IOSR Journal of Computer Engineering (IOSR-JCE)
E-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 20, Issue 2, Ver. II (Mar. - Apr. 2018), PP 67-72
www.iosrjournals.org

EM Clustering Model for Evaluating SaaS on Cloud Computing Environment

Dhanamma Jagli, Dr. (Mrs.) Seema Purohit, Dr.N. Subhash Chandra
[1]*Research Scholar,* [2,3] *Research Guide, JNTU Hyderabad.*
Corresponding Author: Dhanamma Jagli

Abstraction: *A Cloud computing is a dynamic computing sharing resource. The cloud computing paradigm has emerged and that is transforming the IT industry at huge. In cloud computing, all resources are available as services and accessible through the Internet. Especially Software-As-A-Service (SaaS) is a service delivery model that support end users to access any software or an application as a service via the Internet without installing at local. The usage of SaaS has been increased by many users and thus leads to need to evaluate the quality of SaaS to select the best one that suits to cloud service users. In this paper, a quality model is implemented by using data mining Estimation and Maximization (EM)-clustering model for evaluating the quality of software as a service (SaaS) in the cloud computing environment.*

Keywords: *SaaS, Cloud Computing, EM-Clustering.*

Date of Submission: 21-03-2018 Date of acceptance: 06-04-2018

I. Introduction

Cloud computing is a tremendous resource sharing computing adopted by many organizations in the last decade. The main concept of cloud computing has used any resource as a service. i.e everything as a service (XaaS).they are mainly three service models: IaaS (Infrastructure as a Service), PaaS (Platform as a Service) and SaaS (Software as a Service).the SaaS is a deployment model where the service end user need not install the software on their local machine even though they use it as its local. The SaaS has been used by many vendors as well as providing by many cloud service providers because of its advantages.However, the usage of the SaaS has been increased throughout the globe in the computing world. Hence, many challenges were introduced, one of the main challenges for Cloud service users is how to select right service as per their requirements and which one would meet their expectations. In order to deal with this particular challenge, in this paper is a new quality model is proposed to evaluate the Quality of SaaS in the cloud computing environment based on the data mining clustering algorithm.

This paper Initial describes the importance of data mining EM-clustering algorithm- , then it explains about research methodology adopted for specified challenge. Finally, it explains about EM-Clustering implementation using R-Studio with SaaS quality related data.

II. Literature Review

Jerry Gao, Pushkala Pattabhiraman, Xiaoying Bai w. T. Tsai presented their research work [7] as new formal graphic models and metrics to evaluate SaaS performance and scalability features. The results shown best potential application and effectiveness of the proposed model for evaluating SaaS scalability and performance attributes only. But not on other attributes, which are also playing an important role for good quality. Zia ur Rehman proposed work [8] discussed and proposed a multi-criteria cloud service selection methodology in general. Very important parameters like reliability, trust, reputation, etc are not given importance even though they are very critical in the cloud computing environment. Qiang He, Jun Han, etc proposed work [9] is used to evaluate the attribute multi-tenancy cloud-based software applications with less scalability. It may not suitable if number of end users are increasing. Tung-Hsiang Chou and Wan-Ting Liu research work [10] presented that some of the SaaS dimensions integrated along with service dimensions of SERVEQUL to maintain the standard for customer's service. So that presented work is only benefited with very few attributes of SaaS, not applied to quality parameters.

III. Research Methodology

The scenario to be analyzed is collected from a sample data set of cloud service users, by asking rating for SaaS quality attributes on the scale of 1 to 5. From the literature review, it has been identified that in order to

evaluate SaaS, quality six attributes are playing a vital role in evaluating the quality like Availability, pay for use, data managed by the provider, scalability, reusability, and service customizability.

