International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 9, Issue 8 (January 2014), PP. 19-24

Comparative Analysis of EM Clustering Algorithm and Density Based Clustering Algorithm Using WEKA tool.

Prajwala T R[1], Sangeeta V I[2]
Assistant professor, Dept. of CSE, PESIT, Bangalore

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. The two clustering algorithms considered 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 it's input.

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

I. 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 labelled. In supervised learning the variables can be split into two groups: explanatory variables and one (or more) dependent variables[1]. The target of the analysis is to specify a relationship between the explanatory variables and the dependent variable. In unsupervised learning situations all variables are treated in the same way, there is no distinction between explanatory and dependent variables. Unsupervised systems are not provided any training examples.

Supervised learning includes classification and regression techniques. Classification technique involves identifying category of new dataset. Regression is a statistical method of identifying relationship between variables of dataset[11].

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[10].

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 platform[5].

1. Clustering Technique

Clustering is the unsupervised classification of patterns - observations, data items, or feature vectors into groups (clusters) which have same features. The two properties of a cluster are

i. High intra cluster similarity.
ii. Low inter cluster similarity.

Consider the following example,

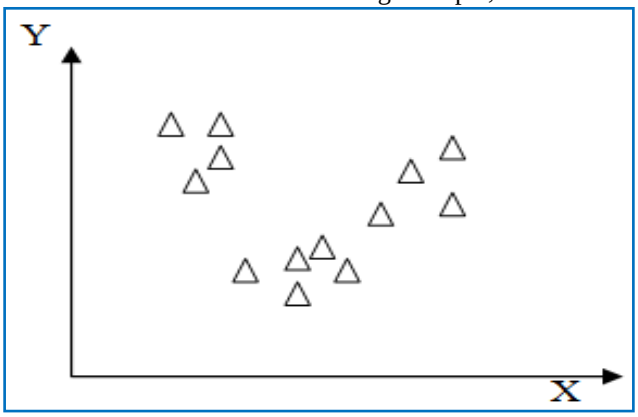

Figure 1:set of elements in dataset

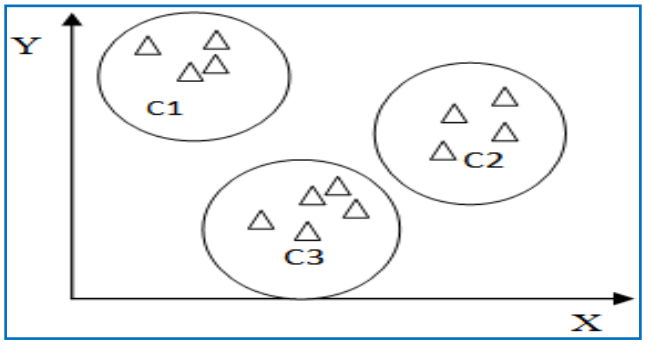

Figure 2:clustered elements in dataset

Figure 1 shows the set of data elements. Based on the positions in the figure1,the data elements are grouped into clusters C1,C2 and C3 as shown in figure 2[12].

2. EM(Expectation–maximization) algorithm

It is general method of finding the maximum likelihood estimate of data distribution when data is partially missing or hidden[3]. The two steps are:

1. 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 (x_k). 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 C_i.

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

Where,

x is input dataset.

M is the total number of clusters

t is an instance and initial instance is zero.

2. 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 (μ_j) 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.

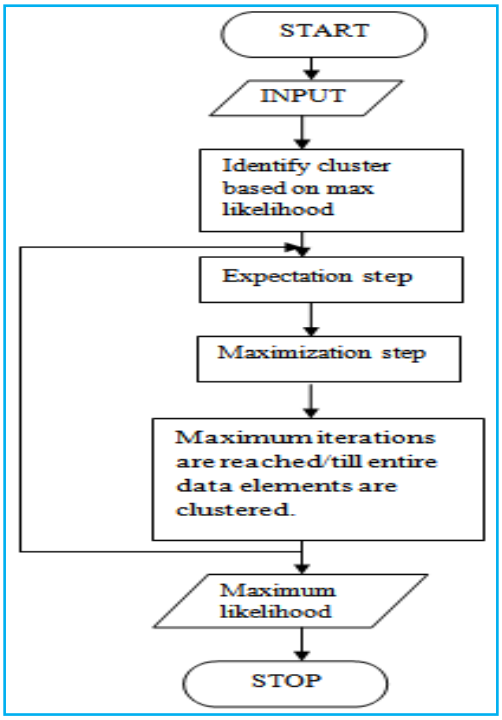

Figure 3: Flowchart for EM algorithm

Where, x is input dataset.

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

M is the total number of clusters t is an instance and initial instance is zero[8].

4. Density based clustering

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 [2],

i. For any point in a cluster, the local point density around that point has to exceed some threshold
ii. The set of points from one cluster is spatially connected

Two global parameters are[6]:

i. Є(Eps):Maximum radius of the neighbourhood
ii. MinPts: Minimum number of points in an Є -neighbourhood of that point

Core object is object with at least MinPts objects within a radius 'Є -neighborhood'. Border object is object that on the border of a cluster.

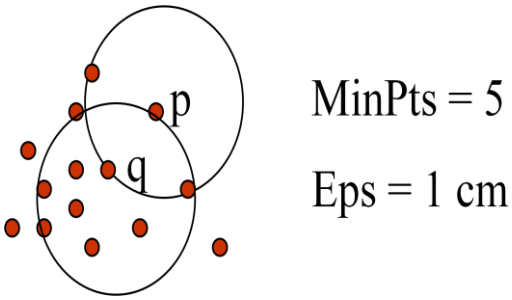

Figure 4: Illustration of global parameters of Density based clustering algorithm

4.1 Density-reachable and Density connectivity

Є -Neighborhood – Objects within a radius of Є from an objectDensity reachable- An object q is directly density-reachable from object p if p is a core object and q is in p's Є –neighborhood[6].

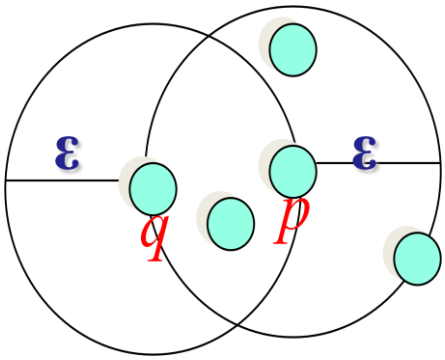

Figure 5:Illustration of density reachablity

Density-Connected-A pair of points p and q are density-connected if they are commonly density-reachable from a point o[12].

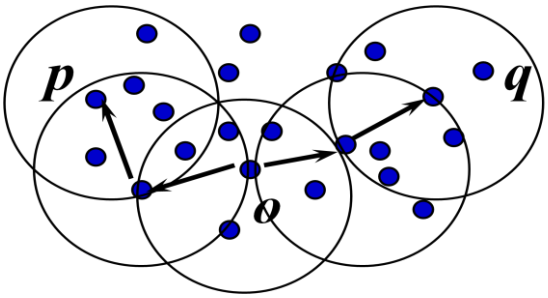

Figure 6: Illustration of density connection of points

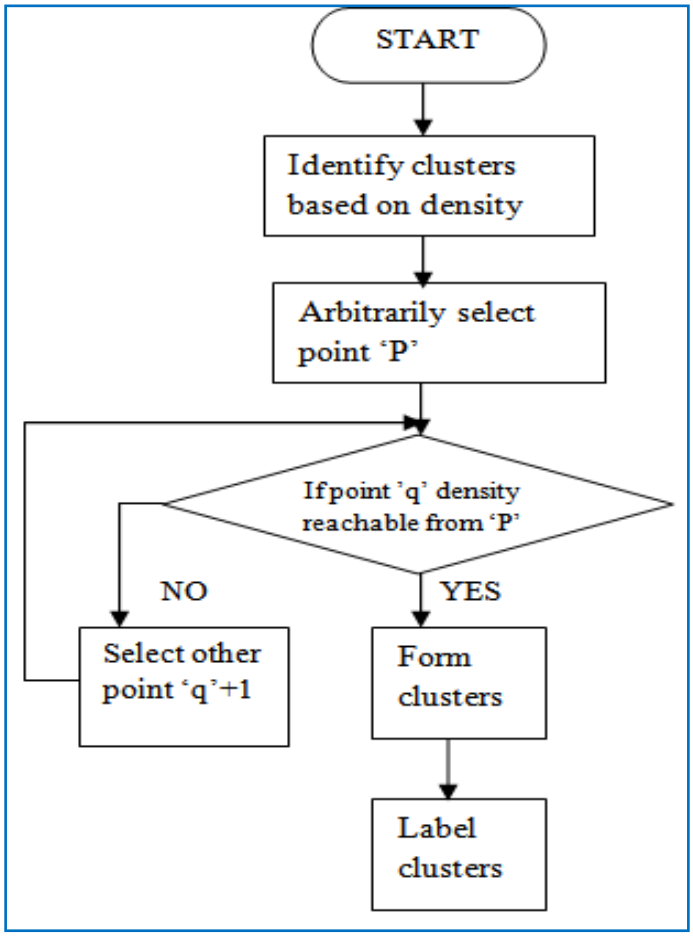

Figure 7: flowchart for Density based algorithm

5. 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[5].

The EM algorithm is run using Weather dataset. The figure 6 shows the output for EM algorithm. There are five attributes namely 'outlook', 'Humidity', 'temperature', 'windy', 'play'. There are 14 instances.

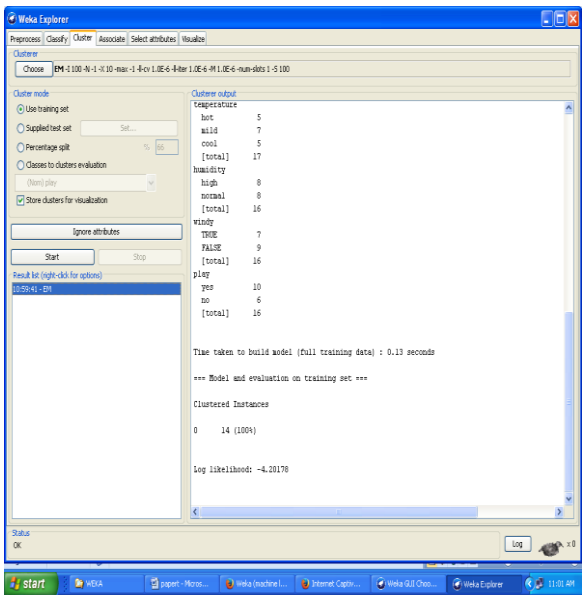

Figure 8: EM clusterer output.

The Density based algorithm is run using Weather dataset. The figure 7 shows the output for EM algorithm. There are five attributes namely 'outlook', 'Humidity', 'temperature', 'windy', 'play'. There are 14 instances.

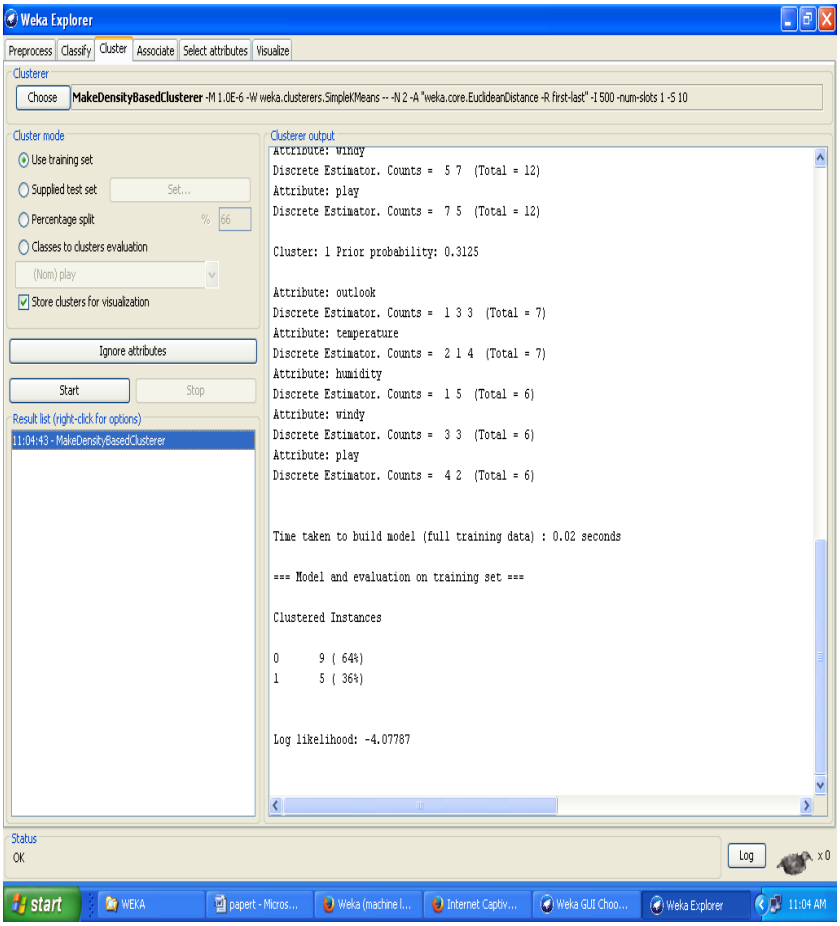

Figure 9: Density based clusterer output

Comparison between EM and Density based algorithm is shown in **Table 1**.

	Log-Likeli-hood	Time taken to build the model	Clustered instances
EM algorithm	-4.2017	0.06 seconds	1
Density based algorithm	-4.0778	0.02 seconds	2

Table 1: comparison between EM and Density based algorithm

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.

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]. Statistical pattern recognition: a review, Pattern Analysis and Machine Intelligence, Kyu-Young Whang, IEEE Transactions, August 2002, P: 4 - 37

[2]. A top-down approach for density-based clustering using multidimensional indexes Jae-Joon Hwan, Kyu-Young Whang, Yang-Sae Moon, Byung-Suk Lee, The Journal of Systems and Software 73 (2004) 169–180

[3]. The study of EM algorithm based on forward sampling. Peng Shangu, Wang Xiwu ; Zhong Qigen Electronics, Communications and Control (ICECC), 2011 , Pages 4597 – 4600

[4]. A fast density based clustering algorithm for spatial database system. Computer and Communication Technology (ICCCT), 2011 2nd International Conference, Pages: 1652 – 1656

[5]. Comparison of clustering algorithms using WEKA tool, Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya, International Journal of Emerging Technology and Advanced Engineering, (ISSN 2250-2459, Volume 2, Issue 5, May 2012.

[6]. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining,2012

[7]. A Density-Based Clustering Structure Mining Algorithm for Data Streams ,Huan Wang, Yanwei Yu , Qin Wang, Yadong Wan, proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications , August 2012, Pages 69-76

[8]. A Fast Convergence Clustering Algorithm Merging MCMC and EM Methods ,David Sergio Matusevich, Carlos Ordonez, Veerabhadran Baladandayuthapani, proceedings of the 22nd ACM international conference on Conference on information & knowledge management, October 2013, Pages 1525-1528

[9]. Data Clustering: A Review, A.K.JAIN Michigan State University

[10]. A Few Useful Things to Know about Machine Learning, Department of Computer Science and Engineering University of Washington

[11]. http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/

[12]. http://www.cs.put.poznan.pl/jstefanowski/sed/DM-7clusteringnew.pdf