www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 5 may, 2014 Page No. 6193-6195

IMPROVED EXPECTATION MAXIMIZATION CLUSTERING ALGORITHM

Garima Sehgal[1], Dr. Kanwal Garg[2]

[1]Research Scholar, Department of Computer Science and Applications, Kurukshetra University, India

sehgalgarima1990@gmail.com[1]

[2] Assistant Professor, Department of Computer Science and Applications, Kurukshetra University, India

gargkanwal@gmail.com[2]

Abstract - This paper focuses on the improvement of Expectation Maximization algorithm. A method attribute selection for experimentation on Expectation Maximization (EM) clustering is used. In attribute selection we used Gain Ration Attribute Eval, Ranker method for EM clustering, which gives better results than the result obtain without using attribute selection method.

Keywords - *Clustering, Expectation maximization, Gain Ration Attribute Eval ,Ranker method.*

1.INTRODUCTION

Data Mining is the data analysis technique where the data is searched in automated way to solve the problems[6]. It discovers patterns from large dataset. There are various techniques in data mining like regression, clustering, classification. In this paper clustering is used because it is applicable in various real time applications such as banking domain etc. 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[3]. Cluster is aggregation of data objects with common characteristics based on the measurement of some kind of information. There are several commonly used clustering algorithms, such as K-means, Density based and Expectation Maximizationl and so on.[1]. This paper focuses on the improvement of the EM algorithm.

1.1 EXISTING EXPECTATION MAXIMIZATION ALGORITHM

Expectation Maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the *E* step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step. EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters.[5]

The EM iteration consist of two steps expectation (E) step and maximization (M) steps:-

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.

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.[4]

Expectation Maximization has various advantages like it gives extremely useful results for real word applications such as banking, medical etc, it can handle learning problem in data mining, converges to local maximum. Besides the various advantages it has a major disadvantage that it takes large amount of time in forming clusters, which increases the cost.

In this paper section 1 gives the introduction about the EM algorithm and its advantages and disadvantages. Section 2 describes attribute selection method. Section 3 defines the dataset and tool which is used Section 4 gives the details of the improved EM algorithm. Section 5 compares the EM algorithm with the improved EM algorithm and Section 6 concludes the paper.

2. PROPOSED METHODOLOGY

The methodology used for improving the EM algorithm is feature selection. By using this, time taken to form clusters can be decreased. Feature selection is also useful in data analysis. It show which features are important and relationship between features for further analysis. Feature selection is particularly important for data sets with large numbers of features e.g. classification problems in molecular biology may involve thousands of features In

supervised learning , feature selection improves the performance of classifier in given dataset but in unsupervised learning, feature selection has very little attention . Our aim is to select best attributes to improve the time of clustering.[6]

2.1 ATTRIBUTE SELECTION

In attribute subset evaluator, It takes the subset of attributes and returns numerical measure that guide search. Gain ratio Attribute Eval , evaluates the worth of an attribute by measuring the gain ratio with respect to the class. Ranker method ranks attributes by their individual evaluations.

2.2. DATASET USED

For performing the comparison of existing and proposed EM algorithm , three dataset has been used. These datasets has been used in previous papers for comparison of various clustering algorithms. EM algorithm took the maximum time in forming clusters, so in this paper we have tried to improve the time of EM algorithm[2]. Table 1 shows the description of the three dataset i.e number of attributes and number of instances. These datasets has been collected from web (www.cs.waikato.ac. nz/ml/weka /datasets .html). Dataset 1 is in .csv format and dataset 2 and dataset 3 are in. arff format.

TABLE 1 : DATASET USED

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

3. PROPOSED IMPROVED EXPECTATION MAXIMIZATION ALGORITHM

1.Input the dataset.
2.Apply select attribute on the dataset using Gain RatioAttributeEval and Ranker method.
3.**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.
4.**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.

4. COMPARISON OF EM ALGORITHM AND IMPROVED EM ALGORITHM

Improved Expectation Maximization Algorithm reduced the time of clustering as compared to the Expectation Maximization Algorithm.

Table 2 shows the time taken by EM and Improved EM using different datasets of different sizes. Three different datasets of different sizes is used.

