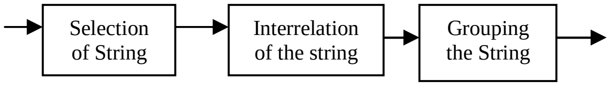

Fig 1 . Loop for the string representation.

The distance between two points or strings is taken as a common metric to access the similarity among the instances of the string. The most commonly used distance measures is the Euclidean metric which defines the distance between the two strings as P & Q.

Let P = (X1(P),X2(P)…) & Q=(X1(P,X2(Q)…). Consider d as the distance between the strings. This may be represented as

$$d(P.Q) = \sqrt{(x1(P)-x1(Q)^2+(x2(P)-x2(Q))^2+...}$$

$$= \sqrt{\left[\sum_{j=1}^{p}(xj(P)-xi(Q))^2\right]}$$

$$= \sqrt{(xi(P)-xi(Q)^2}$$

There are several methods used to calculate the distance between the two strings or cluster. The distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of other cluster.

Three different clustering methods are used for calculating the distance between clustering.

2.2 K-means clustering:

k-means is one of the most popular methods used to solve the clustering problems phases is to define k centroids, one for each cluster. The next phase is to take each points distance is generally considered to determine the distance between data points and the centorids. When all the points are included in some clusters, the first step is completed and an early grouping is done. At this point we need to recalculate the new centroids , as the inclusion of new points may lead to a change in the cluster centroids. Once we find k new centroids , a new binding is to be created between the same data points and the nearest new centroids, generating a loop. As a result of this loop, the k centroids may change their position in a step by step manner. Eventually , a situation will be reached where the centroids do not move anymore. This signifies the convergence criterion for cluistering [3, 10,11].

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.

2.3 Genetic algorithm

GA are examples of evolutionary computing methods and are optimization type algorithms. GA are inspired by Dawin's theory about evolution. Algorithm is started with a set of solutions represented by chromosomes called population[7].

In GA, a population of strings encodes the candidates solution to an optimization problem which envolves towards better solutions. Solutions are represented in binary as strings of 0's and 1's, so the output are generated randomly or individually. in this process multiple individually are selected and it is evaluated automatically[9].

2.3.1 Basic's of GA:

1.[START] Generates random items of each n chromosomes which is suitable solutions of the problem.

2.[FITNESS] Evaluates the efficiency of f(x) of each chromosomes x in the string.

3.[NEW POPULATION] Creates a new substring by repeating following steps until the new substring is created.

1.[SELECTION] Selecting thse two substrings from the given strings according to their nearest relationships of the given string. (the better fitness, the bigger chance to be selected).

2.[CROSS OVER] A cross over probability cross over the parents to form a new offspring. If no cross over was performed, off spring is an exact copy of parent.

3.[MUTATION] A mutation probability creates a new substring for the offspring presented in the chromosomes.

4.[REPLACE] New generated substring will be processed to run the algorithm.

5.[TEST] If the test condition is satisfied, STOP and return the best solution in current string.

6.[LOOP] continues if the condition is not satisfied. then the process will Go to step2.

GA is used for the purpose of finding the fixed number k of the cluster, where GA's is executed first to give initial values of k-means to start with rather than choosing random ones and expected to minimize the number of iterations that k-means needs in order to converge the local minima.

2.3.2 Principles of GA:

1. Encoding of the data as the binary string.
2. Randomly generating the substring for the given string. This one includes a genetic representation for the group of solutions.
3. Knowing the fitness of the substring value for each string. Then it will directly depend on the distance to the optimum value of the string.
4. Selection of the string from the substring will share the fitness value of the data.
5. Genomes crossover and mutations has been processed for each values.
6. And then start again from the point3.

Using GA's into clustering an initial population of random cluster is set. The k cluster centers encoded in each chromosomes are initialized to k randomly chosen points from the data set. This process is repeated from each of the P chromosomes in the population. At each generation, each individual is evaluated on the basics of its fitness. New individuals can be created using two main genetic operators are crossover and mutations. At the beginning of a run of a GA a large population of random chromosomes are created. Each one, when decoded will represent a different solution to the problem. Repeat the process until a new population of N members has been created.

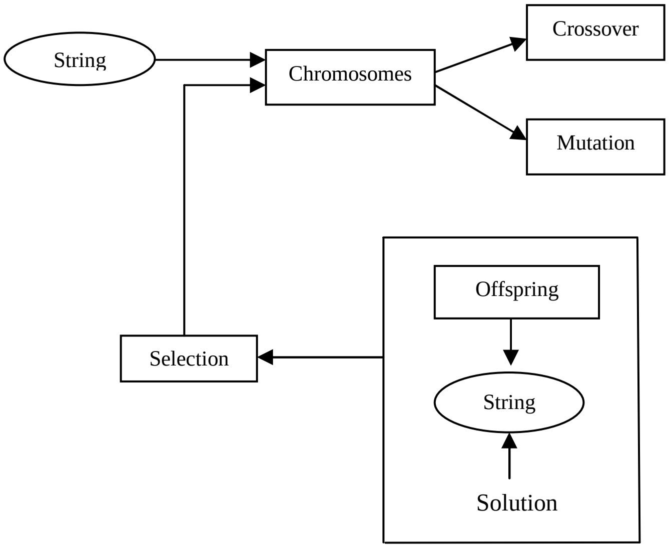

Fig 2. General structure of GA

Fig 2 represents the general structure of GA . the string patterns are given as the input, the chromosomes in the given string are compared , if the given strings are similar the string is selected and the solution is given.

