International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 10 3161 – 3166

Efficient K-Mean Clustering Algorithm for Large Datasets using Data Mining Standard Score Normalization

Sudesh Kumar
Computer science and engineering
BRCM CET Bahal Bhiwani (Haryana)
Bhiwani, India
ksudesh@brcm.edu

Nancy
Computer science and engineering
BRCM CET Bahal Bhiwani (Haryana)
Bhiwani, India
Nancypubreja09@gmail.com

***Abstract*—**In this paper, the clustering and data mining techniques has been introduced. The data mining is useful for extract the useful information from the large database/dataset. For extract the information with efficient factor, the data mining Normalization techniques can be used. These techniques are Min-Max, Z-Scaling and decimal Scaling normalization. Mining of data becomes essential thing for easy searching of data with normalization. This paper has been proposed the efficient K-Mean Clustering algorithm which generates the cluster in less time. Cluster Analysis seeks to identify homogeneous groups of objects based on the values of their attribute. The Z-Score normalization technique has been used with Clustering concept. The number of large records dataset has been generated and has been considered for analyze the results. The existing algorithm has been analyzed by WEKA Tool and proposed algorithm has been implemented in C#.net. The results have been analyzed by generating the timing comparison graphs and proposed works shows the efficiency in terms of time and calculation

Keywords- *Normalization, Data Mining, Clustering, Modified K-Mean, Centroids*

*****

I. INTRODUCTION

Data mining technology is used to give the user an ability to extract meaningful patterns from large database.After data mining implementation, Data Analysis need to be used for analyze the results for identify mining efficiency. Data analysis (DA) is an efficient method of analyzing large sets of data in a variety of fields, for internal, external and forensic audits. Most DA engagements involve working on existing data extracted by the IT departments of the audit client. Preparing the data for analysis can be a time-intensive task. Data mining (DM) is defined as the process of automatically searching large volumes of data for patterns such as association rules. It is a generic term used to explain a variety of tasks involving the analysis of data.Data Analysis is an analytical and problem-solving process that identifies and interprets relationships among variables. It is used primarily to analyze data based on predefined relationships, while DM, as it pertains to computer science, is used to identify new relationships in an otherwise bland dataset. More often than not, DA is considered as the knowledge to operate one of the DA tools e.g., Microsoft Excel, etc. Like auditing, DA needs a specific mindset as opposed to having merely the capability to use a given tool. It requires an analytical and problem-solving mindset with the ability to identify and interpret the relationships among the variables. Successfully solving a DA problem requires a deep understanding of the definition and application of various elements of DA. For analyze the results properly, the data transformation is core concept of data mining.

Preprocessing: The measurement unit used can affect the data analysis. For example, changing measurement units from meters to inches for height, or from kilograms to pounds for weight, may lead to very different results. In general, expressing an attribute in smaller units will lead to a larger range for that attribute, and thus tend to give such an attribute greater effect or "weight."

To help avoid dependence on the choice of measurement units, the data should be normalized or standardized. This involves transforming the data to fall within a smaller or common range such as [-1, 1] or [0.0,1.0] Normalization is particularly useful for classification algorithms involving neural networks or distance measurements such as nearest-neighbor classification and clustering. If using the neural network back propagation algorithm for classification mining, normalizing the input values for each attribute measured in the training tuples will help speed up the learning phase. For distance-based methods, normalization helps prevent attributes with initially large ranges (e.g., income) from outweighing attributes with initially smaller ranges (e.g., binary attributes). It is also useful when given no prior knowledge of the data [1]. There are many methods for data normalization. The number of Normalization techniques exist in data mining elaborated as:

a. Min-max normalization

This performs a linear transformation on the original data. Suppose that min_A and max_A are the minimum and maximum values of an attribute, A. Min-max normalization maps a value, v_i, of A to v'_i in the range [new min_A,new max_A] by computing the formula.

