iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** VI **Month of publication:** June 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 3 Issue VI, June 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

An Empirical Analysis and Designing of Data Mining Clustering Algorithm

Sukhman Jot Kaur[1], Er.Tarun Bagga[2]
[1]Research scholar, [2]Lecturer, Department of Computer science and engineering, HEC Jagadhri, Haryana, India

Abstract- Data mining is the technique of running the data through the algorithms for discovering the meaningful correlations and patterns among the data that would remain hidden otherwise. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).
Keywords-Data mining, Clustering, Clustering algorithms, Weka tool

I. INTRODUCTION

Data mining is the technique of running the data through the algorithms for discovering the meaningful correlations and patterns among the data that would remain hidden otherwise. The operations of data mining occur in the background. Users of data mining can just see the results. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

Data mining is related to the subarea of statistics called exploratory data analysis, which has similar goals and relies on statistical measures. It is also closely related to the subareas of artificial intelligence called knowledge discovery and machine learning. The important distinguishing characteristic of data mining is that the volume of data is very large, although ideas from these related areas of study are applicable to data mining problems, scalability with respect to data size is an important new criterion. An algorithm is scalable if the running time grows (linearly) in proportion to the dataset size, given the available system resources (e.g., amount of main memory and disk). Old algorithms must be adapted or new algorithms must be developed to ensure scalability. Finding useful trends in datasets is a rather loose definition of data mining. In a certain sense, all database queries can be thought of as doing just this. Indeed, we have a continuum of analysis and exploration tools with SQL queries at one end, OLAP queries in the middle, and data mining techniques at the other end. SQL queries are constructed using relational algebra (with some extensions) OLAP provides higher level querying idioms based on the multidimensional data model and data mining provides the most abstract analysis operations. Data mining tasks can be as complex `queries' specified at a high level, with a few parameters that are user-definable, and for which specialized algorithms are implemented.

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

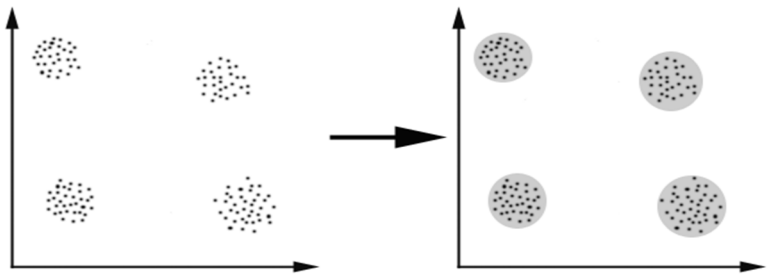


Fig: - Graphical Example of Clustering

II. LITERATURE REVIEW

To carry on this review work a variety of papers has been explored. Major findings out of these review papers are explained in the upcoming paragraphs.

Osama Abu Abbas (2008)[6] explained Comparisons between data clustering algorithms. Clustering is division of data into groups of similar objects. Each group called cluster consists of objects that are similar amongst themselves and dissimilar compared to objects of other groups. There are various clustering algorithms used for clustering of data.[6] concluded that as no. of cluster, k become greater; the performance of SOM becomes lower. The performance of k-means and EM algorithms is better than hierarchical clustering algorithms. But there is no comparison between k-mean and EM algorithms in this research.

Sharmila,R C Mishra(2013)[7] studied the Performance evaluation of clustering algorithms. Data mining is the technique of running the data through the algorithms for discovering the meaningful correlations and patterns among the data that would remain hidden otherwise. Clustering is division of data into groups of similar objects . Each group called cluster consists of objects that are similar amongst themselves and dissimilar compared to objects of other groups.[7] concluded the advantage and disadvantages of various clustering algorithms.

Zaid Makani,Sana Arora,Prashasti kanikar(2013)[3] researched on A Parallel Approach to Combined Association Rule Mining.A parallel approach to combined mining has been implemented that not only generates rules which are “actionable” but also does so in a time period that is lesser than that of the traditional approach.[3] concluded the two approaches of combined mining- serial and parallel have been applied on survey dataset.

Neelamadhab Padhy, Dr. Pragnyaban Mishra , and Rasmita Panigrahi(2012)[2] worked on The Survey of Data Mining Applications And Feature Scope.[2] focused on variety of techniques, approaches and different areas of the research which are helpful and marked as the important field of data mining technologies.[2] briefly reviewed the various data mining applications.

Pratibha Mandave, Megha Mane and Prof. Sharada Patil(2013)[1] Data Mining using Association Rule Based on Apriori algorithm and Improved Approach with Illustration.[1] explain one of the useful and efficient algorithms of association mining named as APRIORI algorithm.[1] concluded that association rule is one of the efficient techniques of data mining for finding out frequent item set in transaction database.

It is difficult to mine rare association rules with single minimum support based approaches such as Apriori and Fp- growth as they suffer from rare item problem. R.Uday Kiran and P. Krishna Reddy (2010)[7] presented improved approaches to mine rare association rules in transactional databases. Rare association rules consist of rare items. They analyzed the multiple minimum supports based approach and proposed improved approaches and using experiments they showed that the improved FP-growth like approach was able to extract rare frequent patterns from the databases in which items frequencies vary widely.

There is a lack of comprehensive study on controlling false positives in association rule mining. Guimei Liu et al. (2011)[8] presented different methods to deal with the false positive errors in association rule mining.

III. PROPOSED ALGORITHMS

A. Input the dataset.

B. Apply select attribute on the dataset using Gain RatioAttributeEval and Ranker method.

C. *The Expectation (E) Step*

Each object assign to clusters with the center that is closest to the object. Assignment of object should be belonging to closest cluster.

D. *The Maximization (M) step*

For given cluster assignment, for each cluster algorithm adjust the center so that, the sum of the distance from object and new center is minimized.

The select attribute step is applied in the algorithm which decreases the time taken to form clusters.

IV. EXPERIMENT AND RESULT

A. Comparison

For performing the comparison analysis three datasest has been used, which are defined in annexure I, II, III. These datasets has been taken from web, these datasets can directly apply to data mining tools and predict the result.

Dataset name	No. Of Attributes	No. Of Instances
Dataset 1	5	150
Dataset 2	9	1253
Dataset 3	9	2924

1) *Comparison Between K-means, Expectation Maximization and Farthest First Clustering Algorithms Using Dataset 1:* A dataset of 5 attributes and 150 instances has been applied to the WEKA version 3.7.10 as referred in table 5.1(Annexure I) and the results in respect to time, number of clusters are explored as follows.

a) K-means Algorithm

Time taken to build model - 0.02 Seconds
Number of Clusters – 2
Number of iterations – 8

b) Expectation Maximization Algorithm

Time taken to build model – 1.98 Seconds

Number of Clusters – 7

Number of iterations – 39

c) Farthest First Algorithm

Time taken to build model – 0.01 seconds

Number of clusters – 2

In terms of time Farthest First Clustering Algorithm took the least time in forming clusters whereas EM took the longest time.

2) *Comparison Between K-means, Expectation Maximization and Farthest First Clustering Algorithms Using Dataset* **2:** A dataset of 9 attributes and 1253 instances has been applied to the WEKA version 3.7.10 as referred in Table 5.1(Annexure II) and the results with respect to time, number of clusters are explored.

a) K-means Algorithm
Time taken to build model - 0.03 Seconds

Number of Clusters – 2

b) Expectation Maximization Algorithm

Time taken to build model – 124.95 Seconds

Number of Clusters – 11

c) Farthest First Algorithm

Time taken to build model – 0.03 seconds

Number of clusters – 2

In terms of time Farthest First algorithm took the least time whereas the EM took longest time.

3) Comparison Between K-means, Expectation Maximization and Farthest First Clustering Algorithms Using Dataset 3: A dataset of 9 attributes and 2924 instances has been applied to the WEKA version 3.7.10 as referred in table 5.1(Annexure III) and the results with respect to time, number of clusters are explored as follows.

a) K-means Algorithm

Time taken to build model - 0.16 Seconds

Number of Clusters – 2

Number of iterations – 9

b) Expectation Maximization Algorithm

Time taken to build model – 966.41 Seconds

Number of Clusters – 21

Number of iterations – 76

c) Farthest First Algorithm

Time taken to build model – 0.09 seconds

Number of clusters – 2

In terms of time Farthest First algorithm took the least time whereas the EM took longest time.

B. Comparison Of Expectation Maximization Algorithm with Improved Expectation Maximization Algorithm Using Dataset 1

A small dataset of 5 attributes and 150 instances has been applied to the WEKA version 3.7 and the results related to time, number of clusters are explored as follows.

Time taken to form clusters using Expectation Maximization Algorithm – 1.98 seconds

Time taken to form clusters using Improved Expectation Maximization Algorithm – 1.62

Time decrease of .36 seconds is observed with Improved Expectation Maximization Algorithm

1) Graphical Representation

Volume 3 Issue VI, June 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

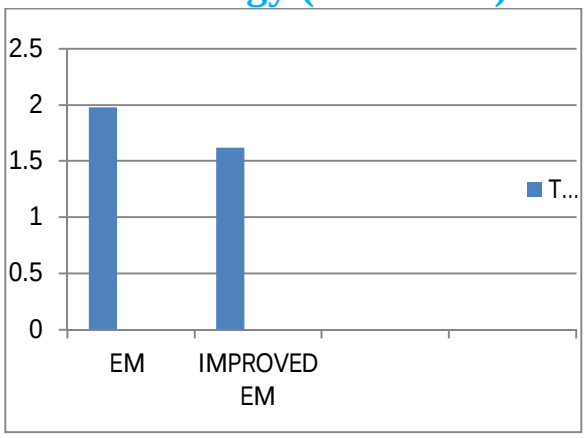


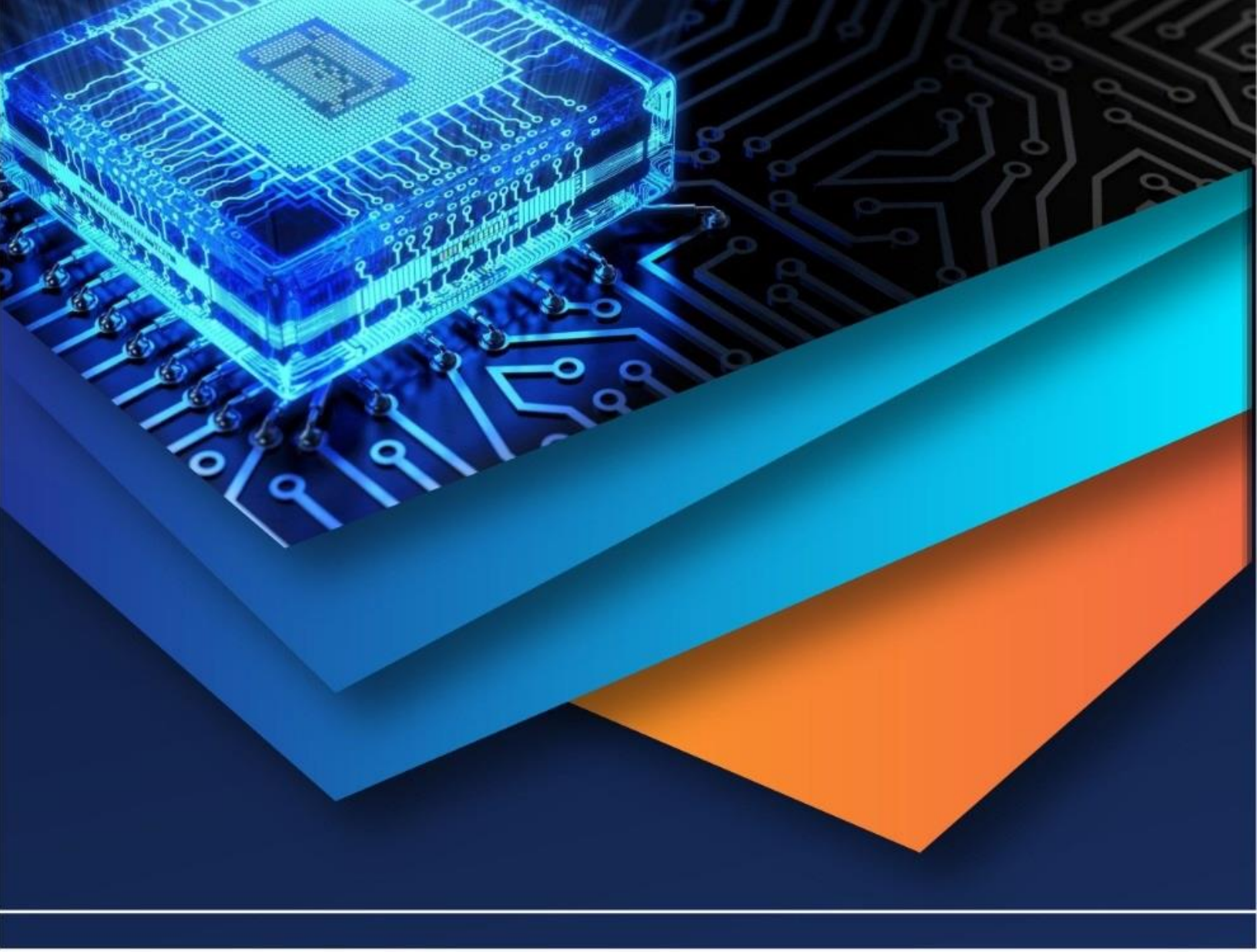
Fig – Graphical representation of time taken to form clusters Using EM and Improves EM clustering Algorithm using Dataset 1.