III. PROPOSED WORK

This paper presents a Parallel Implementation of Genetic algorithm using k-means clustering algorithm (PIGAKM) which uses multiple subpopulations and dynamic parameters. Simple genetic algorithm involves only one initial population with fixed genetic 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. To solve the draw backs of simple genetic algorithm, a new technique is proposed in this paper called Parallel Implementation of Genetic algorithm using K-means clustering is developed using multiple substrings and dynamic parameters.

Parallel Implementation of Genetic algorithm:

In a simple GA, there is only one string in each generation and all the genetic operations are applied on it. It may require lot of functions to solve the problem within the time. The general way to speed up the GA processes is to evaluate the individual string parallelly not sequentially. It uses a client-server method where as one processor is the client which stores a single string and another processor is a server which evaluates the individual based on mutation and cross over operation.

The result of parallel GA is similar as the sequential GA. Here, the strings are split into separated substring to perform the diversity of the execution of the each substrings. The number of substrings is involved in parallel and an exchange of individuals through the set of substrings. Migration of individuals between different substring, followed by application of genetic operators, obtain generation of new individuals. The rate of migration allows the algorithm to control the level of diversity to be maintained inside the substring.

The genetic operational parameters, specifically crossover probability and mutation probability are designed in a dynamic way on the basis of selected adjustment functions. When the evolution process started the crossover probability and the mutation probability are used to search the solution space quickly and increase the diversity of the string. When the number of generations executed the values of the genetic operator will be reduced , to keep the evolution process stable the global optimum is finally converged.

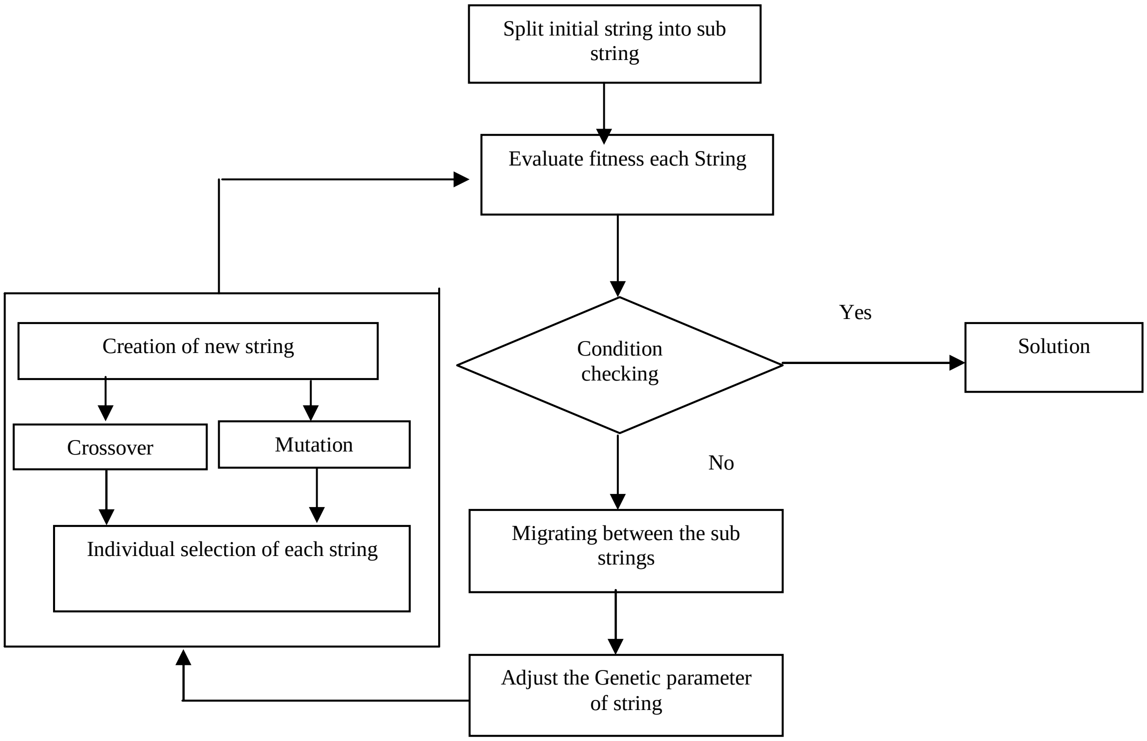
Split initial string into sub string
Evaluate fitness each String
Condition checking
Yes
Solution
No
Migrating between the sub strings
Adjust the Genetic parameter of string
Creation of new string
Crossover
Mutation
Individual selection of each string

Fig 3 . Structure of Parallel Implementation of Genetic Algorithm

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. The length of the code should be determined by

$$L \quad = \quad 1n\left(\frac{x_{\max} - x_{\min}}{\sigma} + 1\right) / \ln 2$$

Where Xmax and Xmin denote the maximum value and the minimum value of the variables respectively, and is the precision required.

