Performance Evaluation of EM and K-Means Clustering Algorithms in Data Mining System

Shaik Firoj Basha
Research Scholar, Department of Computer Science
Sri Venkateswara University, Tirupati, Andhra Pradesh, India
shaikfirojbashasvu@gmail.com

Dr. S. Ramakrishna
Professor, Department of Computer Science, Sri
Venkateswara University, Tirupati, Andhra Pradesh, India
drsramakrishna@yahoo.com

***Abstract*—**In the Emerging field of Data Mining System there are different techniques namely Clustering, Prediction, Classification, and Association etc. Clustering technique performs by dividing the particular data set into associated groups such that every group does not have anything in common.Clustering algorithms have emerged as an alternative powerful meta-learning tool to accurately analyze the massive volume of data generated by modern applications. Actually the main goal is to classify data into clusters such that objects are clustered in the same cluster when they are related according to particular metrics. Classification is the organization of data sets into some predefined sets using various mathematical models. This research discusses the comparison of algorithms K-Means and Expectation-Maximization in clustering. Empirically, we focused on wide experiments where wecompared the best typical algorithm from each of the categories using a large number of real or bigdata sets. The effectiveness of the Expectation-Maximization clustering algorithm is measured through a number of internaland external validity metrics, stability, runtime and scalability tests.

Keywords- *Data Mining, Classification, Clustering, K-Means Algorithm, Expectation-Maximization Algorithm.*

*****

I. INTRODUCTION

Data mining is a technique used in Big Data analytics for discovering hidden correlations and pattern in data from data warehouses which cannot be obtained using traditional techniques. Classification is one of the important forms of data analysis, but most of the classification algorithms are only suitable for small data sets, with the increasing amount of data and dimensions, it is very important to establish an efficient classification algorithm for large data sets. Data clustering is one of the most traditional and important issues in computer science. In recent years, due to emerging applications such as data mining and document clustering, data clustering has attracted a new round of attention in computer science research communities. Big Data Analytics is the approach of examining the stored data to identify some hidden patterns and correlations among the data. Big data analytics can be used in any field where a large data is generated. Fields ranging from technology to medical field, from petroleum to space research program. Big data analysis gives various useful data which is very important from economic as well as non-economic point of view[1,4]. Big data is defined by three characteristics generally known as three V's. These three V's stand for Velocity, Variety and Volume.

Velocity - It is the rate at which the data is coming in to an organization.
Variety - It relates the varied type of data like Structured, Unstructured and Semi-structured.
Volume - The size of data that is flowing in to an organization.

II. LITERATURE REVIEW

Data mining is a technique used in Big Data analytics, for discovering hidden correlations and pattern in data from data warehouses which cannot be obtained using traditional techniques. Analysts use data mining for formulating various statistical methods and recognition of pattern in the data. One of the data mining techniques is clustering. Clustering is dividing the data set into groups such that data points with similar properties are grouped together[6,9]. There are various algorithms that can perform clustering. These algorithms are broadly classified into the following categories:

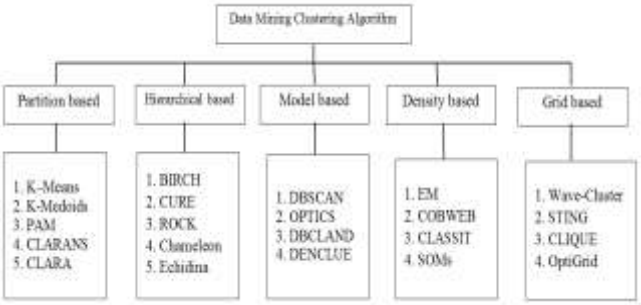

Figure 1: Clustering Algorithms

III. PARTITION BASED CLUSTERING TECHNIQUE

Partition based clustering technique divides data points in a dataset into various partitions (cluster). It optimizes cluster quality by iteratively performing on any objective function. Each Cluster has minimum one data point. Also no data point can be present in more than one cluster[8,10]. Disadvantages of this algorithm are we need to define number of clusters upfront in most of algorithm.

A. *K-Means clustering algorithm*

It one of most simple clustering algorithm which is usedtosolve problem of clustering by forming clusters iteratively.The main idea is to define K centroids. In K-means algorithmwe need to define numbers of Clusters (i.e. K Cluster) atbeginning. Then any K points from dataset are selected to becentroid. Then for each point calculate centroid-data pointdistance. Based on these distance, the point is associated withnearest centroids. All the data points are divided into numberof clusters based on distance of data points from centroid ofcluster. Centroid is unique point for each partition. Centroid isthe point from where distance is calculated for each data point.This distance can be calculated using Manhattan distance,Euclidean distance, cosine similarity etc. Once all the datapoints are placed, all K centroids are calculated again. Newcentroid is mean of all point in cluster[2,7]. Then all data points arereassigned to cluster with respect to new centroids bycalculating centroid-data point distance. This is doneiteratively till certain criterion is satisfied.Objective function is sum of square distance. i.e.x_{ij} is j^{th}data point of i^{th}cluster. $m_i$is centroid of i^{th} cluster.

