ISSN 2278-3083

Volume 4, No.2, March - April 2015

International Journal of Science and Applied Information Technology

Available Online at http://www.warse.org/ijsait/static/pdf/file/ijsait01422015.pdf

# Incremental Data Clustering using a Genetic Algorithmic Approach

**Amit Anand[1], Tejan Agarwal[2], Rabishankar Khanra[3], Debabrata Datta[4]**

[1]Department of Computer Science, St. Xavier's College, Kolkata, India, technamrit_amit@yahoo.com
[2]Department of Computer Science, St. Xavier's College, Kolkata, India, tejanagarwal@gmail.com
[3]Department of Computer Science, St. Xavier's College, Kolkata, India, khanrarabisankar@gmail.com
[4]Department of Computer Science, St. Xavier's College, Kolkata, India, debabrata.datta@sxccal.edu

## ASBTRACT

Data clustering can be considered as a guided classification of patterns into groups, popularly known by the term clusters. The problem of clustering is represented by different analyst, researchers and scientists in much different form. These representations reflect that clustering is one of the most important stages in the field of data or information analysis, especially when dealing with data in warehouses for mining. Till date a lot of clustering techniques have been introduced in the market. However, in this paper we have tried to discuss here a new kind of clustering method based on Genetic Algorithms.

**Key words**: Big Data, Data Warehouse, Incremental Clustering, Genetic Algorithm, KDD

## 1. INTRODUCTION

The world is dealing with millions of transaction and other online activities happening across the globe those ultimate results in huge amount of data flow from end to the other. These data are typically known by term 'Big Data' or 'Historical Data'. With these historical data, organizations can extract even that information which they really never collect directly. This can be achieved by the use of data mining. Data mining refers to the process of identifying valid, novel, useful and understandable relations and patterns in the existing data. This identification of useful insights is often referred to as data discovery, data archaeology, information harvesting etc. The term “data mining” is mostly popular among the statisticians, database researchers, and business organizations which use this technique for their benefit. The term Knowledge Data Discovery (KDD) is generally used to refer to the overall process of discovering useful knowledge from data, where data mining is a particular step in this process. Data mining process analyses enormous data and transformed them into such a form which is easily understandable by the user [1]. These historical data are stored in large storage units called Data warehouses. Organizations dealing with historical data have recognized the importance of the knowledge hidden in their large data warehouses.

When speaking of a data warehousing environment, analysts and knowledge workers work on two characteristics namely, analysis and multiple updates. These characteristics or requirements tend to new approaches in the field of Data Mining or knowledge data discovery (KDD). Data mining has been defined as the application of data analysis and discovery algorithms that - under acceptable computational efficiency limitations - produce a particular enumeration of patterns over the data. Data mining includes a number of steps such as clustering, classification and summarization. Our area of concentration is clustering. In data warehouse, data is not updated immediately when insertions and deletions on the operational databases occur. Updates are collected and applied to the data warehouse periodically in a batch mode, e.g., each night [2]. Due to the very large size of the databases, it is unfeasible to cluster entire data for every updates, since it requires ample amount of time to process data that were stored previously with the one which are recently added. Hence, it is highly desirable to perform these updates incrementally.

In this paper, the main discussion is on a new way of clustering which is based on the principles of genetic algorithms. This algorithm ART_INC clusters the data dynamically as and when updates are applied to the warehouse. In other words it can be said that using this algorithm the warehouse needs to be clustered just once and when new data elements are identified, they are added to their respective clusters dynamically. Thus, it saves time by the denying requirements of clustering the data from scratch each time when the warehouse is updated.

## 2. GENETIC ALGORITHM

The concept of Genetic Algorithms was first proposed by John Holland in the year of 1970. In general we deal with a number of search techniques and apply it to a variety of custom problem in order to achieve optimized result. Similarly, GAs are also heuristic based search techniques that rely on the biological principal of "Natural Selection", which in its simplest form allows the stronger individuals to take over the weaker ones. This phenomenon is more commonly known by the phrase "Survival Of The Fittest". GAs also simulates this phrase over consecutive generation that consists of population of character strings that are analogous to the chromosomes in human DNA. The first generation of the strings is randomly generated and a series of operations are then performed on them to generate more successive generations. The initial generation is known as the parents and subsequent populations that are derived from them are known as child.

The use of Genetic algorithms for problem solving is not new. Genetic algorithms have been successfully applied in the field of optimization technique, machine learning etc [3]. . The standard GA applies genetic operators such as selection, crossover and mutation on a randomly generated population for the computation of the whole generation of new string. These operations are applied iteratively until two consecutive generations generated have the same chromosomes. The probability of the new chromosomes generated depends on their fitness for the problem calculated by the fitness function and so the quality of the new chromosomes enhances in successive generations [4]. GAs combine the good information hidden in a solution with good information from another solution to produce new solutions with good information inherited from both parents, hopefully leading towards optimality. The ability of genetic algorithms to explore and exploit a growing amount of theoretical justification, and successful application to real-world problems strengthens the conclusion that GAs are a powerful, robust optimization technique. However, getting an optimal or exact solution to a problem is very difficult but the researches have shown that GAs represent an intelligent approach of solving heuristic based problem and can lead to a fairly good solution which may not be optimal but much better in comparison to the results obtained using the primitive techniques.

