International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 07 | July -2017 www.irjet.net p-ISSN: 2395-0072

# Mining Big Data using Genetic Algorithm

**Surbhi Jain**

*Assistant Professor, Department of Computer Science, India*

---

**Abstract –** *In today's era, the amount of data available in the world is growing at a very rapid pace day by day because of the use of internet, smart phones, social networks, etc. This collection of large and complex data sets is referred to as Big Data. Primitive database systems are unable to capture, store and analyse this large amount of data. It is necessary to improve the text processing so that the information or the relevant knowledge which was previously unknown can be mined from the text. This paper proposes need for an algorithm for the clustering problem of big data using a combination of the genetic algorithm with some of the known clustering algorithms. The main idea behind this is to combine the advantages of Genetic algorithms and clustering to process large amount of data. Genetic Algorithm is an algorithm which is used to optimize the results. This paper gives an overview of concepts like data mining, genetic algorithms and big data.*

***Key Words***: **Genetic Algorithms, Big Data, Clustering, Chromosomes, Mining**

## 1. INTRODUCTION

In current Big Data age the data is becoming more and more available owing to advances in information and communication knowhow, enterprises are gaining meaningful information, relevant knowledge and vision from this huge data based on decision making. Big data mining is the ability of taking out valuable information from huge and complex set of data or data streams i.e. Big Data. One of the important data mining techniques for big data analysis is clustering. There are difficulties for applying clustering techniques to big data due to enormous amount of data rising on daily basis. There are a lot of clustering techniques available the most common of which is the K-means algorithm. It is used to analyze information from a dataset. But as we are saying that because of big data we have plethora of data available, thus available clustering algorithms are not very efficient. As Big Data refers to terabytes and petabytes of data, we need to have clustering algorithms with high computational costs. We can think of designing an algorithm which can combine the features of some of the clustering algorithms and genetic algorithm to process big data.

To extract some meaningful information from the source data is the process called Mining. It is a set of computerized techniques that are used to extract formerly unknown or buried information from large sets of databases. A Successful Data Mining makes possible to uncover patterns and relationships, and then to use this "new" information for making proactive knowledge-driven business decisions. There are a lot of algorithms which are being used for mining the information from plain text. The algorithms used to solve the optimization problems are the Genetic Algorithms. These algorithms work on search based inputs. The algorithms eventually leads to generate useful solutions for such kind of problems.

## 2. GENETIC ALGORITHMS

Genetic Algorithms are a clan of computational prototypes inspired by evolution theory of Darwin. According to Darwin the species which is fittest and can adapt to changing surroundings can survive; the remaining tends to die away. Darwin also stated that "the survival of an organism can be maintained through the process of reproduction, crossover and mutation". GA's basic working mechanism is as follows: the algorithm is started with a set of solutions (represented by chromosomes) called population. Solutions from one population are taken and used to form a new population (reproduction). This is driven by optimism, that the new population will be superior to the old one. This is the reason they are often termed as optimistic search algorithms. The reproductive prospects are distributed in such a way that those chromosomes which represent a better solution to the target problem are given more chances to reproduce than those which represent inferior solutions.

They search through a huge combination of parameters to find the best match. For example, they can search through different combinations of materials and designs to find the perfect combination of both which could result in a stronger, lighter and overall, better final product.

As an example we can consider "Face Recognition Systems" which are used for drawing sketches based on visualizations. This system is majorly used for investigation purposes where in sketch of some criminal is to be made on the basis of description given by some eye witness to the crime. The initial population is nothing but a lot of facial features which are already there in the system. The features may include a lots of varieties of noses, ears, lips, eyes etc. They may differ in color, size or anything else. As the witness starts giving descriptions the features which are most likely to match can be selected (Selection). The selected features can then follow the steps of cross-over and mutation to produce more likely features. As in eyes of one face and lips of another can be chosen to go for cross over to produce a new individual which has both the features matching with the criminal. The process continues till the witness recognizes the final face as the one desired.

## 3. BIG DATA

Big data is a term for data sets that are so large or complex that primitive data processing application software is inadequate to deal with them. Big data represents a new period in data study and utilization. It is a leveraging open source technology- a robust, secure, highly available, enterprise-class Big Data platform. Challenges include capture, storage, analysis, querying, and updating data safely and securely. While the term "big data" is relatively new, the doing of collecting and storing plethora of information for eventual analysis is ages old.

The significance of big data is not based on how much data we have, but how we use that data. We can take data from any source and analyze it to find responses that enable us to produce results in reduced cost and time with smart decision making. Here in this paper we are trying to combine big data with genetic algorithms for generating efficient analysis of data. The reason for the interest in genetic algorithms is that these are very powerful and broadly applicable search techniques. As said earlier also, Big Data refers to large-volume, complex, growing data sets with numerous, self-directed sources. Big Data are now rapidly expanding in all fields like science and engineering, including physical, biological and biomedical sciences with the fast development of networking, data storage, and the data collection capacity.

With the new technology of Big Data, the computations can be speeded up. In very usual cases, if our system starts getting heavy because of loads of data which is becoming too big for our system to be managed, we add RAM or vacate some space by deleting certain processes. Big data on the contrary, adds more systems to the pool and there by promote parallelism. This however leads to fault tolerance as a consequence. More the number of systems, more is the probability of system failures. Fortunately, big data handles this automatically by duplicating data on the systems so that if one system fails, its data can be redirected to some other system.

## 4. DATA MINING

The knowledge from the data sets is extracted using Data Mining technology. It is used to search and analyze data. The data to be mined varies from a small data set to an enormous sized data set i.e. big data. In Data Mining, the source data is kept in the format of databases i.e. in the form of tables if we are considering relational databases. We only have to apply the algorithms to extract data from databases. The Data Mining environment produces voluminous data. The information retrieved in the data Mining step is transformed into the structure that is easily understood by users. Once data has been extracted and then transformed, it is loaded into systems from where we can read it. The various methods like genetic algorithms, support vector machines, decision tree, neural network and cluster analysis to disclose the hidden patterns inside the huge amounts of data set are all included in data mining.

For handling such large amount of data sets, various algorithms which define various structures and approaches implemented to handle Big Data are needed. They also defines the various tools that were developed for analyzing them. Data mining and Text Mining are often used synonymously which however is not right. Although both are mining techniques, but there is a very thin line of difference between the two. Data mining refers to the process of extraction of useful text from the databases which is not known prior, while text mining refers to extraction of useful and knowledgeable data from the plain text i.e. the naturally occurring text. Unlike data mining, this text need not be transformed into any other format.

## 5. CLUSTERING

Clustering refers to categorizing similar kind of objects. It is a method of exploring the data, a technique of finding out patterns in the dataset. It falls in the category of unsupervised learning i.e. we don't know in advance how data should group the data objects (of similar types) together. It is one of the most vital research field in the data mining. In clustering we aim at making collections of objects in such a manner that the objects having same attributes

belong to same group and objects with different behaviors in dissimilar groups. With the formation of groups, we can easily identify areas where the object space is dense and where it is sparsely filled and hence can determine the distribution patterns. We can find the stimulating patterns directly from the data sets without needing to have much of background knowledge. One of the popular approaches of clustering is Partitioning. Partitioning works by transferring objects by moving them from one cluster to another cluster starting from a certain point. The number of clusters for this technique should be pre-defined for this technique (like in k-means algorithm).

## 6. GENETIC ALGORITHM FOR CLUSTERING

The voluminous data that is available to us can be divided into small groups where each group can be considered as population. By applying genetic operators iteratively on the population we can find out the optimum solution for the current scenario. Search process, as we all know, is a problem-solving method wherein we cannot determine the sequence of steps leading to the solution in advance. It is based on how nicely and wisely we have applied the search operators. An ideal search should be capable of carrying out search process locally as well as in a random manner. Random search explores the entire solution and is proficient in avoiding reaching to a local optimum while local search helps in exploring all the local possibilities and reaching the best solution.