Algorithm:

Step 1: Define number of Clusters and then select same number ofdata points as centroids.

Step 2: Calculate distance of a point from all centroid. Assign the point to cluster with minimum centroid-pointdistance.

Step 3: Repeat step 2 for all points.

Step 4: Calculate the mean of all point in a cluster and assign it asnew centroid for that cluster.

Step 5: Repeat from step 2, until desired clusters or certain criteriaare satisfied.

Since initial centroid are selected randomly results of clustersdepends on initial centroids. There are many methods todetermine initial centroids to optimizeclusters. Complexity ofk-means algorithm is O(tnk), where k is number of clusters, tis number of iterations and n being number of data sets. Butt,k<<n. therefore, complexity is O(n).

Advantages:

1. It simple to implement K-means algorithm.
2. It is efficient algorithm with complexity O(n).
3. Produces denser clusters than the hierarchical methodespecially when clusters are spherical.

IV. DENSITY-BASED CLUSTERING TECHNIQUE

As there are so many clustering algorithms; this section introduces a categorizing framework that groups the various clustering algorithms found in the literature into distinct categories. The proposed categorization framework is developed from an algorithm designer's perspective that focuses on the technical details of the general procedures of the clustering process. Here, data objects are separated based on their regions of density, connectivity and boundary they are closely related to point-nearest neighbors. A cluster, defined as a connected dense component, grows in any direction that density leads to. Therefore, density-based algorithms are capable of discovering clusters of arbitrary shapes[5,12]. Also, this provides a natural protection against outliers. Thus the overall density of a point is analyzed to determine the functions of datasets that influence a particular data point. DBSCAN, OPTICS, DBCLASD and DENCLUE are algorithms that use such a method to filter out noise (outliers) and discover clusters of arbitrary shape.

A. *EXPECTATION-MAXIMIZATION ALGORITHM*

Expectation-Maximization algorithm is designed to estimate the maximum likelihood parameters of a statistical model in many situations, such as the one where the equations cannot be solved directly. EM algorithm iteratively approximates the unknown model parameters with two steps: the E step and the M step. In the E step (expectation), the current model parameter values are used to evaluate the posterior distribution of the latent variables. Then the objects are fractionally assigned to each cluster based on this posterior distribution. In the M step (maximization), the fractional assignment is given by re-estimating the model parameters with the maximum likelihood rule. The EM algorithm is guaranteed to find a local maximum for the model parameters estimate. The major disadvantages for EM algorithm are: the requirement of a non-singular covariance matrix, the sensitivity to the selection of initial parameters, the possibility of convergence to a local optimum, and the slow convergence rate[3,11]. Moreover, there would be a decreased precision of the EM algorithm within a finite number of steps. The details of the EM algorithm are given below.

B. *Expectation-Maximization Algorithm pseudo-code*

*Input:*the dataset (x), the total number of clusters (M), the accepted error for convergence (e) and the maximum number of iterations.

E-step: compute the expectation of the complete data log-likelihood.

$$Q(\theta, \theta^T) = E\ [\log p(x^g, x^m|\theta) x^g, \theta^T]$$

*M-step:*select a new parameter estimate that maximizes the Q-function,

$$\theta^{t+1} = \arg\max_{\theta} Q(\theta, \theta^T)$$

*Iteration:*increment t = t+1; repeat steps until the convergence condition is satisfied.

Output: A series of parameter estimates $\{\theta^0, \theta^1, \ldots\ldots, \theta^T\}$, which represents the achievement of the convergence criterion.

V. IMPLEMENTATION OF THE ALGORITHMS AND RESULTS

We have taken a weather dataset from the repository which is having 5 attributes outlook, temperature, humidity, windy, play along with 14 instances to test the clustering performance of these two algorithms. We compared the performance of the EM Algorithm and K-Means Clustering Algorithm using data mining tool. The results of these algorithms are shown below.