$$v'_i = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Min-max normalization preserves the relationships among the original data values. It will encounter an out-of-bounds error if a future input case for normalization falls outside of the original data range for A. For Example, minimum and maximum values for the attribute income are $12,000 and $98,000, respectively. It would like to map income to the range [0.0, 1.0]. By min-max normalization, a value of $73,600 for income is transformed to (73,600-12,000/98,000-12,000)*(1.0-0) +0=0.716.

b. Standard Score Normalization(Z-Score)

In z-score normalization (or zero-mean normalization), the values for an attribute, A, are normalized based on the mean (i.e., average) and standard deviation of A. A value, v_i, of A is normalized to v'_i by computing:

$$v'_i = \frac{v_i - \bar{A}}{\sigma_A}$$

Where σ_A and $\bar{A}$ are the mean and the standard deviation respectively of attribute A

c. Decimal Scaling

Normalization by decimal scaling normalizes by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of A. A value, v_i, of A is normalized to v'_i by computing the formula Where j is the smallest integer such that max $|v'_i| < 1$.

$$v'_i = \frac{v_i}{10^j}$$

Clustering: Clustering is the process of partitioning a set of data into a set of meaningful sub-classes, called clusters. It helps users to understand the natural grouping or structure in a dataset. Clustering is the most important concept for generate the groups and it is an unsupervised-learning problem. The main purpose is finding the structure in the collection of unlabeled data. Totally, the clustering involves partitioning a given dataset into some groups of data whose members are similar in some way. The usability of cluster analysis has been used widely in data recovery, text and web mining, pattern recognition, image segmentation and software reverse engineering.

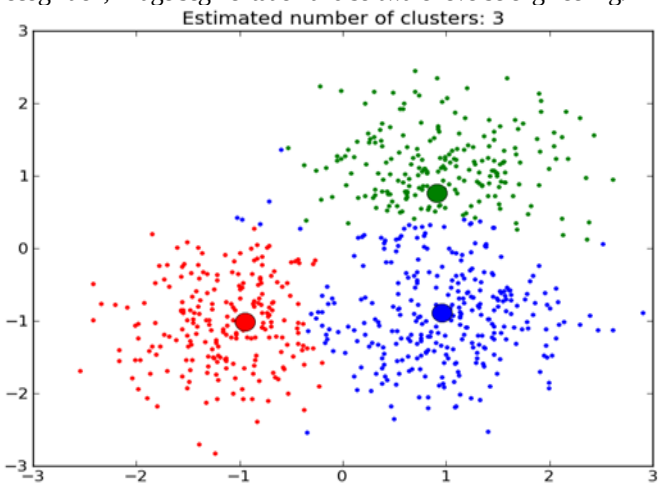

Figure 1: Clustering Concept

K-mean algorithm: K-Means Clustering algorithm is an idea, in which there is need to classify the given data set into K clusters; the value of K (Number of clusters) is defined by the user which is fixed. In this first the centroids of each cluster is selected for clustering and then according to the chosen centroids, the data points having minimum distance from the given cluster, is assigned to that particular cluster. Euclidean Distance is used for calculating the distance of data point from the particular centroids.

II. LITERATURE REVIEW

Author explained an approach for personalizing web directories with the aid of web usage data given by author Pierrako et al. (2010) [2]. Here, the Web directory is viewed as a thematic hierarchy and personalization is realized by constructing user community models on the basis of usage data. In contrast to most of the work on Web usage mining, the usage data that are analyzed here correspond to user navigation throughout the web, rather than a particular web site, exhibiting as a result a high degree of thematic diversity. Following this methodology, it enhances the clustering and probabilistic approaches presented in previous work and also presents a new algorithm that combines these two approaches. The experiments also assess the effectiveness of the different machine learning techniques on the task.

Author explained the concept of web usage data clustering using Dbscan algorithm and set similarities given by author Santhisree& Damodaram (2010) [3]. Here a new Rough set Dbscan clustering algorithm which identifies the behavior of the user's page visit, order of occurrence of visits. Web data Clusters are formed using the rough set Similarity Upper Approximations. It also presents the experimental results on MSNBC web navigation dataset, and proved that Rough set Dbscan clustering has better efficiency and performance clustering in web usage mining is finding the groups which share common interests compared to Rough set agglomerative clustering.

Author explained the concept of clustering based URL normalization technique for web mining given by Nagwani et al. (2010) [4]. URL normalization is an important activity in web mining. URL normalization also reduces lot of calculations in web mining activities. A web mining technique for URL normalization is proposed in this paper. The proposed technique is based on content, structure and semantic similarity and web page redirection and forwarding similarity of the given set of URLs. Web page redirection and forward graphs can be used to measure the similarities between the URL's and can also be used for URL clusters. The URL clusters can be used for URL normalization.

A data structure is also suggested to store the forward and redirect URL information and author explained the URL (Uniform Resource Locator) normalization is an important activity in web mining given by Nagwani, N.K.(2010) [5]. Web data can be retrieved in smoother way using effective URL normalization technique. URL normalization also reduces lot of calculations in web mining activities. A web mining technique for URL normalization is proposed in this paper. The proposed technique is based on content, structure and semantic similarity and web page redirection and forwarding similarity of the given set of URLs. Web page redirection and forward graphs can be used to measure the similarities between the URL's and can also be used for URL clusters. The URL clusters can be used for URL normalization. A data structure is also suggested to store the forward and redirect URL information.

Author has been assimilated the Knowledge about cluster analysis with an emphasis on the challenge of clustering high dimensional data given by Er. Arpit Gupta,Er.Ankit Gupta, 2011[6]. The principal challenge in extending cluster analysis to high dimensional data is to overcome the "curse of dimensionality," and they described the way in which high dimensional data is different from low dimensional data, and how these differences might affect the process of cluster analysis and also described several recent approaches to clustering high dimensional data, including our own work on concept-based clustering. All of these approaches have been successfully applied in a number of areas, although there is a need for more extensive study to compare these different techniques and better understand their strengths and limitations.

Author explained web usage mining for predicting user's browsing behaviors by using Fuzzy Possibilistic Clustering Mean (FPCM) algorithm given by Punithavalli et al. (2011) [7]. World Wide Web is a huge storehouse of web pages and links. It offers large quantity of data for the Internet users. User's accesses are recorded in weblogs. Web usage mining is a kind of mining techniques in logs. Because of the remarkable usage, the log files are growing at a faster rate and the size is becoming very large. This leads to the difficulty for mining the usage log according to the needs. The hierarchical agglomerative clustering to cluster user's browsing patterns is used. Here it enhances the two levels of Prediction Model to achieve higher

hit ratio. It uses Fuzzy Possibilistic algorithm for clustering. The experimental result shows that the proposed techniques results in better hit ratio than the existing techniques.

Cluster analysis divides data into groups (clusters) for the purposes of summarization or improved understanding. For example, cluster analysis has been used to group related documents for browsing, to find genes and proteins that have similar functionality, or as a means of data compression and author explained the concept of K-Means clustering, its advantages and disadvantages given by the author Sovan Kumar Patnaik, Soumya Sahoo(2012) [8]. Author has also explained the biggest advantage of the k-means algorithm in data mining applications and its efficiency in clustering large data sets. However, its use is limited to numeric values. Due to filtering capacity of K-mean, this algorithm is only used in case of numeric data sets. The Agglomerative and Divisive Hierarchical Clustering algorithm was adopted the dataset of categorical nature initially. Due to complexity in both of the above algorithm, this paper has presented a new approach to assign rank value to each categorical attribute for K-mean Clustering. The categorical data have been converted into numeric by assigning rank value. It is a categorical dataset can be made clustering as numeric datasets. It is observed that implementation of this logic, k- mean yield same performance as used in numeric datasets.