As discussed earlier a genetic algorithm is capable of effectively searching the problem domain and solving complex problems by simulating natural evolution. It perform search and provide near optimal solutions for objective function of an optimization problem. A set of chromosomes is referred to as a population wherein a chromosome (represented as strings) refers to the parameters in the search space, encoded by a combination of cluster centroids.

First step is to create a random population, which represents different solutions in the search space. Next, a few of chromosomes are selected as per the principle of survival of the fittest, and each is assigned into the next generation. Chromosomes are nothing but binary encoded strings, which represents probable solutions to the optimization problem. Each string is then evaluated on the fitness function (objective function), giving a measure of the solution quality called the fitness value. A new candidate solution population can be created after recombination (crossover and mutation) is being performed upon candidate solution selection. Individual representation and population initialization, fitness computation, selection, crossover and mutation are thus the basic steps of genetic algorithm for data clustering. Given is the algorithm for the same:

**Input:**

k: the no of clusters
d: the data set containing n objects
p: population size Tmax: Maximum no. of iterations

**Output:**

A set of K clusters
1) Initialize every chromosome to have k random centroids selected from the set of data.
2) For T=1 to Tmax
    (i) For every chromosome i
        a. Allocate the object data to the cluster with the closest centroid.
        b. Recomputed k cluster centroids of chromosome i as the mean of their data objects.
        c. Compute the chromosome i fitness.
    (ii) Generate the new group of chromosomes using GA selection, crossover and mutation.

The spine for a Genetic Algorithm to work is the Fitness function F (x). The prime focus of this function is to give the successive results after applying GA.

Firstly, it is derived from the objective function and then used in successive genetic operations like crossover, mutation. Fitness means quality value which is the degree of the reproductive efficiency of individual string (chromosomes). A score is given to each individual chromosome with the help of fitness functions. The proposal is to generate a Genetic Algorithm based clustering algorithm which is expected to provide an optimal clustering, better than that of K-Means approach. This may however induces a little more time complexity.
The major benefit of using genetic algorithms is that they are easily parallelized. Parallel implementation of GA is apprehended using two commonly used models namely:

- Coarse-grained parallel GA
- Fine-grained parallel GA

In the first model every node is given a population split to process while in the second model each individual is

provided with a separate node for fitness evaluation. Adjoining nodes communicate with each other for selection and remaining operations.

### 6.1 PARALLEL Implementation for Clustering using GAs

At first, the input data set is fragmented according to the block size by the input format. Each fragment is then given to a mapper to perform the First phase clustering, the results of which are passed on to a single reducer to perform the Second phase mapper.

**Step 1: Population initialization**

Each mapper forms the initial population of individuals after receiving the input fragments. Each individual is a chromosome of size *N*. Every segment of the chromosome is a centroid. Centroids are randomly selected data points from the received data split. For every data point in each chromosome clustering is performed and the data set is assigned to the cluster of the closest centroid. Then the fitness is evaluated.

**Step 2: Mating & Selection**

Cross-over and mutation techniques are used for mating. For cross-over, we generally use arithmetic cross-over which generates one offspring from two parents. The centroid of the offspring is the arithmetic average of the corresponding centroid of parents. Swap mutation technique is used for mutation. In this, 9's compliment of the data points is taken. The offspring from older population are selected to produce a new population. For selection, an approach known as Tournament selection is used wherein the individual is selected by performing a tournament based on fitness evaluation among several individuals chosen at random from the population.

**Step 3: Termination**

A new population thus generated replaces the older population which would again form a newer population using mating and selection procedure. This whole procedure would be reiterated again and again until the termination condition is met. The termination condition can be anything like achieving a specified number of iterations or reaching a particular solution. The fittest individual of the final population of each mapper is handed on as the result to the reducer. The Second phase clustering on the mapping results of all mapper is then performed by the reducer.