TABLE 1: PERFORMANCE COMPARISON OF ALGORITHMS

Name of the Algorithm	Number of Clusters Formed	Name of the Clusters	Clustered Instances	Percentage of Clustering
Expectation-Maximization	1	Cluster(0)	14	100%
K-Means	2	Cluster(0) Cluster(1)	9 5	64% 36%

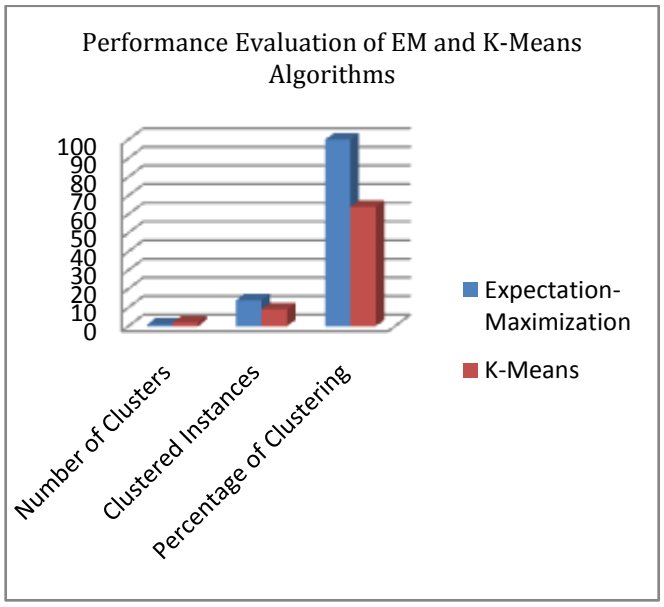

Figure 2: Graphical Representation Performance of EM and K-Means Algorithms

VI.CONCLUSION

The main goal of this research is to provide readers with a proper analysis of the different classes of available clustering techniques for big data by experimentally comparing them on real big data. Theother important characteristic of big data is Velocity. This requirement leads to a high demand for online processing of data, where processing speed is required to deal with the data flows. Variety is the third characteristic, where different data types, such as text, image, and video, are produced from varioussources, such as sensors, mobile phones, etc. In this research we compared the two algorithms Expectation-Maximization and K-Means by using the weather dataset to evaluate the clustering performances. After execution of these two algorithms we found that the EM algorithm gives good clustering performance when compared to K-Means. This kind of clustering techniques aims to produce a good quality of clusters. Therefore, they would hugely benefit everyone from ordinary users to researchers and people in the corporate world, as they could provide an efficient tool to deal with large data such as critical systems (to detect cyber-attacks). The directions of the future work depend on the clustering models of the algorithms with real world big datasets.

REFERENCES

[1]. B.J. Frey and D. Dueck, “Clustering by Passing Messages between Data Points,” Science, vol. 315, pp. 972-976, 2007.
[2]. Christoph F. Eick, Nidal Zeidat, and Zhenghong Zhao, "Supervised Clustering – Algorithms and Benefits", 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004), 2004.
[3]. Dr.M.Jayakameswariah, Mr.M.Veeresh Babu, Dr.S.Ramakrishna, Mrs.P.Yamuna, "Computation Accuracy of Hierarchical and Expectation Maximization Clustering Algorithms for the Improvement of Data Mining System",International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 12, ISO 9001:2008, Page 1580-1585, e-ISSN: 2395 -0056, p-ISSN: 2395-0072, Dec -2016.
[4]. H. Huang, Y. Gao, K. Chiew, K, L. Chen and Q. He, “Towards Effective and Efficient Mining of Arbitrary Shaped Clusters,” IEEE 30th ICDE Conference, pp. 28- 39, 2014.
[5]. Mehmed Kantardzic., Jozef Zurada. “Next Generation of Data-Mining Applications.” New York : Wiley-IEEE Press 3 (2005).
[6]. Mitra, S., Pal, S. K., & Mitra, P,”Data mining in soft computing framework: A survey”. IEEE Transactions on Neural Networks, 13, 3–14, 2002.
[7]. M.Jayakameswaraiah and S.Ramakrishna, "A Study on Prediction Performance of Some Data Mining Algorithms", International Journal of Advanced Research in Computer Science and Management Studies, Volume 2, Issue 10, , ISSN: 2321-7782, October 2014.
[8]. R. Krakovsky and R. Forgac,”Neural network approach to multidimensional data classification via clustering”, Intelligent Systems and Informatics (SISY),IEEE2011, IEEE 9th International Symposium on, 169–174, 2011.
[9]. Ranshul Chaudhary, Prabhdeep Singh, Rajiv Mahajan,” A SURVEY ON DATA MINING TECHNIQUES”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 1, ISSN (Online): 2278-1021, January 2014.
[10]. Wai-Ho Au, Member, IEEE, Keith C. C. Chan, Andrew K.C. Wong, Fellow, IEEE,and Yang Wang, Member, IEEE ,”Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data”, Sep. 15, 2004.
[11]. Yinghua Lv, TinghuaiMa, MeiliTang, JieCao, YuanTian ,Abdullah Al-Dhelaan , MznahAl-Rodhaan, “An efficient and scalable density-based clustering algorithm for datasets with complex structures”, Elsevier, 23 march 2015.
[12]. Y. Chen, E.K. Garcia, M.R. Gupta, A. Rahimi, and L. Cazzanti, “Similarity-Based Classification: Concepts and Algorithms,” J. Machine Learning Research, vol. 10, pp. 747-776, 2009.