ISSN 2278-3091

Volume 3, No.4, July - August 2014

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://warse.org/pdfs/2014/ijatcse02342014.pdf

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

ABSTRACT

With the exponential growth of the internet the world is dealing with a variety of online activities which ultimately result in a flow of millions of bytes of data across the globe of which the users are hardly aware. To promote this phenomenon, esteemed organizations often use the technique of data mining to identify the user requirements more precisely. Traditional data mining algorithms are no more much efficient in the present era as they take ample processing time and space. To overcome these complexities and to identify the user determinants including organizational readiness and user characteristics this work presents a genetic algorithm based data mining technique. The use of this proposed data mining technique to the historical data that were collected and stored in a warehouse through a survey or by any such mean results to be useful for discovering meaningful and valuable data about the required entity.

Key words: Big Data, Data Mining, Genetic Algorithm, Clustering Technique

1. INTRODUCTION

The term "Big-data" generally means a collection of large amount of unstructured data stored in a data repository. These data are in general so enormous in amount that it is difficult to store, manage and analyze them. "Big-data" are the collection of data from heterogeneous data source. The "Big-data" architecture basically consists of three segments namely, the storage segment, the processing segment and the analysis segment and this architecture follows a distributed approach which is different from other traditional data storage models [1].

In the present era with the ever increasing internet technology, organizations gather vast and never ending amount of data daily in the ordinary course of their business. Much of this information is collected for day to day operational reasons but now many organizations have realized that these data have much more additional values. Such data are referred to as 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 Discovery in Databases (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 [2].

The construction of a data warehouse, which involves data cleaning and data integration, can be viewed as an important pre-processing step for data mining. However, a data warehouse is not a requirement for data mining. Building a large data warehouse that consolidates data from multiple sources, resolves data integrity problems, and loads the data into a database, can be an enormous task, sometimes taking years and costing millions of dollars [Gray and Watson, 1998a]. If a data warehouse is not available, the data to be mined can be extracted from one or more operational or transactional databases, or data marts. Alternatively, the data mining database could be a logical or a physical subset of a data warehouse. Data mining uses the data warehouse as the source of information for knowledge data discovery (KDD) systems through an amalgam of artificial intelligence and statistics-related techniques to find associations, sequences, classifications, clusters, and forecasts [Gray and Watson,1998b]. The data that are to be entered are then "cleaned" and moved into the warehouse. The data continue to reside in the warehouse until they reach an age where one of three actions is taken: the data are purged; the data, together with other information, are summarized; or the data are archived. An aging process inside the warehouse moves current data into old detail data [3]. The idea of Genetic algorithms, originally proposed by John Holland in 1970, are basically heuristic search algorithms that are based on the principle of biological process "Natural selection" in which the stronger individuals are likely to take over the weaker ones.[4] Genetic Algorithm simulates the survival of the fittest among individuals over consecutive generations for solving a problem. Each generation consists of a population of character strings that are similar to the chromosome that we see in our DNA. An initial generation is randomly generated, and several operations take place on chromosomes of that generation to produce new generations. It represents an intelligent approach of random search in a search space to solve a problem.

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. [5]. A fitness function associated with every string measures the fitness of the new chromosomes for the problem. 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 [3]. 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.

2. Background Of The Work

A cluster is an ordered list of objects, which have some common objects. The objects belong to an interval. As described in [5], clustering is a method in which we make cluster of objects that are somehow similar in characteristics. 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 [5]. 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. Clustering is ultimately a process of reducing a mountain of data to manageable piles and data discovery becomes easier.

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 we have used the most frequently used distance measure metric called Euclidean distance. The Euclidean distance defines the distance between two points, viz., $p = (p_1, p_2, \ldots)$ and $q = (q_1, q_2, \ldots)$ as follows –

$d = [\ \Sigma(p_i - q_i)^2]^{1/2}$, where **'i'** is the range of the points to be considered [9].

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.

3. Applying Genetic Algorithm To The k-means 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 are not the best fitted ones for the solution to the problem. Research work has been done to use the basic concepts of this algorithm for any modified applications [7]. 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. This work has designed an algorithm called ART 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). 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 may be depicted as follows –

```
Algorithm:- ART(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 – End.
```

A small glimpse of the data samples used for the experimental purpose using this algorithm is tabulated in Table 1. Table 1 shows that if case 1 is considered then it is found that the value of the random number (r) generated using Roulette's Wheel algorithm is 26.809368193176642, which is very close to the mean value of the cluster k_3 which is 65.0. As a result the algorithm selects the cluster k_3 as the fittest cluster. The algorithm repeats the same phenomenon for case 2 and case 3 as well. Also, it is seen that even if the algorithm is made to run using a set of data for different number of times (tested here for 15 times using a same data set) then also the algorithm works in the above fashion i.e. it selects the cluster (as the fittest one) whose mean value is closest to the value of r (generated in Roulette's wheel algorithm). Therefore, from the above analysis it is quite apparent that the selection of the fittest cluster using this Genetic Algorithm is solely dependent on the value of r and is thus very much probabilistic in nature.

Table 1: Experimental Results

Case #	No. of elements	Data set	No. of clusters	Clusters	Mean Values	Fittest Cluster
1	10	{10, 22, 45, -9 65, -7, 98, -14, 2 52}	3	K_1 : { -9, -7, -14, 2 }	m_1 = -7.0	K_2
				K_2 : { 10, 22 }	m_2 = 16.0	
				K_3 : { 45, 65, 98, 52 }	m_3 = 65.0	
2	10	{65, 2, -7, -98 332, 10, -54, 11, 3, -24}	3	K_1 : { 332 }	m_1 = 332.0	K_1
				K_2 : { 65, 2, -7, 10, 11, 3 }	m_2 = 14.0	
				K_3 : { -98, -54, -24 }	m_3 = -58.66	
3	5	{63, -74, 20, 11, 2}	2	K_1 : { 63, 20, 11, 2 }	m_1=24.0	K_1
				K_2 : { -74 }	m_2=-74.0	

4. CONCLUSION AND FUTURE WORK

Thus using ART we have successfully dealt with the problem domains of the traditional k-means algorithm and are able to eradicate it as much as possible by employing ART. Also, it is now possible to select the fittest cluster using Genetic Algorithm (GA) technique which results in a much better solution in contrary to what we would have obtained by the implementation of traditional clustering techniques. A point to be noted that the algorithm

ART has now been tested for 4 clusters of data sets and attempt is being made to make the algorithm much more efficient so that it can produce N number of clusters as desired by the user. Further testing on various databases is in progress to test the robustness of our algorithm is in progress. Splitting continuous attribute into multiple intervals rather than just two intervals based on a single threshold is also considered to improve the performance. It is shown that our approach outperformed other approaches on both prediction accuracy and the standard deviation.

REFERENCES

[1] Chanchal Yadav, Shuliang Wang , Manoj Kumar, “Algorithm and approaches to handle large Data – A Survey”, International Journal of Computer Science and Network, Vol 2, Issue 3, 2013.

[2] Joyce Jackson, “Data Mining: A Conceptual Overview”, Proc. at Communications of The Association for Information Systems, Vol 8, 2002.

[3] 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.

[4] Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar, Partha Pratim Sarkar, “Mining Frequent Itemsets Using Genetic Algorithm”, International Journal of Artificial Intelligence & Applications, Vol. 1, No. 4, October 2010.

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

[6] Longbing Cao, Yong Feng, Jiang Zhong, “Advanced Data Mining and Applications”, Proc. at Second Intl. Conf, ADMA, 2006.

[7] 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.

[8] 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.

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