International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

Enhance Clustering Mechanism using Expectation Maximization Algorithm for Gaussian Model

Rajbala Suthar[1]
[1]*Chaudhary Devilal University, Sirsa, India*

Abstract: EM is frequently used for data clustering in machine learning & computer illusion. In normal language dispensation two well-known instances of algorithm are Baum-Welch algorithm & inside-outside algorithm for unsupervised induction of probabilistic free grammars. Data mining technologies are open to all people with IoT technologies for decision making support & system optimization. Data mining involves discovery novel, interesting, & potentially useful model from data & applying algorithms to extraction of no hide information Due to increasing amount of data available online, World Wide Web had becoming one of most valuable resources for information retrievals & knowledge discoveries.
Keywords: Data mining, web mining, web intelligence, knowledge discovery, fuzzy logic

I. INTRODUCTION

It involves use of complicated data study tools to discover previously unknown, valid patterns & relationships in large data sets. These tools could include statistical models, machine learning methods like neural networks or decision trees. Accordingly, data mining [5] consists of more than collecting & managing data; it also includes study & calculation. Objective of data mining is to recognize valid, potentially helpful & understandable correlations & patterns in existing data. Finding useful patterns in data is known as different names.

II. LITERATURE REVIEW

A. Bhawna Nigam (2011) Document Classification Using Expectation Maximization within Semi Supervised Learning

As amount of online document increases, demand for document classification to aid analysis & management of document is increasing. Text is cheap, but information, in form of knowing what classes a document belongs to, is expensive. Chief purpose of this chapter is to explain expectation maximization recited of data mining to arrange document & to learn how to improve accuracy while using semi-supervised approach. Expectation maximization algorithm is useful within both supervised & semi-supervised methods.

B. David Jensen & Jennifer Neville (2012) Data Mining in Social Networks[1]

Several techniques for learning statistical models have been developed recently by researchers in machine learning & data mining. All of these secrets must address a similar set of representational algorithmic choices & must face a set of statistical challenges unique to learning from relational data. However, little this work had been made good use of research in other areas, such as social network analysis & statistics. Cross-disciplinary efforts & joint research efforts should be encouraged to promote rapid development & dissemination of useful algorithms & data representations. This work should focus on unique statistical challenges raised by relational data.

C. Nikita Jain1, Vishal Srivastava2 Nov (2013) Data mining techniques: A Survey paper [2]

Data mining based on Neural Network & Inherited Algorithm is researched in detail & key technology & ways to achieve data mining on Neural Network & Genetic Algorithm are also surveyed. This paper also conducts a formal review of area of rule extraction from ANN & GA. If conception of computer algorithms being based on evolutionary of organism is surprising, extensiveness within which these methodologies are applied in so many areas is no less than astonishing. At present data mining is a new & important area of research & ANN itself is a very suitable for solving problems of data mining because its characteristics of better robustness, self-organizing adaptive, distributed storage & high degree of fault tolerance. Commercial, educational & scientific applications are increasingly dependent on these methodologies.

D. Aarti Sharma (2014) Application of Data Mining – A Survey Paper [3]

Data mining is a powerful & a new field having various techniques. It converts raw data into useful information in various research fields. It helps in finding patterns to decide future trends in medical field.

E. Muhammad Husain Zafar (2015) A Clustering Based Study of Classification [4]

A grouping of data objects such that objects within a group are similar to one another & different from (or unrelated to) objects in other groups. This paper intends to study & compare different clustering algorithms. These algorithms include K-Means, Farthest First, CURE, Chameleon algorithm. All algorithms are compared to each other on basis of their pros &cons, similarity measure & working, functionality & time complexity. In his paper they present brief & easy comparison between different clustering algorithms. They also evaluate these algorithms on different datasets & present results through tables.

III. OBJECTIVES

A. To make study of existing clustering mechanisms along with there limitations.
B. To study the existing EM clustering model and make traditional EM based simulation.
C. To simulate the proposed EM clustering model using MATLAB and applying graphical user.
D. To take a dataset to make comparative analysis of EM based clustering and other clustering techniques using WEKA Tool.
E. To study the scope of EM clustering and conclude the probability of usage of clustering mechanisms.

IV. RESEARCH METHODOLOGY

Data mining techniques for effective automated discovery of earlier unknown, valid, novel, useful & understandable patterns in large databases. Patterns must be actionable so that they might be used in an enterprise's decision making process. Description of document classification using Expectation algorithm

In statistics, an expectation–maximization (EM) algorithm is an iterative method to find maximum probability or maximum a posteriori (MAP) estimates of parameters in statistical models, where model depends on unobserved latent variables. Prospect Maximization alternates between performing an expectation step, which creates a function for expectation of log-probability evaluated using current estimate for parameters, & maximization step, which computes parameters maximizing expected log-probability found on E step.

A. Study of existing EM that is making it Magical

EM could occasionally get stuck in a local maximum as you estimate parameters by maximizing log-likelihood of observed data, there are three things that make it magical is ability to simultaneously optimize a large number of variables & ability to find good estimates for any missing information in data at same time. Other work it Magical in context of clustering data that lends itself to modelling by a Gaussian mixture, ability to create both traditional hard clusters & not-so-traditional soft clusters

B. Clustering Three Dimensional Data

With regard to ability of EM to simultaneously optimize a large number of variables, consider case of clustering three dimensional data

C. Using Gaussian cluster in 3d space

Gaussian cluster in three dimensions space is feature by following ten variables: six unique elements of 3×3 covariance matrix which could be symmetric & positive-definite, 3 unique elements of mean, & prior associated with Gaussian. –Now let's say you expect to see six Gaussians in your data. When you would need values for fifty five variables remember unit-summation constraint on class priors which reduces overall number of variables by one to be estimated by algorithm that seeks to discover clusters in data.

D. Implementation of EM algorithm for Gaussian mixture model

A mixture model has been explain by assuming that every observed data point had a corresponding unobserved data point, or latent variable, specifying mixture component that each data point belongs to. The algorithm is an iterative algorithm that starts from some initial estimate of Θ & then proceeds to iteratively update Θ until convergence is detected. Each iteration consists of an E-step & an M-step.E-Step: Denote current parameter values as Θ. Compute w_{ik} (using equation above for membership weights) for all data points x_i, $1 <= i <= N$ & all mixture components $1 <= k <= K$.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

M-Step: P Now use membership weights & data to calculate new parameter values.

V. IMPLEMENTATION & RESULT

A. Implementation of EM algorithm for Gaussian mixture model

A mixture model would be explain more simply by assuming that each observed data point has a corresponding unobserved data point, or latent variable, specifying the mixture component that each data point belongs to.

B. EM algorithm for Gaussian mixture model

Finding a maximum likelihood solution typically requires taking the derivatives of the likelihood function with respect to all the unknown values the parameters and the latent variables and simultaneously solving the resulting equations. In statistical models with latent variables, this usually is not possible. Instead, the result is typically a set of interlocking equations in which the solution to the parameters requires the values of the latent variables and vice versa, but substituting one set of equations into the other produces an unsolvable equation.

C. Maximization function

```
function model = maximization(X, R)
[d,n] = size(X);
k = size(R,2);
sigma0 = eye(d)*(1e-6); % regularization factor for covariance
s = sum(R,1);
w = s/n;
mu = bsxfun(@rdivide, X*R, s);
Sigma = zeros(d,d,k);
for i = 1:k
   Xo = bsxfun(@minus,X,mu(:,i));
   Xo = bsxfun(@times,Xo,sqrt(R(:,i)'));
   Sigma(:,:,i) = (Xo*Xo'+sigma0)/s(i);
end
model.mu = mu;
model. Sigma = Sigma;
model. weight = w;
```

D. Mixgaussem

```
function [label, model, llh] = mixGaussEm(X, init)
%  Code perform EM algorithm to fit Gaussian mixture model.
% Input: X: d x n data matrix & initialize k (1 x 1) number of components or label (1 x n, 1<=label(i)<=k) or model structure
% Output:
%   label: 1 x n cluster label, model: trained model structure,  llh: loglikelihood
fprintf('EM for Gaussian mixture: running ... \n');
tol = 1e-6;
maxiter = 500;
llh = -inf(1,maxiter);
R = initialization(X,init);
```