The author has proposed an efficient, modified K-mean clustering algorithm to cluster large data-sets whose objective is to find out the cluster centers which are very close to the final solution for each iterative steps is given by author Anwiti Jain, Anand Rajava,t Rupali Bhartiya[9]. Clustering is often done as a prelude to some other form of data mining or modeling. Performance of iterative clustering algorithms depends highly on the choice of cluster centers in each step. The algorithm in this paper is based on the optimization formulation of the problem and a novel iterative method. The cluster centers computed using this methodology are found to be very close to the desired cluster centers. The experimental results using the proposed algorithm with a group of randomly constructed data sets are very promising. The best algorithm in each category was found out based on their performance.

III. PROPOSED METHODOLOGY

The existing Clustering algorithm has been analyzed through WEKA Tool and result has been generated. The duplicate, irrelevant information need to be cleaned in relational database. The Prime Objectives are:

1. To Study of existing data analysis clustering technique.
2. To Analyze complexity and outlier issue in Algorithm.
3. To Find point of complexity in Algorithm.
4. To Develop the Clustering Algorithm for large and small datasets.
5. To Study the proposed algorithm and its advantage.
6. To Implement the algorithm and perform analysis.
7. Generate Results [15].

Flow Chart:

A flowchart is a type of diagram that represents an algorithm, workflow or process, showing the steps as boxes of various kinds, and their order by connecting them with arrows. This diagrammatic representation illustrates a solution model to a given problem. Flowcharts are used in analyzing, designing, documenting or managing a process or program in various fields.

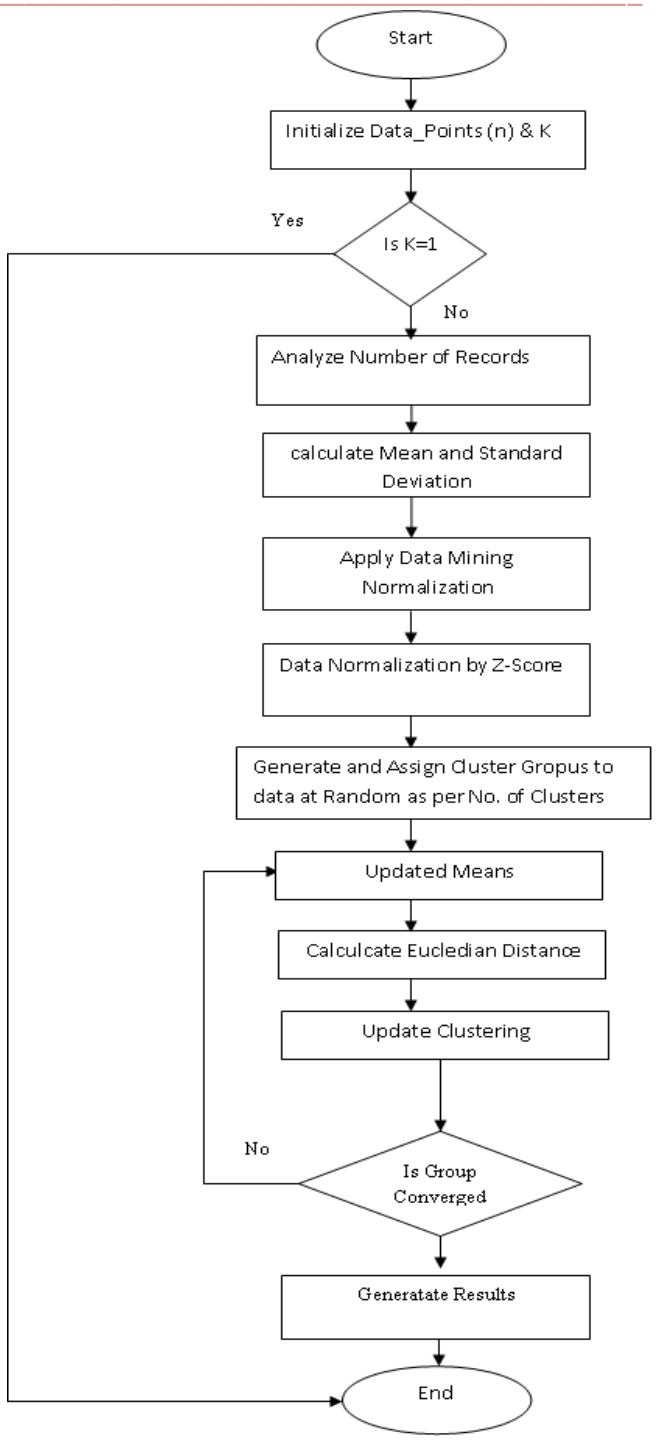

Figure 2: K-Mean Proposed Flow Chart

The result has been analyzed by WEKA tool with backend relational database. WEKA Explorer is an application that provides the functionality of Dataset Management, loading data, feeding them to classifiers, filters, storing the results of classification, apportioning data between training and testing subsets [15].

IV. ALGORITHM

The algorithm described below presents the design of the simulator.

1. Initialize the data points (n) and Number of Clusters (K)

ISSN: 2321-8169
3161 – 3166

2. Checkpoint Cluster Value (K)
3. If number K=1, then Exit
   Else
4. Calculate Std and Mean of (Data_Points) by the Equation

$$std(E) = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(e_i - \bar{E})^2}$$

5. Convert the Data_Points as Per the Z-Score Normalization
6. Generate the Cluster Group (Cg) at Random
7. Assign the Cg to the Data_Points.
8. If Cg Converged, move to Step 13
   Else
   Move to Step 9
9. Update the means
10. Update the Euclidean Distance
11. Update Cg Cluster Groups
12. Move to Step 8
13. Generate Results
14. End

*Where Std indicates Standard Deviation

V. RESULT

In this, C# program has been executed and different dataset has been taken as input. After this, the proposed algorithm will take no. of clusters as input. As the computation is complete, clustering result will shown in window. This clustering result also contain the total computation time. With the variation in clusters and number of records as well, the time calculated and graphs has been drawn. The 20K and 30 K records have been considered. These records has been analyze in WEKA tool as well as Proposed Algorithm in C#.net platform.

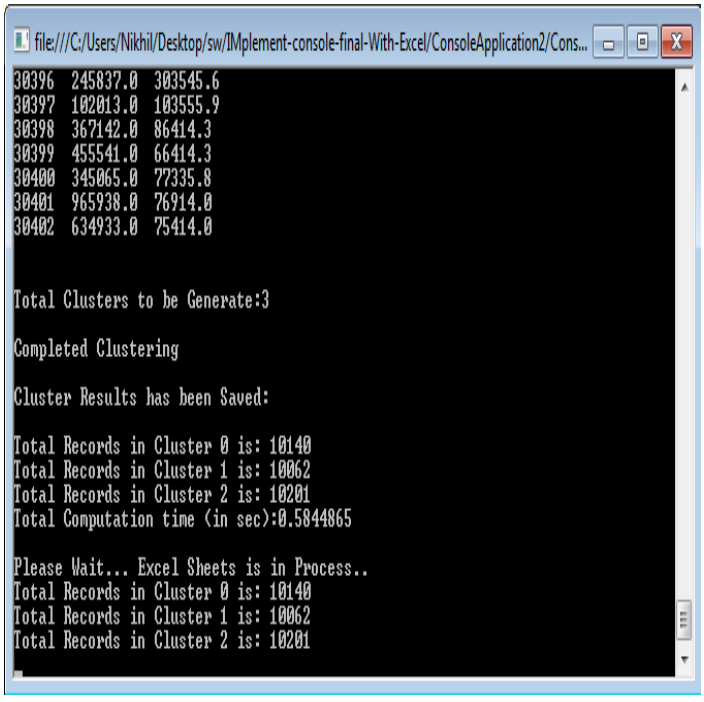

Figure 3: Timing for 3 Clusters for 30K records in C#

.