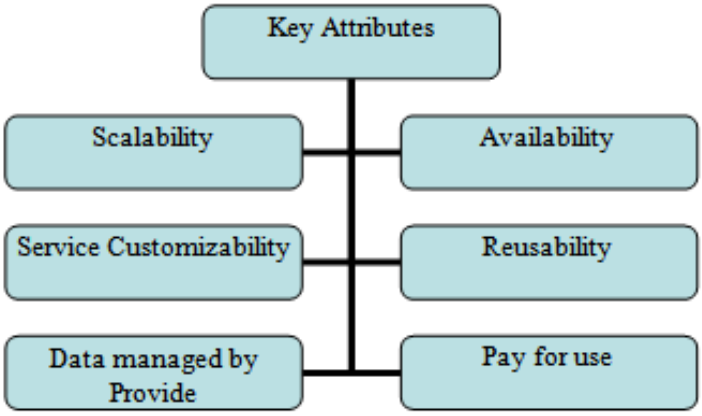

Figure 1: SaaS Kay Attributes

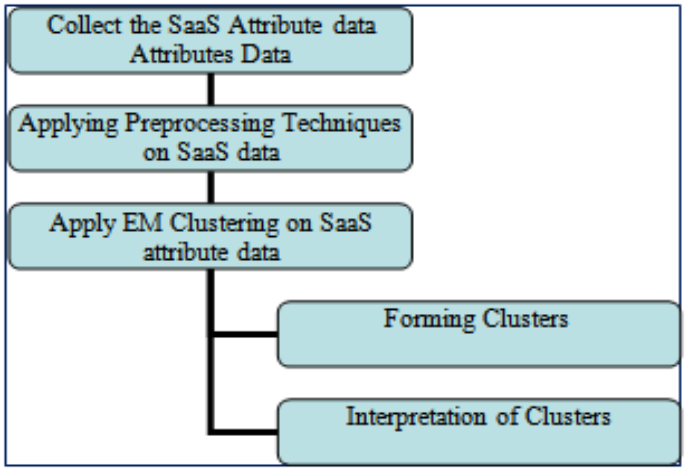

Figure 2: workflow of Proposed Model

IV. EM-Clustering

The Expectation-Maximization Algorithm

The EM algorithm is an effective iterative process to calculate the Maximum Likelihood (ML) estimate in the occurrence of absent or unseen data. In ML estimation, it wishes to estimate the model parameter(s) for which the observed data are the maximum likely. Each iteration of the EM algorithm consists of two processes: The E-step, and the M-step. In the expectation, or E-step, the missing data is estimated given the observed data and the current estimate of the model parameters. This is attained using the conditional expectation. In the M-step, the likelihood function is maximized under the assumption that the missing data are known. The estimate of the missing data from the E-step is used in lieu of the actual missing data. Convergence is assured since the algorithm is guaranteed to increase the likelihood at each iteration.

Probabilistic models, such as hidden Markov Models or Bayesian networks are a commonly used to model biological data. Much of their popularity can be attributed to the existence of efficient and robust procedures for learning parameters from observations. Often, however, the only data available for training a probabilistic model are incomplete. Missing values can occur, for example, in medical diagnosis, where patient histories generally include an only limited number of tests. The expectation maximization algorithm enables parameter estimation in probabilistic models with incomplete data.

EM-Clustering forms clusters through defining a mixture of Gaussians that fitting a given data set. For each Gaussian has an allied mean and covariance matrix. But, since spherical Gaussians is considered, a variance scalar is used in place of the covariance matrix. The previous probability for each Gaussian is the fraction of points in the cluster defined by that Gaussian. These parameters can be primed by randomly picking

means of the Gaussians, or by using the output of K-means for initial centers. The algorithm touches to a locally optimal solution by iteratively updating values for means and variances.

Advantages

In spite of the fact that EM can intermittently become immovable in a local maximum as it estimates the parameters by maximizing the log-likelihood of the experiential data. Here are three gears that make it enchanted 1.) The ability to instantaneously enhance a huge number of variables,2) the ability to discover worthy estimates for some missing information in the data.3) In the context of clustering, multi-dimensional data that provides itself to modeling by a Gaussian mixture, the ability to create both the traditional "hard" clusters and traditional "soft" clusters. Hard clusters mean a disjoint partition of the data and soft clusters means permitting for a data point to belong to two or more clusters at a time.