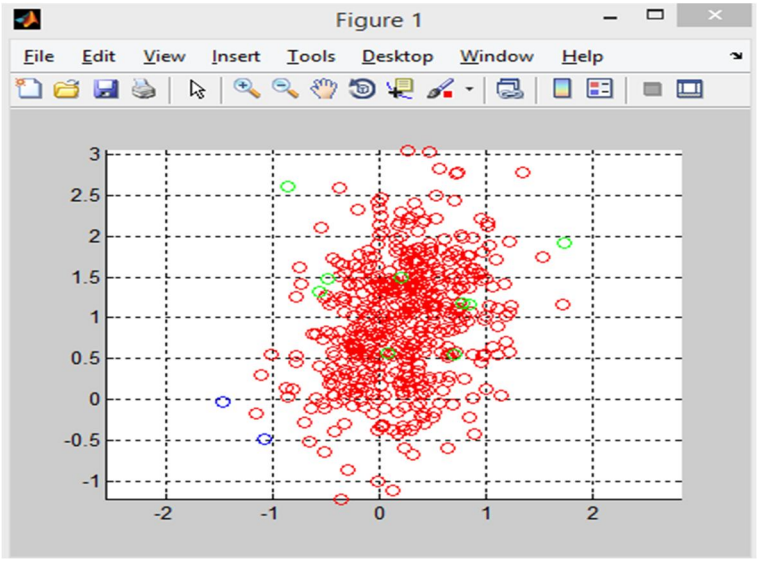

Fig: 1 Show results 1

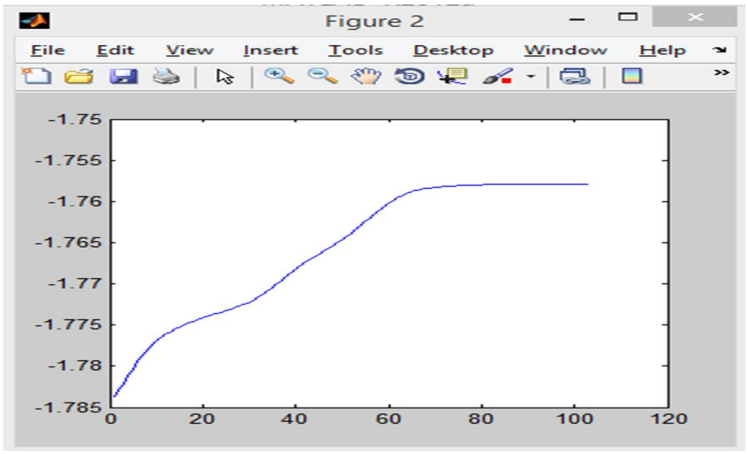

Fig: 2 Show results 1

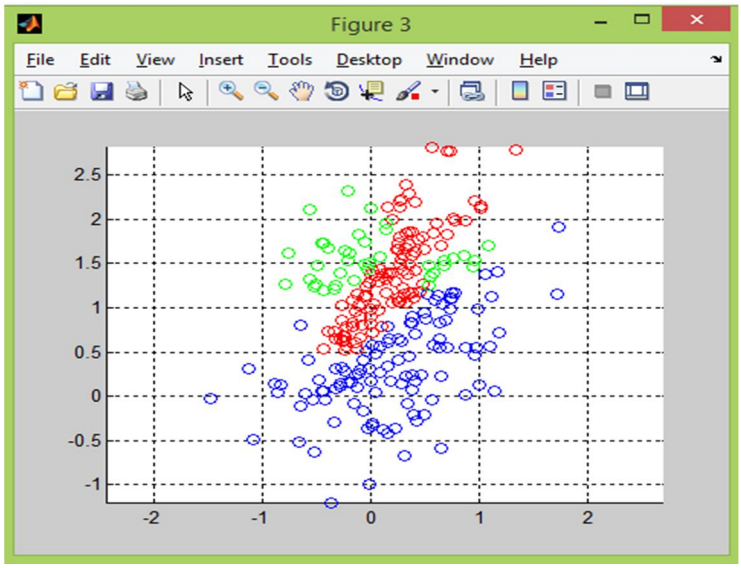

Fig: 3 Show result 3

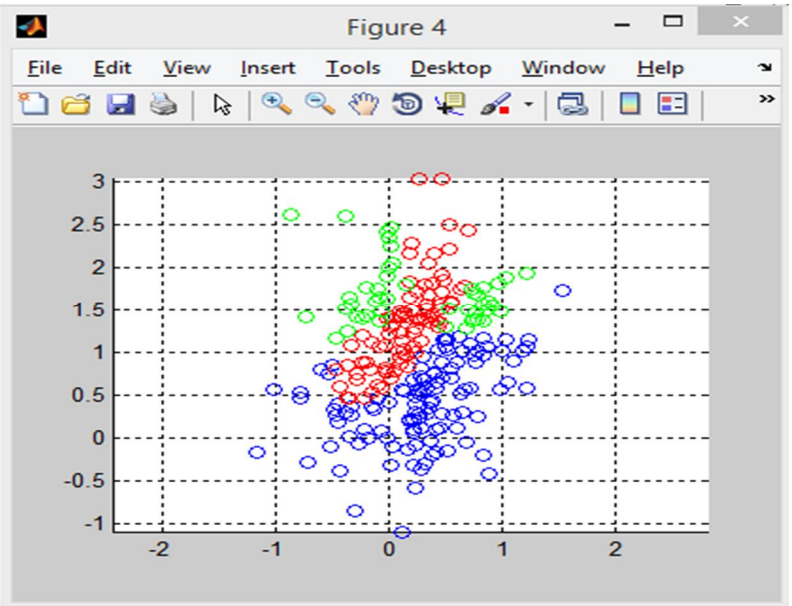

Fig: 4 Show result 4

VI. CONCLUSION

In general goal of data mining course is to extract in order from a data piece & convert this into an logical structure. In order to make wise decisions both for people & for things in IoT, data mining technologies are open to all people within IoT technologies for decision making support & system optimization. Data mining involved discovering novel interesting & potentially useful models from data & applying algorithms in extirpation of no hide information Due to increasing amount of data available online, World Wide Web had becoming one of most valuable resources for information retrievals & knowledge discoveries. Ware right solution for knowledge innovation on Web data extracted from could be used to raise performances for Web information retrievals, question answering, & Web based data warehousing.

REFERENCES

[1] David J. Woodruff (1996) "Estimation of Item Response Models Using EM Algorithm for Finite Mixtures" International Journal of Computer Science Issues, Vol. 5, Issue 2, No 3, March1996
[2] Mário A. T An(2003) "EM algorithm for wavelet-based image restoration ieee transactions on image processing" vol. 12, no. 8, august 2003.
[3] Bettina Grun (2008) "Fitting finite mixtures of linear mixed models within EM algorithm" International Journal of Scientific & Research Publications , Volume 4, Issue 3 June 2008 ISSN 2250- 3153