## 3. BACKGROUND OF THE WORK

Clustering is a fundamental form of data analysis that is applied in a wide variety of domains, from astronomy to zoology. With the radical increase in the amount of data collected in recent years, the use of clustering has expanded even further, to applications such as personalization and targeted advertising. Clustering is now a core component of interactive systems that collect information on millions of users on a daily basis [5]. The ultimate aim of the clustering is to provide a grouping of similar records. Clustering is a technique in which, the information that is logically similar is physically stored together. In order to increase the efficiency of search and the retrieval in database management, the number of disk accesses is to be minimized [3]. In clustering, since the objects of similar properties are placed in one class of objects, a single access to the disk can retrieve the entire class. However, with the never ending data in today's era it is becoming impractical to store all relevant information in memory at the same time, often necessitating the transition to incremental methods called Incremental Clustering.

Incremental Clustering basically incorporates the database activities that help in adding the recently updated or newly added data elements to the appropriate clusters that were obtained as a result of clustering on the data in the warehouse earlier than the most recent update to the database had taken place. In this way it requires less time and has improved efficiency in comparison the traditional clustering methods since it denies the claims of clustering the data again from the scratch.

In clustering similar data sets are identified with the help of distance between the two clusters. This distance consists of all or some elements of the two clusters. It is taken as a common metric to analyze the similarity between among the component of a population. In this paper, the most frequently used distance measure metric called Euclidean distance has been used. The Euclidean distance defines the distance between two points, viz., $p = (p_1, p_2, \dots)$ and $q = (q_1, q_2, \dots)$ as follows –

$d = [\Sigma(p_i - q_i)^2]^{1/2}$ , where 'i' is the range of the points to be considered [7].

In this work, the k-means algorithm has been used as a clustering method. The k-means algorithm is a simple iterative method to partition a given dataset into a user specified number of clusters, k. The basic steps of the k-means algorithms are as follows –

Step 1: To select 'k' points as the initial centroids.
Step 2: To assign all the points to the closest centroid.
Step 3: To re-compute the centroid of each cluster.
Step 4: To repeat steps 2 and 3 until the centroids don't change.

## 4. APPLYING GENETIC ALGORITHM TO THE CLUSTERING ALGORITHM

The standard k-means algorithm is sensitive to the initial centroids and poor initial cluster centres as a result of which the cluster generated may not be the best fitted ones for the solution to the problem. In order to overcome the sensitivity problem the concept of the Genetic Algorithm (GA) has been used in this work to the traditional k-means algorithm for the centre point selection that is locally optimal [8]. This work has designed an algorithm called ART_INC that have been tested with some experimental data to achieve the target.

This algorithm basically, at first, applies the traditional k-means algorithm to the given data set to divide the data set into k number of clusters. After successfully deriving k clusters ($k_1$, $k_2$, $k_3$, ..., $k_n$), it computes the mean value for the respective clusters as ($m_1$, $m_2$, $m_3$, …, $m_n$). After this, the incremental clustering task is performed. This is achieved by taking new elements and keeping the old clusters as it is. These new elements in the clusters are put by comparing the Euclidean distance from the mean of each cluster. The element is added to the cluster for which the distance is the minimum. The process is repeated until two consecutive clusters are the same. These clusters are then passed to the GA sub-routine that generates a uniform random number ‘r’ and then examines each cluster with respect to the value of ‘r’ to select the best fit cluster among the k clusters. Here a gene represents a cluster centre of ‘n’ attributes and a chromosome of ‘k’ genes represents a set of ‘k’ cluster centres. The pseudocode representing the algorithm ART_INC may be depicted as follows –

**Algorithm**:- **ART_INC(n,d[n],p)**

```
//n is the number of elements
//d[n] is the data set
//p is the number of clusters
Step 1 – Initialize the cluster array k[ ][ ] to -1
Step 2 – For all the elements in the array repeat steps 3 and 4
Step 3 – Calculate the Euclidean distance of every element from the mean
Step 4 – Store the elements in the appropriate cluster
Step 5 – Calculate the mean of every cluster and store it in an array (say m[ ])
Step 6 – Repeat steps 2 to 6 until two consecutive clusters are same
Step 7 – Calculate the total of the means stored in m[ ]
Step 8 – Find a random number in the GA module
Step 9 – Multiply this random number with the total of the mean
Step 10 – Find the cluster whose mean is immediately greater than the value obtained in Step 9
Step 11 – The cluster thus obtained is the fittest cluster
Step 12 – Take more input elements
Step 13 – Find the distance from the mean of each cluster
Step 14 – Enter the element to the cluster for which the distance is the minimum
Step 15 – Repeat steps 13 and 14 until all the elements are put into a cluster
Step 16 – Calculate the total of the means stored in m[ ]
Step 17 – Repeat steps 8 to 12 for the clusters obtained in the previous step.
Step 18 – End
```

Table 1, as given below, shows how the algorithm works when GA is applied no normal k-means clustering. The corresponding algorithm, known as ART, was discussed and illustrated in [1].

