RESEARCH ARTICLE OPEN ACCESS

Enhancing Clustering Mechanism by Implementation of EM Algorithm for Gaussian Mixture Model

RajvinderKaur*, ManinderKaur**
*Department of CSE, DIET, Regional Centre PTU

----------**********----------

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 DATA MINING APPLICATIONS

Data mining is highly useful in following domains

1. Market Analysis & Management
2. Corporate Analysis & Risk Management
3. Fraud Detection

Data mining could also be used infield of invention control, client preservation, science exploration, sports, astrology, & Internet Web Surf-Aid.

Process

The **Knowledge Discovery in Databases process** is defined in stages:

a) Selection
b) Pre-processing
c) Transformation
d) Data Mining[5]
e) Interpretation/Evaluation

It exists in various differentiations on this theme, like Cross Industry Standard Process for Data Mining which defines six phases:[5]

a) Business Understanding
b) Data Understanding
c) Data Preparation

d) Modeling
e) Evaluation
f) Deployment

Or with a simplified process like (1) data mining, (2) pre-processing (3) results validation. In Polls conducted in year 2002, 2004, & 2007 shows that CRISP-DM methodology is a leading methodology that is used by data miners. However, many people reported using CRISP-DM. Many researchers have published data mining is Azevedo& Santos have conducted a comparison of CRISP-DM & SEMMA in 2008.

III TOOLS & TECHNOLOGY USED

Matlab is known as Language of Technical Computing. It is considered as a high-level language with interactive environment. Matlab enables us to perform computationally tasks quicker as compare to other programming languages such as C, C++, and Fortran.

Matrix is a rectangular array of numbers in MATLAB environment. Its Meaning is attached to 1x1 matrices. These are scalars. In order to matrices with one row or column there are vectors. The MATLAB has different ways to store numeric & nonnumeric data. It is best to consider everything as a matrix in beginning. Operations in MATLAB have been designed to be natural.

IV ARRAY CREATION IN MATLAB

MATLAB is abbreviation for matrix laboratory. Usually programming languages work with numerical value one at a time. But MATLAB has been designed to operate on complete matrices and arrays primarily.

Functions in Matlab

MATLAB usually provides huge number of functions which are used to perform calculative tasks. These Functions are working same as subroutines or methods in programming languages other than Matlab.

Let user workspace consists of variables x and y, such as

```
x = [3 8 9];
y = [5 4 7];
```

In order invoke a function user has to enclose its input arguments in parentheses:

```
max(x);
```

And if there is presence of input arguments user has to separate them with commas:

```
max(x,y);
```

Following statement return output from a function after assigning it to a variable:

```
maxx = max(x);
```

IV PROPOSED WORK

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

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

Using Gaussian cluster in 3d space

Cluster using Gaussian model in 3d area 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.

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 Gaussian mixture model is 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.

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

V RESULT & DISCUSSION

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.

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.

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;
```

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);
```