### 6.2 GENETIC K-MEANS ALGORITHM

Apart from parallel implementation using Genetic Algorithms, we can also have an algorithm that combines the advantage of Genetic algorithm and K-means algorithm for clustering. It is expected to provide an optimal clustering, better to that of K-Means approach, but probably with a little more time complexity.
The major steps of the algorithm of GK-means are:

1) Set the population.
2) Compute fitness of every individual by following equation.
   Fitness (i) =2. (pi - 1)/Q-1
   i=individual, p=position, Q=total individuals
3) If satisfied with the fitness condition, then assign solution, Else
4) Calculate sub population and migrate
5) Counting the $i^{th}$ individual depends on the rate si, which is relative to its level of fitness that is
   Si = fitness (i) / summation (fitness (i));
6) Translate population and assets individual wellness.
7) Perform crossover and mutation on each sub population
8) If termination condition satisfies, stop; else go to step 5.

The major drawback of k-means algorithm is that it can't process large amounts of data. If we have minimum amount of data then k mean is easy to process but for large amount of data it will not give desired results. Since we are talking about Big Data, so surely k-means is not the solution to our problem. GK-means on the contrary will take less memory and time to process big data and will give desired results as well. The Genetic k-means gradually converges to the global optimum as desired.

## 7. DISADVANTAGES OF GA

A major difficulty in applying Genetic Algorithms is how to handle constraints. Genetic operators often produce infeasible offspring while manipulating chromosomes. A Penalty technique is used to keep a check on the number of infeasible solutions produced in each generation. This helps in enforcing the genetic search towards an optimal solution. Apart from this, a few other disadvantages are:

1) These are challenging to understand and to describe to end users.

2) The problem abstraction and the means to represent individuals is quite difficult.
3) How to determine the best fitness function is a difficult work.
4) How to do crossover and mutation is another difficulty.
5) The large over-production of individuals and the random character of the search process is another drawback.

## 8. FUTURE SCOPE

The paper compares and reviews the methods available for clustering data based on genetic algorithms. A more robust and time saving algorithm can be designed such that big data can be effectively mined overcoming all the challenges being faced by Genetic Algorithms.

## 9. CONCLUSION

This paper provides the reader a review of all the jargons related to analysing big data. The concepts like Text Mining, Big Data and Genetic Algorithm concept, samples, scope, methods, advantages, challenges etc. are all discussed here. The paper reviews various methods that are available for text mining. The paper concludes that since the prime focus is on to mining big data, so algorithm followed has to be space effective and time effective. The paper presents need for an algorithm that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective

## 10. REFERENCES

[1] Senthilnath, J., S. N. Omkar, and V. Mani. "Clustering using firefly algorithm: performance study." Swarm and Evolutionary Computation 1, no. 3 (2011).
[2] Ahmed and Saeed. A Survey of Big Data Cloud Computing Security. *International Journal of Computer Science and Software Engineering (IJCSSE)*, Volume 3, Issue 1, December 2014.
[3] Arora, Deepali, Varshney, Analysis of K-Means and K-Medoids Algorithm For Big Data, *International Conference on Information Security & Privacy* (ICISP2015), 2015.
[4] Dash and Dash, 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, 2012.
[5] Gaddam, Securing your Big Data Environment, Black Hat USA 2015.
[6] http://www.sas.com/en_us/insights/big-data/internet-of-things.html
[7] Inukollu, Arsi and Ravuri, Security Issues Associated With Big Data in Cloud Computing. *International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014.*
[8] Jiawei Han and Micheline Kamber, "Data Mining Concepts & Techniques", Second Edition, *Morgan Kaufmann* Publishers
[9] "Text Mining Technique using Genetic Algorithm", International Journal of Computer Applications (0975 – 8887) Volume #. 63, February 2013
[10] McAfee, Andrew, and Erik Brynjolfsson. "Big data: the management revolution." Harvard business review 2012
[11] Deepankar Bharadwaj, Dr. Arvind Shukla, Text Mining Technique on Big data using Genetic Algorithm, International Journal of Computer Engineering and Applications, Volume X, Issue IX, Sep. 16
[12] Mitsuo Gen, Runwei Cheng, Genetic Algorithms and Engineering Optimization, John Wiley and Sons, 2000