TABLE 2- TIME TAKEN TO FORM CLUSTERS BY EM AND IMPROVED EM CLUSTERING ALGORITHM

DATA SET USED	TIME TAKEN BY EM ALGORITHM	TIME TAKEN BY IMPROVED EM ALGORITHM
Data Set 1	1.98	1.62
Data Set 2	124.95	115.08
Data Set 3	966.41	945.41

Graphical representation

The figure 1 shows the graphical representation of time taken to form clusters by EM and Improved EM clustering algorithm.

The major disadvantage of EM algorithm is improved by the improved EM algorithm.

Results shows that the time taken by the improved EM algorithm to form clusters is less as compared to the time taken by EM to form clusters. As the size of the dataset increases, difference between the time taken to form clusters by EM and improved EM increases. This proves that the improved EM clustering algorithms shows better results when the size of dataset is large.

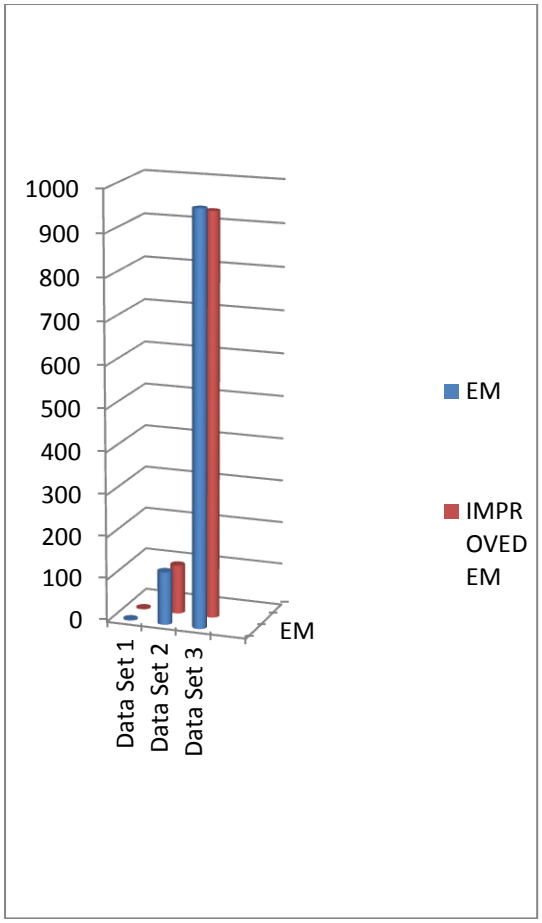

FIGURE 1 – GRAPHICAL REPRESENTATION OF TIME TAKEN TO FORM CLUSTERS USING EM AND IMPROVED EM CLUSTERING ALGORITHM.

5. ANALYSIS AND INTERPRETATION

The datasets are applied to the WEKA(3.7.10). Results shows that the time taken by the improved EM algorithm to form clusters is less as compared to the time taken by EM to form clusters. As the size of the dataset increases, difference between the time taken to form clusters by EM and improved EM increases. This proves that the improved EM clustering algorithms shows better results when the size of dataset is large.

6. CONCLUSION

An Improved EM has been designed. Performance of the improved EM algorithm and EM algorithm is compaed on the basis of time taken to form clusters.
Improved EM algorithm took less time time to form clusters. Results shows that the difference between the time taken to form clusters increases as the size of the dataset increases. This shows that the improved EM algorithm is better for large datasets.

REFERENCES

1. Amita Verma , Ashwani Kuma“ Performance Enhancement of K-means Clustering Algorithms for High Dimensional Data sets” International Journal of AdvanceResearch in Computer Science and Software Engineering, Volume 4 , Issue 1, January 2014.

2.. Garima Sehgal, Dr. Kanwal Garg “ Comparison of Various Clustering Algorithms” International Journal of Computer Science and Information Technology, Volume 5 , Issue 3, April 2014.

3. Osama Abu Abbas “ Comaprison between Data Clustering Algorithms” The International Arab Journal of Information Technology, Volume 5, July 2008.

4. Rupali Bhondave, Madhura Kalbhor “Improvement of Expectation Maximization Clustering” International Journal of Computer Science and Mobile Computing, Volume 3, Issue 4, April 2014.

5. Sharmila and R.C Mishra “Performance Evaluation of Clustering Algorithms” International Journal of Engineering Trends and Technologies, Volume 4, Issue 7, July 2013

6. Weka: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka.