International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

# Comparing EM Clustering Algorithm with Density Based Clustering Algorithm Using WEKA Tool

**Dr. Abdelrahman Elsharif Karrar[1], Moez Mutasim[2]**

[1] College of Computer Science and Engineering, Taibah University, Saudi Arabia

[2] University of Science and Technology, Sudan

**Abstract:** ***Machine learning is type of artificial intelligence wherein computers make predictions based on data. Clustering is organizing data into clusters or groups such that they have high intra-cluster similarity and low inter cluster similarity. This paper deals with two clustering algorithms which are EM and Density based algorithm. EM algorithm is general method of finding the maximum likelihood estimate of data distribution when data is partially missing or hidden. In Density based clustering, clusters are dense regions in the data space, separated by regions of lower object density. The comparison between the above two algorithms is carried out using open source tool called WEKA, with the Weather dataset as its input.***

**Keywords:** Machine learning, Unsupervised learning, supervised learning, EM clustering, Density based clustering, WEKA

## 1. Introduction

Machine learning is type of artificial intelligence wherein computers make predictions based on data. Machine learning broadly classified into supervised classification and unsupervised classification. In supervised systems, the data as presented to a machine learning algorithm is fully labeled. In supervised learning the variables can be split into two groups: explanatory variables and one (or more) dependent variables. [1]

One of unsupervised learning technique is clustering. Clustering is organizing data into clusters or groups such that they have high intra-cluster similarity and low inter cluster similarity.

There are different types of clustering techniques namely K-means clustering, Hierarchical clustering, Exception-maximization clustering and density based clustering. [2]

WEKA is one of the open source tool, is a collection of machine learning algorithms for solving real-world. It is written in Java and runs on almost any platforms. [3]

## 2. Clustering

Clustring is a data mining technique to group the similar data into a cluster and dissimilar data into different clusters. Clustering can be considered the most important unsupervised learning technique so as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.

Clustering is the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters.
A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups. Moreover, most of the data collected in many problems seem to have some inherent properties that lend themselves to natural groupings.

Clustering algorithms are used extensively not only to organize and categorize data, but are also useful for data compression and model construction. [4]

Here we deal with two clustering algorithms which are EM and Density based algorithm, and we use weka software to compare between these two algorithms.

## 3. EM(Expectation–maximization) Algorithm

It is general method of finding the maximum likelihood estimate of data distribution when data is partially missing or hidden. [5]

The two steps are:
**E (Exception) step**: This step is responsible to estimate the probability of each element belong to each cluster P(C_j|x_k). Each element is composed by an attribute vector (xk). The relevance degree of the points of each cluster is given by the likelihood of each element attribute in comparison with the attributes of the other elements of cluster Cj.

$$P(C_j|x)=\frac{|\sum_j(t)|^{-\frac{1}{2}}\ exp^{n_j}\ P_j(t)}{\sum_{k=1}^{M}\ |\sum_j(t)|^{-\frac{1}{2}}\ exp^{n_j}\ P_k(t)}$$

Where,
x is input dataset.
M is the total number of clusters
t is an instance and initial instance is zero

a) **M (maximization) step**: This step is responsible to estimate the parameters of the probability distribution of each class for the next step. First is computed the mean (μj) of class j obtained through the mean of all points in function of the relevance degree of each point. The covariance matrix at each iteration is calculated using Bayes theorem. The probability of occurrence of each class is computed through the mean of probabilities (C_j) in function of the relevance

degree of each point from the class.

$$P_j(t+1) = \frac{1}{N}\sum_{k=1}^{N} P(C_j|x_k)$$

Where,
x is input dataset.
M is the total number of clusters
t is an instance and initial instance is zero. [6]