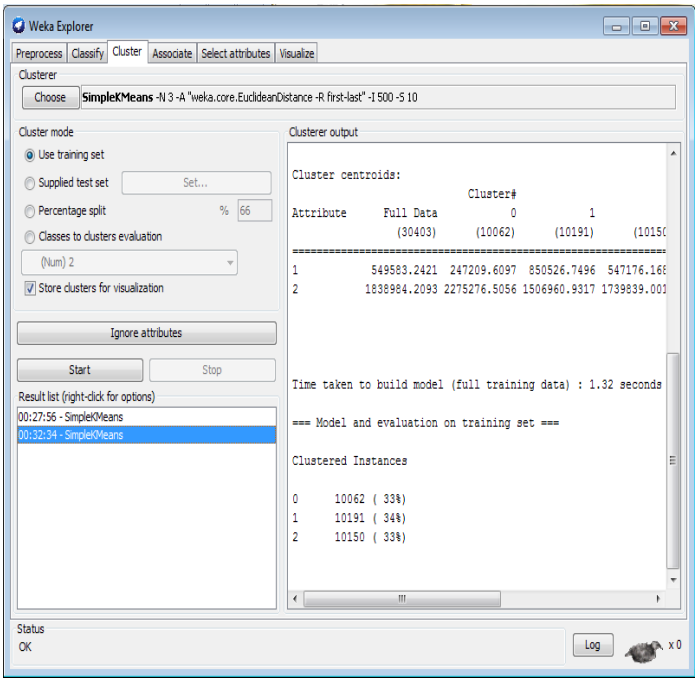

Figure 4: Timing for 3 Clusters for 30K records in WEKA

The below windows show the results for 30K records in WEKA and C# as well. The proposed algorithm has been taken the less computation time for generate the clusters. The below graph are the timing graphs for WEKA and K-Means algorithm and shows the comparison

These results show that the complexity has been reduced in the K-Mean algorithm using the data mining techniques

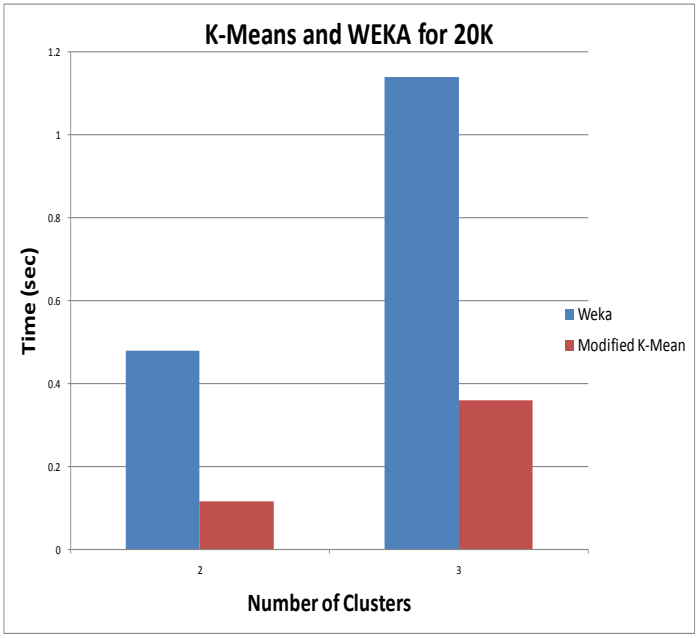

Figure 5: Timing Comparison for 20K records

ISSN: 2321-8169
3161 – 3166

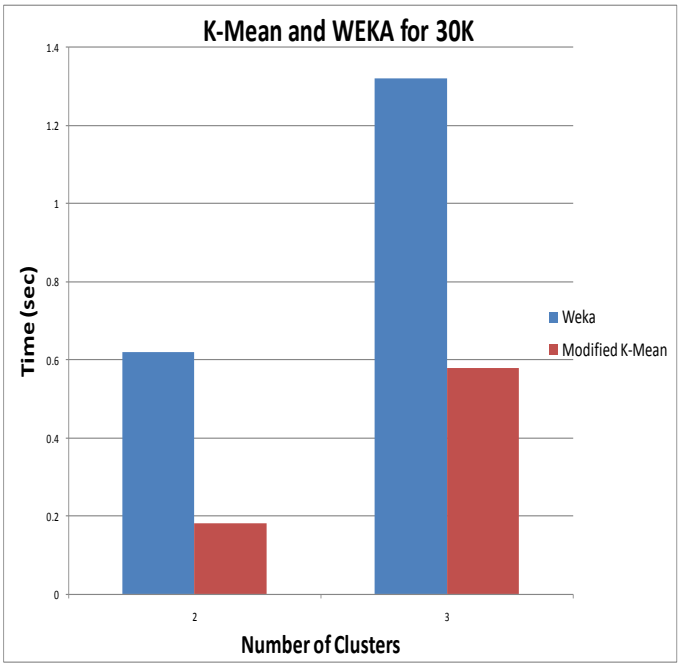

Figure 6: Timing for 30K records

Clusters Generated:

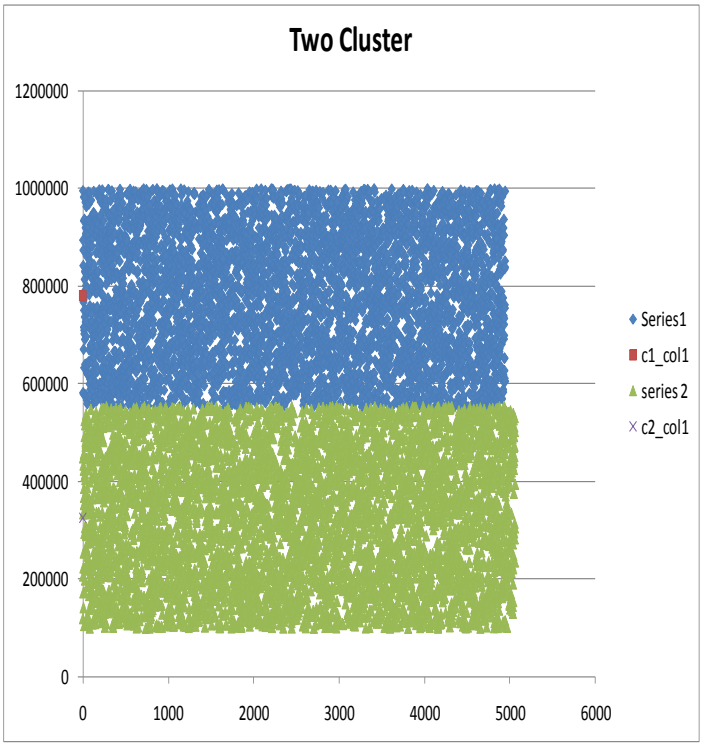

Figure 7: Clusters Output by Proposed Algorithm

Number of Records		20000
Number of Clusters	WEKA(sec)	Proposed Algorithm(sec)
2	0.48	0.11
3	1.14	0.36
Number of Records		30000
2	0.62	0.18
3	1.32	0.58

Table 1: Comparison Table of K- Mean Clustering Algorithm and Our Proposed Algorithm

VI. CONCLUSION AND FUTURE WORK

The clustering involves partitioning a given dataset into some groups of data whose members are similar in some way. The usability of cluster analysis has been used widely in data recovery, text and web mining, pattern recognition, image segmentation and software reverse engineering. K- Mean clustering algorithm is suffered by the problem of time complexity. The Z-Score Normalization technique has been implemented with K-Means Algorithm on large dataset with variation in number of records. The timing has been compared with WEKA Tool to that of our k-mean algorithm proposed in C#. The timing has been improved and calculation problem has been resolved.