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 :

$$fitness(i) = \frac{2.(p_i - 1)}{Q_i - 1}$$

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, denoted by C and M respectively. Before the execution of the process , there is no guidance used to determine the values of M and N. In a general form , the values of M and N are bigger , the quality of the solution may be higher, but the process of the evolution may be longer. The recommended values of PIGAKM are (L+n), where n is the number of variables.

The probability of selecting the i individual depends on the rate si, which is proportional to its degree of fitness, that is

$$S_i \;=\; fitness(i) / \sum fitness(i)$$

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.

$$m_j = a_j . \sin\left(\frac{\pi}{2} \bullet \frac{(T-t)}{T}\right)$$

Step 4: Create initial string randomly.

$$c_j = b_j.\sin\left(\frac{\pi}{2}\bullet\frac{(T_j - t)}{T_j}\right)$$

Here c represents the creation of 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 perform including selection, crossover and mutation, on each substring and create new generations.

Step 9: Process will complete when the termination condition is satisfied.

IV. RESULTS

The results 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.

Dataset1 : Dataset 1 is consists of 2 scattered points and a 4 specific points in the radius are (0.125, 0.25) , (0.625, 0.25), (0.375 , 0.75) , (0.875, 0.75). This datasets are grouped into 4 clusters.

Iris data : The different categories of Iris have 4 feature values, which represents the length and width in centimeters.

Lymphoma data : The lymphoma data sets are grouped into 3 clusters. 1. Diffuse Large B-Cell Lymphoma (DLBCL) 2. Follicular Lymphoma (FL) 3. B-Cell Lymphocytic Leukemia (B-CLL).

Dataset	**Meter Values**	KM	KM(5)	KM(8)	ARAT	PIGAKM
Dataset	Error	36.34	37.66	36.53	35.46	29.59
	Time	18.64	18.24	18.08	17.42	14.28
Iris	Error	44.16	45.08	44.85	38.63	35.38
	Time	16.38	15.98	15.07	12.70	10.26
Lymphoma	Error	35.03	36.76	35.27	30.52	29.43
	Time	12.08	11.26	11.02	12.49	10.84

Table 1. Performance Table

Table1 shows the results that has been obtained using this three datasets. In the table KM shows the K-means Clustering , KM(5) shows the 5 iterations of the datasets , KM(8) shows the 8 iterations of the datasets , the average rate and average time is represented as ARAT , the result obtained in this process is represented in PIGAKM.

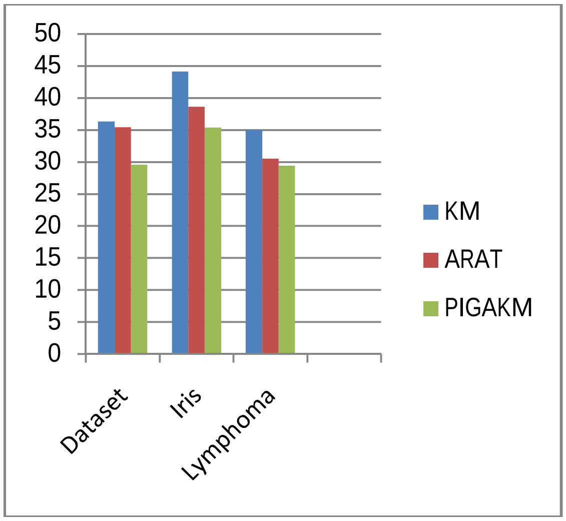

Fig 4 . Average Error Parameter

Fig 4 & Fig 5 represents the graphical chart for the Table 1. Here the KM shows the k-means Clustering values , ARAT shows average time rate and average error rate for the values given in KM. finally PIGAKM shows the result of the proposed work.

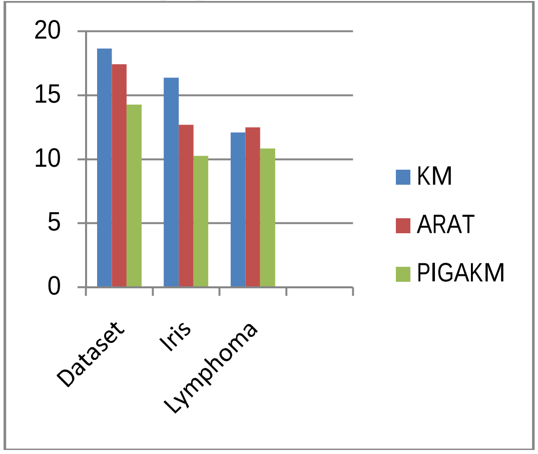

Fig 5. Average Time Parameter

V.CONCLUSION:

This experimental evaluation scheme was created to provide a correct base of performance and also a comparison with other methods. From this 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. This contains interdependence information within the clusters and discriminative information for clustering. The proposed approach is helpful in selecting significant centers, from each clusters. At last , the experimental results of PIGAKM is 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, and J.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.