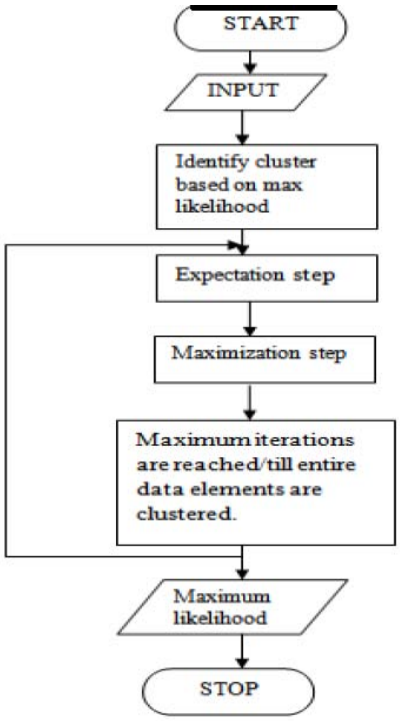


**Figure 1:** Flowchart for EM Algorithm

## 4. Density Based Algorithm

Density-based algorithm is another major clustering algorithm that has been long proposed. It can find arbitrarily shaped clusters and handles noises and yet is a one-scan algorithm that needs to examine the raw data only once. In density-based clustering algorithms, dense areas of objects in the data space are considered as clusters, which are segregated by low-density area (noise). Therefore, density-based method is an attractive basic clustering algorithm for data streams. [7]

The basic idea of density-based clustering is clusters are dense regions in the data space, separated by regions of lower object density. Intuition for the formalization of the basic idea is:

For any point in a cluster, the local point density around that point has to exceed some threshold. The set of points from one cluster is spatially connected Two global parameters are:
○ **Є(Eps)**:Maximum radius of the neighbourhood
**MinPts**: Minimum number of points in an Є - neighbourhood of that point

Density-based clustering regard clusters as dense areas that are separated by low density area. Traditional density-based methods are DBSCAN, OPTICS and DENCLUE.

DBSCAN and its extension, OPTICS2, are both typical density-based methods that grow clusters according to a density-based connectivity analysis in spatial data set. DENCLUE3 is a method that clusters objects based on the analysis of the value distributions of density functions. [7]

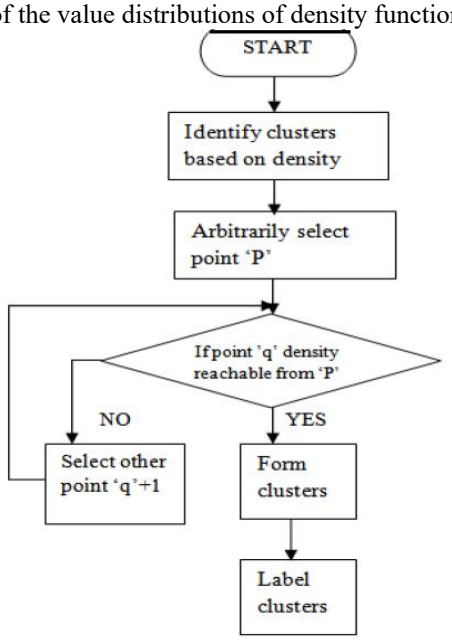


**Figure 2:** Flowchart for Density Based Algorithm

## 5. WEKA

WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-of-the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. [8]

WEKA implements algorithms for data preprocessing, classification, regression, clustering and association rules. It also includes visualization tools. To performing cluster analysis in weka. The dataset is needed to be loaded to weka and it should be in the format of CSV or .ARFF file format. If the dataset is not in arff format we need to be converting it. [8]

## 6. Comparison of EM and density based algorithm using WEKA tool

WEKA (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software. The WEKA workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. [3] The EM algorithm is run using insurance dataset. The figure 3 shows the output for EM algorithm. There are three attributes namely 'id', 'paytype', 'doctype'. There are 68 instances.

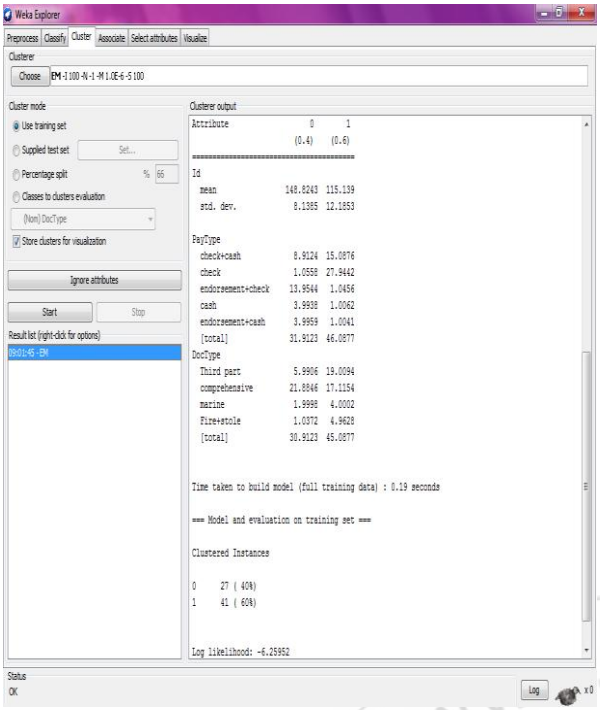


**Figure 3:** EM Clustered Output

The Density based algorithm is run using insurance dataset. The figure 4 shows the output for Density based algorithm. There are three attributes namely 'id', 'paytype', 'doctype'. There are 68 instances

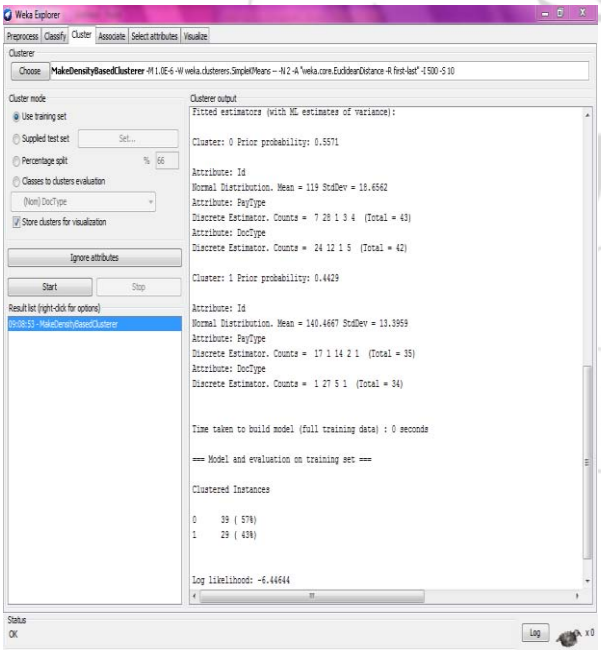


**Figure 4:** Density Based Clustered Output

**Table 1:** Comparison between EM and Density Based Algorithm

| Algorithm Name | Log-Likelihood | Time taken to build the model | Clustered instances |
|---|---|---|---|
| EM algorithm | -6.25952 | 0.19 seconds | 2 |
| Density-Based algorithm | -6.44644 | 0 seconds | 2 |

Referring to table1, Log likelihood here refers to probability of identifying correct group of data elements. In terms of likelihood density-based algorithm is better than EM algorithm. We can infer that Density-based algorithm takes less time than EM algorithm to build the model. Likelihood is often used as a synonym for probability. It is more convenient to work with the natural logarithm of the likelihood function, called the log-likelihood. Log likelihood here refers to probability of identifying correct group of data elements. In terms of likelihood EM algorithm is better than density based algorithm, referred to Table 1. From Table 1 we can infer that Density based algorithm takes less time than EM algorithm to build the model.

## 7. Conclusion

Clustering is organizing data into clusters or groups such that they have high intra-cluster similarity and low inter cluster similarity. EM algorithm is general method of finding the maximum likelihood estimate of data distribution when data is partially missing or hidden. Density based clustering, clusters are dense regions in the data space, separated by regions of lower object density. WEKA an open source tool is used for comparing the above two algorithm. In terms of likelihood EM algorithm is better than density based algorithm, referred to Table 1. From Table 1 we can infer that Density based algorithm takes less time than EM algorithm to build the model.

## References

[1] Kyu-Young Whang,"Statistical pattern recognition: a review, Pattern Analysis and Machine Intelligence", IEEE Transactions, August 2002.
[2] Department of Computer Science and Engineering University of Washington, "A Few Useful Things to Know about Machine Learning".
[3] Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya, "Comparison of clustering algorithms using WEKA tool", International Journal of Emerging Technology and Advanced Engineering, (ISSN 2250-2459, Volume 2, Issue 5, May 2012.
[4] S. Manish Verma, Neha Chack, A. Kumar Diswar, Nidhi Gupta, "A Comparative Study of Various Clustering Algorithms in Data Mining," 2012.
[5] Peng Shangu, Wang Xiwu, "The study of EM algorithm based on forward sampling", Zhong Qigen Electronics, Communications and Control (ICECC), 2011.
[6] David Sergio, Carlos Ordonez, "A Fast Convergence Clustering Algorithm Merging MCMC and EM Methods", October 2013.
[7] A. Amini, R.Saybani, S.Yazdi, "A Study of Density-Grid based Clustering Algorithms on Data Streams," 2011.
[8] K. Mintwal, "Comparison the Various Clustering and Classification Algorithms of WEKA Tools", Volume 3, Issue 12, December 2013.