The proposed partitioning based algorithm assumes Euclidean distance measure as the only measure to produce clusters. Some measure other than distance may be used to produce clusters. These measures can be city block distance, Minkowski metric, Cosine-correlation etc. So, work can be done in this field also.

REFERENCES

[1] Luai Al Shalabi, Zyad Shaaban and Basel Kasasbeh (2006), "Data Mining: A Preprocessing Engine", Journal of Computer Science 2 (9): 735-739.

[2] Dimitrios Pierrakos, GeorgiosPaliouras: "Personalizing Web Directories with the Aid of Web Usage Data, IEEE Transactions on Knowledge and Data Engineering", Vol. 22, Page No 9, September 2010.

[3] K. Santhisree, A. Damodaram et al.: "Web Usage Data Clustering Using Dbscan Algorithm and Set Similarities", pp.220-224, 2010 International Conference on Data Storage and Data Engineering, 2010.

[4] Naresh Kumar Nagwani : "Clustering Based URL Normalization Technique for Web Mining", ace, pp.349-351, 2010 International Conference on Advances in Computer Engineering, 2010.

[5] Nagwani, N.K.:"Clustering Based URL Normalization Technique for Web Mining, Performance analysis of MK-means clustering algorithm with normalization approach", Advances in Computer Engineering (ACE) (2010)

[6] Er. Arpit Gupta, Er.Ankit Gupta: "Research Paper on Cluster Techniques of Data Variations", IJATER, 2011.

[7] R.Khanchana and M. Punithavalli:" Web Usage Mining for Predicting Users' Browsing Behaviors by using FPCM Clustering", International Journal of Engineering and Technology vol. 3, no. 5, pp. 491-496, 2011.

[8] Sovan Kumar Patnaik, Soumya Sahoo: "Clustering of Categorical Data by Assigning Rank through Statistical Approach", International Journal of Computer Applications 2012.

[9] Anwiti Jain, Anand Rajava,t Rupali Bhartiya(2012) "Design Analysis and Implementation of Modified KMean Algorithm for Large Data-set to Increase Scalability and Efficiency ",IEEE

[10] Minky Jindal and NishaKharb, "K-means Clustering Technique on Search Engine Dataset using Data Mining Tool", International Journal of Information and Computation Technology. ISSN 0974-2239 volume 3, Number 6 (2013) , pp. 505-510 © International Research Publications House http://www. irphouse.com /ijict.htm

[11] Dr. Sudhir B. Jagtap and Dr. Kodge B. G., "Census Data Mining and Data Analysis using WEKA", (ICETSTM – 2013) International Conference in Emerging Trends in Science, Technology and Management-2013, Singapore.

[12] Madihah Mohd Saudi and Zul Hilmi Abdullah, "An Efficient Framework to Build Up MalwareDataset", International Journal of Computer, Information, Systems and Control Engineering, Vol: 7 No: 8, 2013.

[13] Rui Wang; Chiu, K:"Optimizing Distributed RDF Triplestores via a Locally Indexed Graph Partitioning, Parallel Processing (ICPP)", 2012 41st International Conference on, on page(s): 259 – 268

[14] G. Singh and N. Kaur, "Hybrid Clustering Algorithm with Modified Enhanced K-Mean and Hierarchical Clustering", International Journal of Advanced Research in Computer Science and Software Engineering 2013.

[15] Sudesh Kumar, Nancy (2014)."K-Mean Evaluation in Weka Tool and Modifying It using Standard Score Method", International Journal on Recent and Innovation Trends in Computing and Communication.