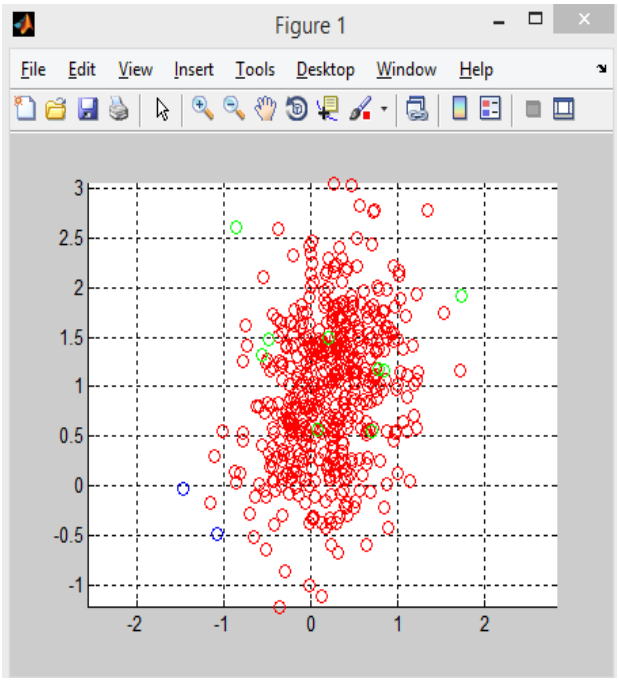

Fig: 1 Show results 1

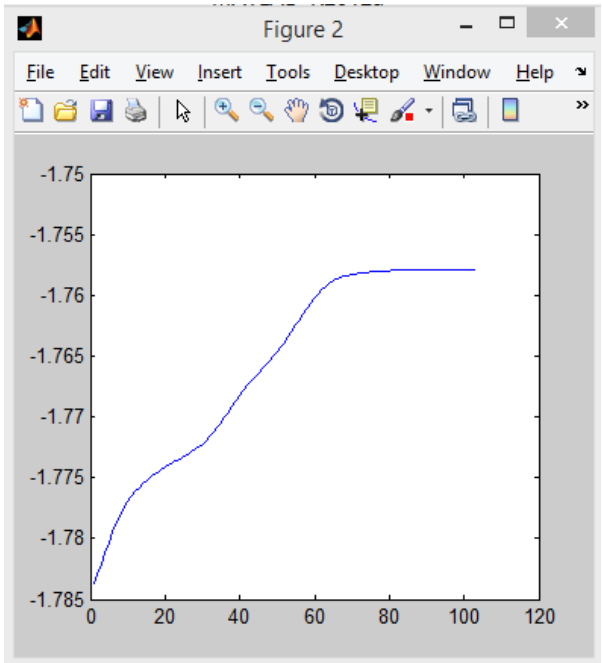

Fig: 2 Show results 1

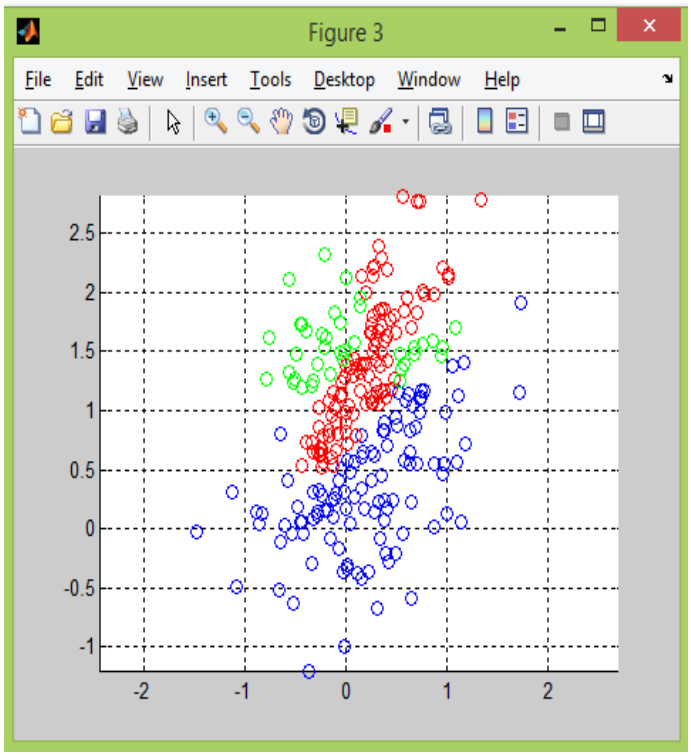

Fig: 3 Show result 3

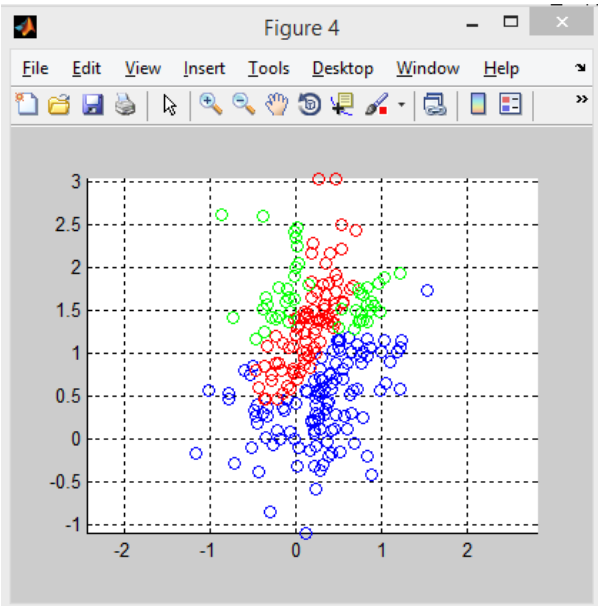

Fig: 4 Show result 4

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.

VII References

1. *Hellerstein, Joe (9 November 2008).* "Parallel Programming in Age of Big Data". *Gigaom Blog.*
2. J. Liu, N. Zhong, Y. Y. Yao, Z. W. Ras, wisdom web: new challenges for web intelligence (WI), J. Intell. Inform. Sys.,20(1): 5–9, 2003.
3. Congiusta, A. Pugliese, D. Talia, & P. Trunfio, Designing GridServices for distributed knowledge discovery, Web Intel. Agent Sys, 1(2): 91–104, 2003.
4. J. A. Hendler& E. A. Feigenbaum, Knowledge is power: semantic web vision, in N. Zhong, et al. (eds.), Web Intelligence: Research & Development, LNAI 2198, Springer, 2001, 18–29.
5. N. Zhong& J. Liu (eds.), Intelligent Technologies for Information Analysis, New York: Springer, 2004.
6. Bouckaert, Remco R.; Frank, Eibe; Hall, Mark A.; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H. (2010). "WEKA Experiences with a Java open-source project".
7. Journal of Machine Learning Research**11**: 2533–2541. original title, "Practical machine learning", was changed ... term "data mining"

was [added] primarily for marketing reasons.

8. Mena, Jesús (2011). Machine Learning Forensics for Law Enforcement, Security, & Intelligence. Boca Raton, FL: CRC Press (Taylor & Francis Group). ISBN 978-1-4398-6069-4.

9. Piatetsky-Shapiro, Gregory; Parker, Gary (2011). "Lesson: Data Mining, & Knowledge Discovery: An Introduction". Introduction to Data Mining. KD Nuggets. Retrieved 30 August 2012.

10. Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, & Algorithms. John Wiley & Sons. ISBN 0-471-22852-4.
OCLC 50055336.

11. "Microsoft Academic Search: Top conferences in data mining". Microsoft Academic Search.

12. "Google Scholar: Top publications - Data Mining & Analysis". Google Scholar.

13. Proceedings, International Conferences on Knowledge Discovery & Data Mining, ACM, New York.

14. SIGKDD Explorations, ACM, New York.