Table 3:Summary data

Sr.No	Pay.per.use	Availability	Reusability	Scalability	Data.managed	Customizability
Min	1.00	2	1	1	1	1
1st Qu	3.00	3	3	3	2	3
Median	4	4	4	4	4	4
Mean	3.76	3.67	3.67	3.63	3.5	3.56
3rd Qu	5	5	5	5	5	5
Max	5	5	5	5	5	5

Limitations of EM

EM-CLUSTERING is useful for several reasons: conceptual simplicity, ease of implementation, and the fact that each iteration improves. The rate of convergence on the first few steps is typically quite good, but it can become excruciatingly slow as it approaches local optima. Generally, EM-CLUSTERING works best when the fraction of missing information is small and the dimensionality of the data are not too large. EM-CLUSTERING can require many iterations, and higher dimensionality can dramatically slow down the E-step.

V. Implementation In R

Clustering is the process in which it identifying similar data have characteristics in common and are cohesive to a group. The EM-CLUSTERING algorithm is an unsupervised clustering method which does not require a training phase, based on mixture models.

The EM algorithm has become a popular tool in statistical estimation problems involving incomplete data, or in problems which can be posed in a similar form, such as mixture estimation.

R Packages

The expectation-maximization algorithm in R will use the package must. This package encompasses critical approaches intended for the implementation of the clustering algorithm, including functions for the E-step and M-step calculation. The must package similarly provides various models for EM-CLUSTERING and hierarchical clustering (HC) that is defined by the covariance structures

Executing the Algorithm

The function "em" have been used for the expectation-maximization method.
This function uses the following parameters:

- ✓ Model- name: the name of the model used;
- ✓ Data: all the collected data, which must be all numeric.
- ✓ Model parameters values: pro, mean, variance and Vinv.
- ✓ Other: less relevant parameters.

After the execution, the function will return:

- ✓ Model-name: the name of the model;
- ✓ z: a matrix whose the element in position presents the conditional probability of the i^{th} sample belongs to the k^{th} mixture component;

- ✓ Parameters: same as the input;
- ✓ others: other metrics;

Data have been collected and created as shown in the following table 1.

Table 1: SaaS Attribute data

Pay-per-use	Availability	Reusability	Scalability	Data managed by provider	S.Customizability
5	4	4	3	4	3
3	5	3	3	3	5
4	3	4	5	5	5
2	5	5	5	5	5
1	4	3	3	5	4
5	3	4	4	5	5
4	2	5	2	4	3
3	2	5	1	4	4
2	2	4	5	4	5
3	3	3	4	2	3
4	5	5	3	2	4
5	5	5	2	2	2
3	5	3	4	2	1
4	4	4	4	4	5
5	3	2	3	5	4
5	5	1	3	3	3
4	2	5	2	5	2
3	3	4	2	4	5
5	3	3	3	3	5
5	4	2	3	2	3
5	4	4	5	2	3
4	5	3	5	2	2
4	5	4	4	3	1

The data have been analysed in the R studio and find the summary as shown in the below.A clustering analysis is performed with more details, applied to a scenario composed of only two attribute scalability and availability. The clustering technique is executed based on the default MCLUST. After execution of the algorithm by limiting to Scalability and availability. The plot of scalability and availability for class 1 points and class 2 points shown in the below figures respectively.

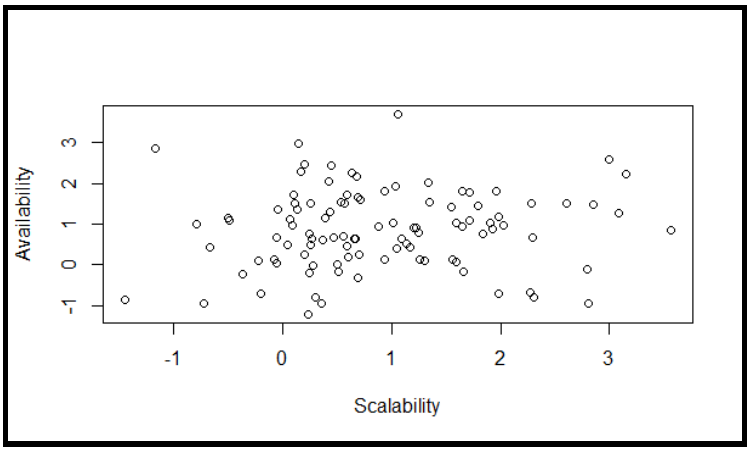

Figure 3:Normal plot for scalability-Availability

```
#SaaS data loaded
#installed mclust Package
# get Scalability and avaialbility normal distributed points for x axis and y axis respectively with mean=1 and 5 and std=1
>library(mclust) #load mclust library
>Scalability=rnorm (n=data [, 4], mean=1, SD=1)
>Scalability1=rnorm (n=data [, 4], mean=5, sd=1)
>Availability=rnorm (n=data [, 2], mean=1, sd=1)
>Availability1=rnorm (n=data [, 2], mean=5, sd=1)
# get the axis x range
# get the axis y range
> rx = range (Scalability, Scalability1)
> rx
[1] -1.439406 7.868007
> ry = range (Availability, Availability1)
> ry
[1] -1.224036 7.529909
# plot the first class points and 2nd class points respectively
> Plot (Scalability, Availability, xlim=rx, ylim=ry)
>plot (Scalability1, Availability1, xlim=rx, ylim=ry)
# create a dataframe matrix
> mix = matrix (nrow=101, ncol=2)
# insert first class i.e Scalability points into the matrix
> mix1= c (Scalability, Scalability1)
> mix1
[1] 3.08249472 3.15266415
# insert second class i.e Availability points into the matrix
> mix2= c (Availability, Availability1)
> mix2
[1] 1.280351497 2.219891626
1.495607342 0.259460489
1.103867933 1.77818513
> mixclust = Mclust (mix1)
> mixclust
'Mclust' model object:
best model: univariate, equal variance (E) with 2 components
> mixclust1 = Mclust(mix2)
> mixclust1
'Mclust' model object:
```