[4] Bhagyashree Umale (2011) "Overview of K-means & Expectation Maximization Algorithm for Document Clustering" International Conference on Quality Up-gradation in Engineering, Science & Technology (ICQUEST-2011)
[5] V.Ramesh, August-2011 "Performance Analysis of Data Mining Techniques for Placement Chance Prediction" Journal., vol. 30, no. 1, Jan. 2011, pp. 92-94, COBISS.SI-ID 6951764
[6] Bhagyashree Umale (2011) "Overview of K-means & Expectation Maximization Algorithm for Document Clustering "International Conference on Quality Up-gradation in Engineering, Science & Technology (ICQUEST-2011)
[7] Bhawna Nigam (2011) "Document Classification Using Expectation Maximization within Semi Supervised Learning" International Journal on Soft Computing (IJSC) Vol.2, No.4, November 2011
[8] David Jensen & Jennifer Neville (2012) "Data Mining in Social Networks "International Journal of Computer Science Issues, Vol. 9, Issue 2, No 3, March 2012
[9] Dr. Pragnyaban Mishra June 2012 "Survey of Data Mining Applications & Feature Scope " International Journal of Scientific & Research Publications , Volume 6, Issue 3,Appril 2012 ISSN 2250-2345
[10] Kalyani M Raval (2012)" Data Mining Techniques" International Journal of Scientific & Research Publications , Volume 3, Issue 3,September 2012 ISSN 4490- 3153
[11] Atul Kumar Pandey (2013) "DataMining Clustering Techniques in Prediction of Heart Disease using Attribute Selection Method" International Journal of Science, Engineering & Technology Research (IJSETR) Volume 2, Issue