**Table 1: Applying GA to k-means clustering**

| **No.of coordinates** | **Data set** | **No.of clusters** | **Clusters** | **Mean Values** | **Fittest Cluster** |
|---|---|---|---|---|---|
| 15 | (10,3)<br>(68,97)<br>(55,55)<br>(21,33)<br>(6,0)<br>(1,68)<br>(74,0)<br>(2,4)<br>(38,10)<br>(11,65)<br>(32,87)<br>(45,10)<br>(0,29)<br>(71,82)<br>(44,31) | 3 | K1 :<br>(10,3)<br>(6,0)<br>(2,4)<br>(38,10)<br>(45,10)<br>(0,29)<br><br>K2 :<br>(68,97)<br>(32,87)<br>(71,82)<br><br>K3 :<br>(55,55)<br>(21,33)<br>(1,68)<br>(74,0)<br>(11,65)<br>(44,31) | m1 =<br>(16.83,<br>9.33)<br><br>m2 =<br>(57.0,8<br>8.67)<br><br>m3 =<br>(34.33,<br>42.0) | K2 |

The next table i.e., table 2 shows how incremental clustering is used to modify and improve upon the previous algorithm as discussed in [1]. At first, 10 elements are taken and they are put into 3 clusters. Now without changing the clusters obtained, 5 more elements are put into the clusters. From the outputs as shown in the next table, it can be observed that the fittest cluster may change after adding extra elements to the clusters already obtained.

**Table 2: Applying GA to incremental k-means algorithm**

| No.of coordinates | Data set | No. of clusters | Clusters | Mean Values | Fittest Cluster |
|---|---|---|---|---|---|
| 15 | (10,3)<br>(68,97)<br>(55,55)<br>(21,33)<br>(6,0)<br>(1,68)<br>(74,0)<br>(2,4)<br>(38,10)<br>(11,65) | 3 | K1:<br>(10,3)<br>(6,0)<br>(74,0)<br>(2,4)<br>(38,10)<br>K2:<br>(68,97)<br>K3:<br>(55,55)<br>(21,33)<br>(1,68)<br>(11,65) | m1 = (26.0,3.4)<br>m2 = (68,97)<br>m3 = (22,55.25) | K2 |
| 5 | (32,87)<br>(45,10)<br>(0,29)<br>(71,82)<br>(44,31) | | K1:<br>(10,3)<br>(6,0)<br>(74,0)<br>(2,4)<br>(38,10)(45,10<br>K2:<br>(68,97)(71,82)<br>K3:<br>(55,55)(21,33)(1,68)<br>(11,65)(32,87)(0,29)<br>(44,31) | m1 = (29.17,4.5)<br>m2 = (69.5,89.5)<br>m3 = (23.43,52.57) | K3 |

## 5. CONCLUSION AND FUTURE WORK

ART_INC has successfully dealt with the problem domains of the traditional k-means algorithm and has been able to eradicate it as much as possible. In addition to this, the proposed algorithm has been successful in performing data clustering in an incremental approach. A major advantage of this algorithm ART_INC is that any number of clusters can be created according to the requirements. This was not possible in the previous version of the algorithm, known as ART. Further testing on various databases is in progress to test the robustness of the present algorithm is in progress.

## REFERENCES

[1] Amit Anand, Tejan Agarwal, Rabishankar Khanra, Debabrata Datta, "Data Clustering using a Genetic Algorithmic Approach", IJATCSE, Vol. 3, No. 4, July – August, 2014, ISSN: 2278-3091.

[2] Martin Ester ,Hans-Peter Kriegel, Jörg Sander, Michael Wimmer, Xiaowei Xu, Incremental Clustering for Mining in a Data Warehousing Environment, Proc. at 24th International Conference on VLDB.

[3] I. K. Ravichandra Rao, "Data Mining and Clustering Techniques", DRTC Workshop on Semantic Web, 8-10 December, 2003.

[4] K. F. Man, K. S. Tang, and S. Kwong, "Genetic Algorithms: Concepts And Applications", IEEE Trans. on Industrial Electronics, Vol. 43, No. 5, pp. 519-534, Oct 1996.

[5] Margareta Ackerman , Sanjoy Dasgupta, Incremental Clustering: The Case for Extra Clusters, Proc. at Advances in Neural Information Processing Systems, 2014

[6] Sangeeta Rani, Geeta Sikka, "Recent Techniques of Clustering of Time Series Data: A Survey", International Journal of Computer Applications Volume 52 - Number 15, August, 2012.

[7] Tian Zhang, Raghu Ramakrishnan, Miron Livny, "BIRCH:An Efficient Data Clustering Method for Very Large Databases", Proc. at ACM SIGMOD, 1996.

[8] Singh, R.V., Bhatia, M.P.S., "Data clustering with modified K-means algorithm", Proc. at International Conference on Recent Trends in Information Technology, 2011.