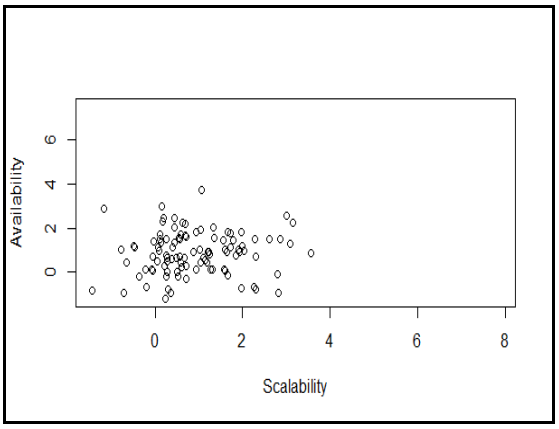

Figure 4:plot for scalability-Availability with Mean value 1

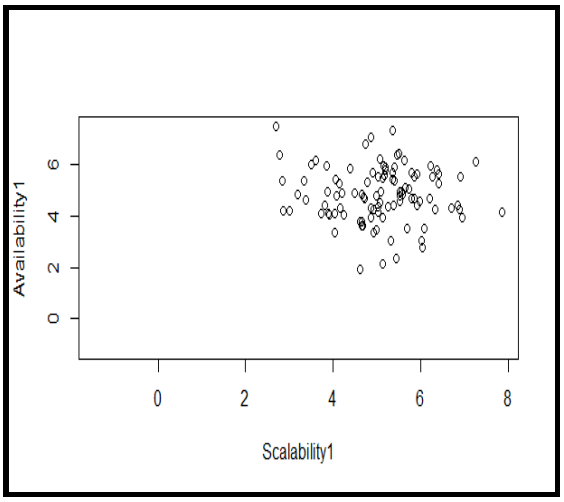

Figure 5:plot for scalability-Availability with Mean value 5

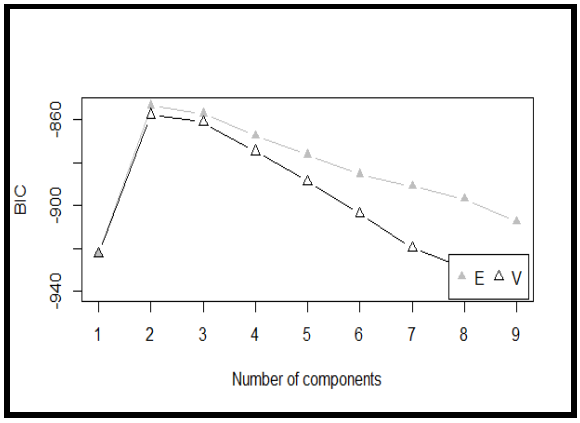

Figure 6: plot with BIC value

Conclusion

In this paper, the quality of any SaaS product with identified quality attributes have been proposed by using model based EM-Clustering algorithm. It is also implemented in the R Studio statical tool. And observed the results of scalability and availability with various plots. EM-Clustering id good for incomplete data sets, being SaaS related quality attributes data some values are not available so that EM-Clustering have been chosen for this approach. In the future, it has been proposed to analyse all other SaaS quality attributes.

References

[1] N. Alldrin, A. Smith, and D. Turnbull, "Clustering with EM and K-means," *Univ. San Diego, California*, pp. 261–95, 2003.

[2] A. Kak and A. Kak, "Expectation Maximization Tutorial Expectation-Maximization Algorithm for Clustering Multidimensional Numerical Data Expectation Maximization Tutorial," vol. 2012, no. November, 2014.

[3] T. M. Mitchell, "Expectation Maximization, and Learning from Partly Unobserved Data (part 2)," *Mach. Learn.*, vol. 15, no. April, 2005.

[4] G. Distribution and I. Criteria, "Package 'EMCluster,'" 2015.

[5] S. Borman, "The Expectation Maximization Algorithm A short tutorial," *Submitt. Publ.*, vol. 25, no. x, pp. 1–9, 2009.

[6] E. M. Em and P. Abbeel, "Maximum Likelihood (ML), Expectation Maximization (EM)," no. Ml, pp. 1–23.

[7] H. Bourlard, Y. Konig, and N. Morgan, "REMAP: recursive estimation and maximization of a posteriori probabilities in connectionist speech recognition." *Eurospeech*, 1995.

[8] S. Borman, "The Expectation Maximization Algorithm A short tutorial," *Submitt. Publ.*, vol. 25, no. x, pp. 1–9, 2009.

[9] C. B. Do and S. Batzoglou, "What is the expectation maximization algorithm," *Nat. Biotechnol.*, vol. 26, no. 8, pp. 897–899, 2008.

[10] A. K. Singh, "The EM Algorithm," *Technometrics*, vol. 48, pp. 1–4, 2005.

[11] Dhanamma Jagli, Seema. Purohit, and N. S. Chandra, "SAASQUAL: A Quality Model For Evaluating SaaS on The Cloud Computing Environment," 2015.

[12] A. gupta Jagli[1], Mrs Dhanamma, "17.Clustering Model for Evaluating SaaS," 2013.

[13] Dhanamma Jagli, Dr. Sunita Mahajan Dr. subhash Chandra, "CBC Approach for Evaluating Potential SaaS on the Cloud," *vesit.edu*, vol. 2, pp. 43–49, 2014.

[14] J. Y. Lee, J. W. Lee, D. W. Cheun, and S. D. Kim, "A Quality Model for Evaluating Software-as-a-Service in Cloud Computing," in *Software Engineering Research, Management and Applications, 2009. SERA '09. 7th ACIS International Conference on*, 2009, pp. 261–266.

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

Dhanamma Jagli "EM Clustering Model for Evaluating SaaS on Cloud Computing Environment "IOSR Journal of Computer Engineering (IOSR-JCE) 20.2 (2018): 67-72.