V. CONCLUSION

The main stages of this research work are comparison of various clustering algorithms, designing an improved clustering algorithm and comparison of the existing and proposed improved clustering algorithm. In the comparison one algorithm took maximum time in forming clusters. In improving the algorithm, the algorithm which took maximum time is improved and in comparison of existing and proposed improved algorithm, the proposed algorithm took less time in forming clusters
Improved Expectation Maximization algorithm has been made to reduce time taken to form clusters. New methods can be applied to reduce further time. On other parameters the improvisation can be applied. Other algorithms can also be improved.

REFERENCES

[1] Pratibha Mandave, Megha Mane and Prof. Sharada Patil, "Data Mining using Association Rule Based on Apriori algorithm and Improved Approach with Illustration" In International Journal of Latest Trends in Engineering and Technology Vol. 3 Issue 2, November 2013 ISSN: 2278-621X.
[2] Neelamadhab Padhy, Dr. Pragnyaban Mishra , and Rasmita Panigrahi, The Survey of Data Mining Applications And Feature Scope," International Journal of Computer Science, Engineering and Information Technology , Vol.2, No.3, June 2012.
[3] Zaid Makani,Sana Arora,Prashasti kanikar, "A Parallel Approach to Combined Association Rule Mining", International Journal of Computer Application,vol 62-No. 15, jan 2013 .
[4] R. Uday Kiran and P. Krishna Reddy, "Improved Approaches to Mine Rare Association Rules in Transactional Databases" In Proceedings of the Fourth SIGMOD PhD Workshop on Innovative Database Research, June 11, 2010 © ACM.
[5] Guimei Liu, Haojun Zhang and Limsoon Wong, "Controlling False Positives in Association Rule Mining" In Proceedings of the VLDB Endowment, Vol. 05, No. 2, August, 2011
[6] Osama Abu Abbas, "Comparisons between data clustering algorithms. " International Arab Journal of information technology,Vol.1,no.3,July 2008.
[7] Sharmila,R C Mishra, "Performance evalution of clustering algorithms. ", International Journal of Engineering Trends and technology,Vol.4,Issue7-July 2013.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY