International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 05, May - 2017

Measuring Performance of Selected Algorithms used for Classification and Regression when Applied Against a Standard Dataset

Vinay S Bharadwaj
M.Tech Student, Dept of ISE,
MSRIT, Bangalore, India

Mrs. Sunitha R S
Assistant Professor, Dept of ISE,
MSRIT, Bangalore, India

Mr. Shashidhara H S
Associate Professor, Dept of ISE,
MSRIT, Bangalore, India

***Abstract*---Data mining algorithms are often used to extract useful information from datasets, which give us deeper insights into what we are focussing on to get from the data. This paper aims at measuring the performance of the few selected algorithms namely, Bayesian Generalized Linear Model, Generalized Linear Model, k-Nearest Neighbours and Partial Least Squares. The various performance parameters measured include sensitivity, specificity, root mean squared error (RMSE), R Squared etc. The dataset used would be from the machine learning library present in R studio namely, mlbench library and the dataset being Pima Indians Diabetes.**

Keywords---- BGLM, GLM, kNN, PLS, mlbench, caret, e1071, train Control, ROC, Kappa

I. INTRODUCTION

With hundreds of data mining algorithms available in present times, it is often confusing to decide upon which particular algorithm needs to be used for a particular data model. The focus here is on algorithms used for classification and regression purpose targeted towards supervised learning. Before we begin our investigation of algorithm performance we need to understand the basics of what is classification and regression and why are they used. Classification in machine learning essentially means that to place a new observation in a set of categories already pre-defined based on the training dataset [3]. So in classification we actually group the output variables into different corresponding classes. Regression in machine learning means that we actually predict the output values based on the training data values. By using these algorithms we can extract huge amount of meaningful data by applying it against the dataset. The model used here also emphasizes on the use of regression techniques for training the data. The four algorithms under consideration here are used for both classification and regression purpose.

II. OVERVIEW OF ALGORITHMS UNDER CONSIDERATION

It is essential to understand in brief about the algorithms under consideration in this paper. Let us look into the basics of these algorithms [4] as to what are they, why are they used and their importance in machine learning etc.

In this paper we have divided the sections into the following: Initially we start off with the overview of the 4 algorithms under consideration, then we discuss the implementation details followed by result analysis and conclusion regarding which of the 4 algorithm best suits the dataset under investigation. We also include the future scope of the paper which is to experiment with various other algorithms available.

A. Bayesian Generalized Linear Model (BGLM)

The approach here is that we specify a conditional distribution and the data is additionally supplemented with prior probability distribution. The prior can take the form specified based on data. There might not be posterior probabilities for all the prior probabilities. Hence we only take the conjugate probabilities and compute their posterior probabilities. We only consider the posterior probabilities for this model. This model is basically used to avoid over fitting when considered for application to large datasets. Finally we do model evidence which clearly tells us how well the model has predicted the behaviour. Generally Laplace equations are used for getting the final results.

B. Generalized Linear Model

The approach here is that the model consists of three components namely, a probability distribution from the exponential family, a linear predictor and a link function. We can have any distribution like gamma, poisson etc. The outcome is basically reliant on the dependent input which also relies on mean of the distribution. The distribution is basically an error distribution model. Here the relationship between a response variable and one or more predictors. The outcome is measured using maximum likelihood or using Bayesian techniques etc.

C. k-Nearest Neighbours

The approach here is that depending on the weights of its neighbours and the Euclidean distance between the training objects, the classification is done. The choice of input parameter k is dependent on the data. The output result depends heavily on the input clusters and the boundaries between them. This algorithm is used both for classification and regression. In regression we determine the inverse distance to compute the k nearest multivariate neighbours.

D. Partial Least Squares

The approach here is that we try to project both the predicted output variables and the observable variables into a new space to find the linear regression between the two. More the input variables in the dataset more accurate will

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 05, May - 2017

be the predicted values. We try to maximise the covariance between the input values and the predicted values to get better results.

III.IMPLEMENTATION DETAILS

Usage of one of the foremost data mining specific open source language namely, R and the IDE used is R Studio. Mainly 4 libraries in R are used namely, mlbench, caret, e1071 and R Weka. The mlbench library consists of built in datasets and this paper uses Pima Indians Diabetes. The dataset consists of 9 attribues. We use the repeated cv method with 10 fold to train the dataset. The user, system and elapsed time is calculated which is nothing but the execution time for each of the four algorithms. Out of the total 9 attributes present in the dataset, we use some of the attributes to estimate the performance of algorithms one at a time. We resample the obtained results and display the summary of results which give us the performance parameters like sensitivity, specificity, root mean squared error (RMSE), R Squared etc. of the algorithms. Following is a glimpse about the dataset under consideration. [11]

Name of the Database:- Pima Indians Diabetes (from mlbench library)

Number of Attributes Present:- 9

Number of Observations:- 768

Attribute Information:

1. Number of times pregnant

2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test

3. Diastolic blood pressure (mm Hg)

4. Triceps skin fold thickness (mm)

5. 2-Hour serum insulin (mu U/ml)

6. Body mass index (weight in kg/(height in m)^2)

7. Diabetes pedigree function

8. Age (years)

9. Class variable (0 or 1)

IV. RESULT ANALYSIS

From the derived results from the console output in R studio, we analyze what obtained result set means, and how to interpret the results obtained. For result analysis we use both tabulated results and visual graph plots for better understand ability.

Following are the results obtained by running the program for getting accuracy.

Metric:-Accuracy

Models: BAYESGLM, GLM, KNN, PLS

Number of resamples: 30

Accuracy:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
BAYESGLM	0.7013	0.7403	0.7778	0.7730	0.8019	0.8442	0
GLM	0.6883	0.7403	0.7792	0.7778	0.8182	0.8961	0
KNN	0.6364	0.6916	0.7386	0.7370	0.7785	0.8442	0
PLS	0.6883	0.7273	0.7532	0.7609	0.7915	0.8442	0

Table 1) Accuracy Result Tabulation

Kappa:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
BAYESGLM	0.3025	0.4003	0.4797	0.4718	0.5488	0.6393	0
GLM	0.2787	0.4040	0.4913	0.4865	0.5742	0.7552	0
KNN	0.1876	0.3076	0.4042	0.3968	0.5003	0.6393	0
PLS	0.2655	0.3731	0.4278	0.4409	0.5243	0.6457	0

Table 2) Kappa Result Tabulation

We have used 5 different types of graphs for presenting the results. The 5 graphs used are box and whisker plots, density plots, dot plots, parallel plots and pair wise scatter plots. Following are the graph plots available for visual interpretation of the results for accuracy metric:

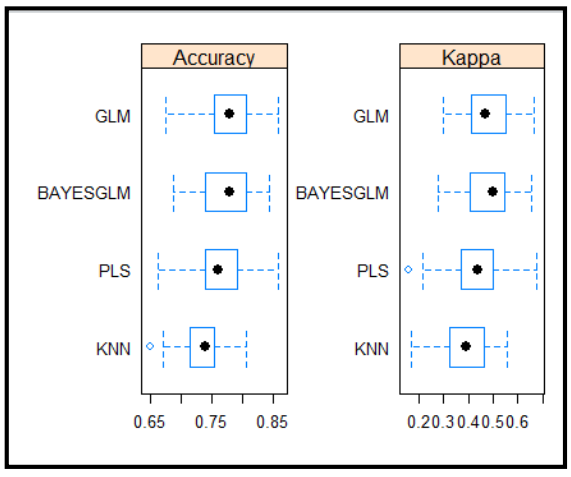

Fig 1) Box and Whisker plot for Accuracy metric

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 05, May - 2017

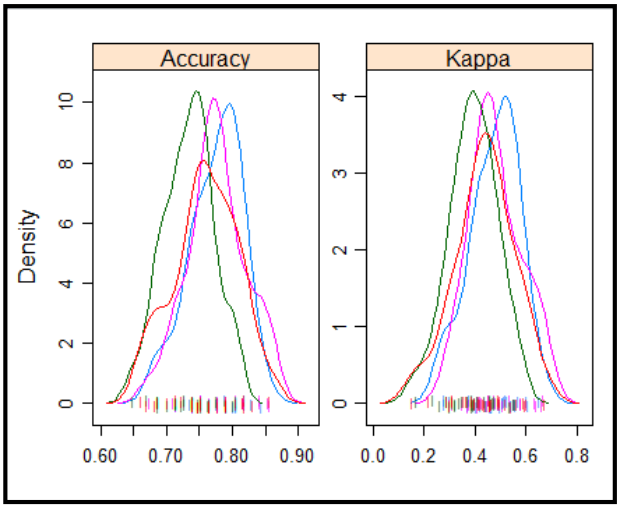

Fig 2) Density plot for accuracy metric

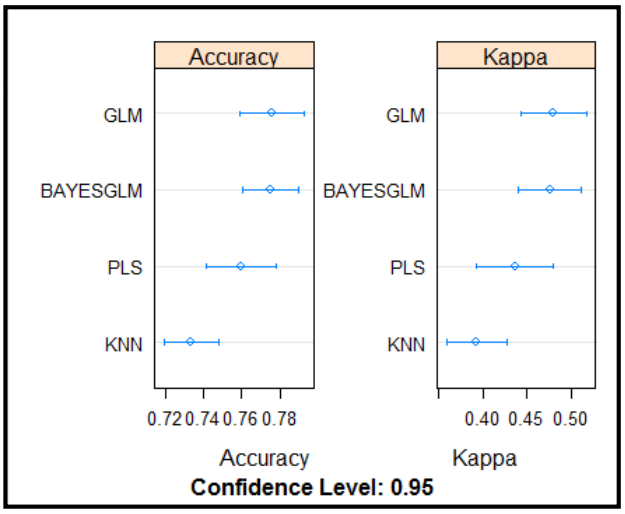

Fig 3) Dot plot for Accuracy metric

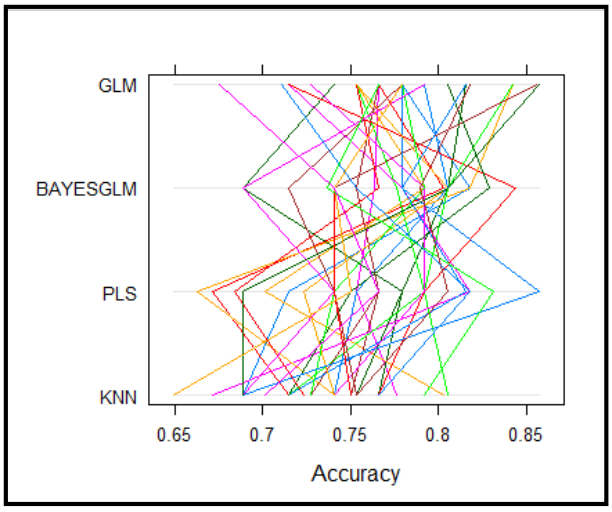

Fig 4) Parallel plot for Accuracy metric

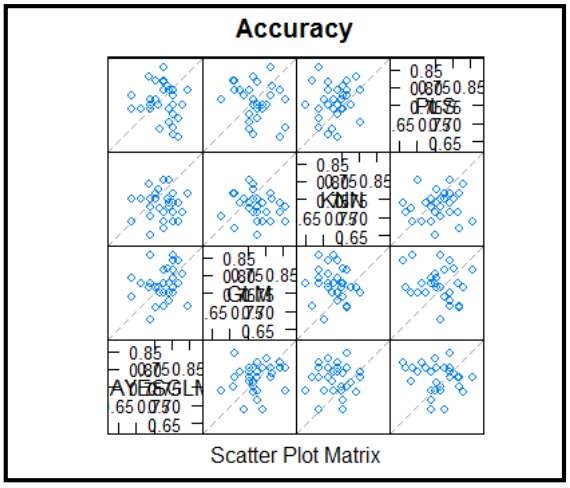

Fig 5) Pair wise scatter plot for Accuracy metric

Measuring execution time as a performance parameter is also very important when measuring the accuracy metric while training the model. We use the average of User, System and Elapsed CPU cycles time for measuring the average execution time.

	User	System	Elapsed
BAYESGLM	1.13	0.03	1.15
GLM	0.94	0.00	0.94
KNN	1.12	0.02	1.14
PLS	1.11	0.00	1.11

Table 3) Execution time for Accuracy metric measurement

Following are the results obtained by running the program for getting ROC, Sensitivity, Specificity:

ROC:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
BAYESGLM	0.6756	0.7989	0.8319	0.8310	0.8639	0.9504	0
GLM	0.7496	0.8026	0.8245	0.8328	0.8665	0.9081	0
KNN	0.6656	0.7341	0.7804	0.7775	0.8269	0.8765	0
PLS	0.6908	0.7822	0.8078	0.8103	0.8339	0.9148	0

Table 4) ROC Result Tabulation

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 05, May - 2017

Sensitivity:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
BAYESGLM	0.72	0.860	0.89	0.8833	0.92	0.94	0
GLM	0.78	0.860	0.88	0.8813	0.90	0.98	0
KNN	0.72	0.800	0.84	0.8380	0.88	0.98	0
PLS	0.72	0.865	0.88	0.8807	0.92	0.98	0

Table 5) Sensitivity Result Tabulation

Specificity:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
BAYESGLM	0.3333	0.5185	0.5556	0.5731	0.6296	0.7778	0
GLM	0.4074	0.5185	0.5556	0.5748	0.6296	0.7692	0
KNN	0.3333	0.4815	0.5556	0.5511	0.6296	0.7308	0
PLS	0.2963	0.4815	0.5185	0.5326	0.6097	0.7407	0

.Table 6) Specificity Result Tabulation

Visual Graph plots for better result analysis for ROC metric are as follows:

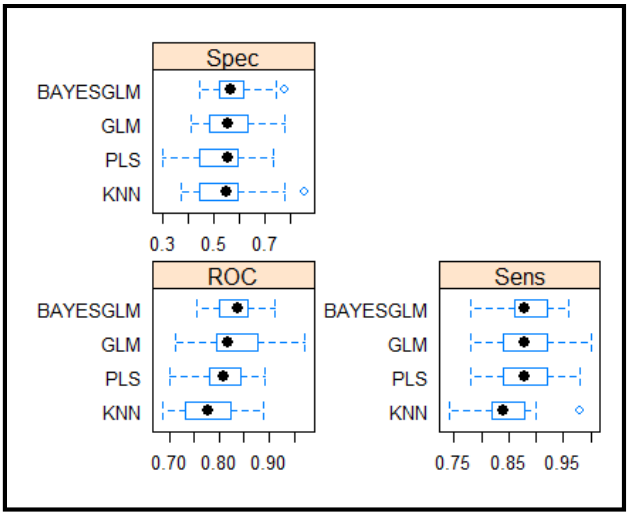

Fig 6) Box and Whisker plot for ROC metric

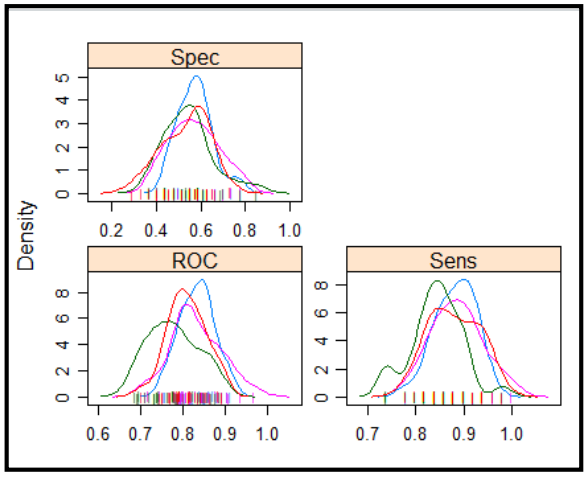

Fig 7) Density plot for ROC metric

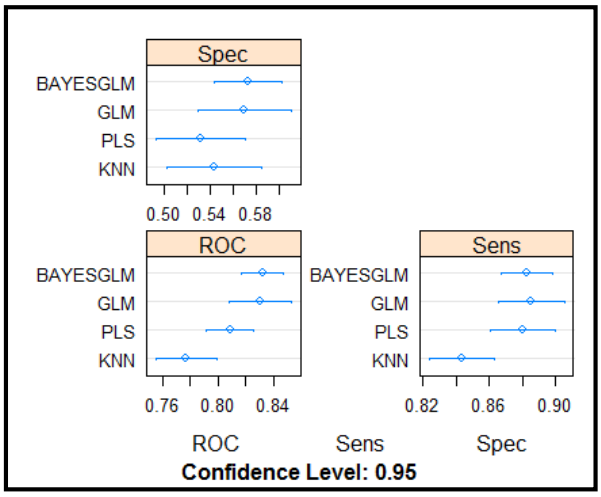

Fig 8) Dot plot for ROC metric

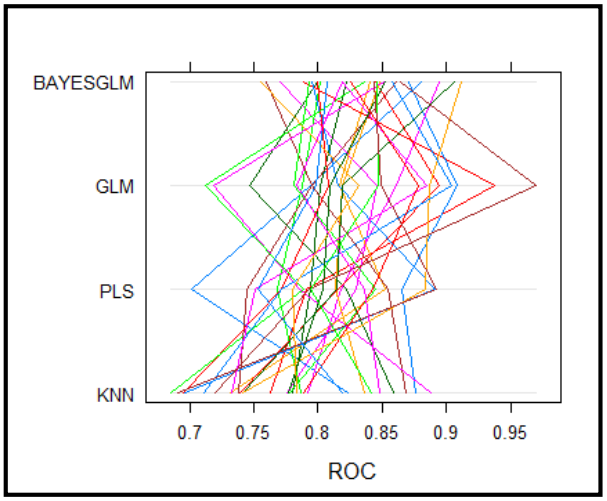

Fig 9) Parallel plot for ROC metric

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 05, May - 2017

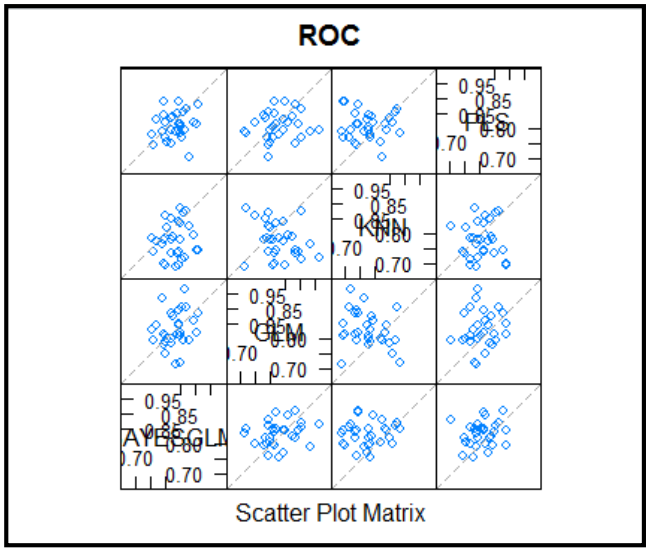

Fig 10) Pair wise scatter plot for ROC metric

Measuring the execution time as a performance parameter while using the ROC as a metric during training the model.

	User	System	Elapsed
BAYESGLM	1.28	0.00	1.33
GLM	0.94	0.00	0.97
KNN	1.21	0.00	1.20
PLS	1.12	0.00	1.13

Table 7) Execution time for ROC metric measurement:

V. CONCLUSION

From examining the result set we come to the following conclusion regarding which algorithm is better out of the 4 chosen algorithms and gives better performance when applied against Pima Indians Diabetes dataset. As we can infer from Table 3 and Table 7 the average execution time are as follows: GLM < PLS < KNN < BAYESGLM in the increasing order. We consider the mean values from the tables for comparison purpose. The mean Accuracy of the algorithms from Table 1are as follows: GLM > BAYESGLM > PLS > KNN in decreasing order. The mean kappa values from Table 2 are as follows: GLM > BAYESGLM > PLS > KNN in decreasing order. We can infer that for both Accuracy and Kappa statistics parameters the order is the same. The mean ROC of the algorithms from Table 4 are as follows: GLM > BAYESGLM > PLS > KNN in decreasing order. The mean Sensitivity of the algorithms from Table 5are as follows: BAYESGLM > GLM > PLS > KNN in decreasing order. The mean Specificity of the algorithms from Table 6 are as follows: BAYESGLM > GLM > KNN > PLS in decreasing order. In this paper we are not going to rank these 4 algorithms, instead we only say which is the best among the 4 algorithms under consideration. Depending on the tabulated results and comparison statistics we conclude that for Pima Indians Diabetes dataset with diabetes as the parameter, the best suited algorithm among the 4 is Generalised Linear Model (GLM) algorithm.

VI.FUTURE SCOPE

The paper presents only 4 algorithms particularly used for classification and regression purposes. Similarly many other algorithms can be used to measure comparative performance of algorithms when applied against standard datasets from respected data sources. Effort can also be made with regards to other types of machine learning algorithms belonging to categories of cluster and predictive algorithms. Usage of open source tools like WEKA or Rapid Mine in place of R studio can be made use for getting the results rapidly and avoid coding to save time although not much flexibility is available without programming.

REFERENCES

[1] Barath Narayanan Narayanan, Ouboti Djaneye-Boundjou and Temesguen M. Kebede, “Performance Analysis of Machine Learning and Pattern Recognition Algorithms for Malware Classification”, 2016

[2] M.H. Hesamian, S. Mashohor, M.I. Saripan,WA Wan Adnan, B.Hesamian, M.M.Hooshyari, “Performance of Various Training Algorithms on Scene Illumination Classification”, 2015 IEEE Student Conference on Research and Development (SCOReD)

[3] Alisa A. Vorobeva, “Examining the Performance of Classification Algorithms for Imbalanced Data Sets in Web Author Identification”, proceeding of the 18th conference of fruct association

[4] A.Swarupa Rani and S.Jyothi, “Performance analysis of Classification Algorithms under Different Datasets”, 2016 International Conference on Computing for Sustainable Global Development (INDIACom)

[5] Neelam Singhal and Mohd. Ashraf, “Performance Enhancement of Classification Scheme in Data Mining using Hybrid Algorithm”, International Conference on Computing, Communication and Automation (ICCCA2015)

[6] K. Dharmaraajan and M.A. Dorairangaswamy, ”Analysis of FP-Growth and Apriori Algorithms on Pattern Discovery from Weblog Data”, 2016 IEEE International Conference on Advances in Computer Applications (ICACA)

[7] Shubhangi D. Patill, Dr. Ratnadeep R. Deshmukh and D.K. Kirange,” Adaptive Apriori Algorithm for Frequent Itemset Mining”, Proceedings of the SMART -2016, IEEE Conference ID: 39669, 5th International Conference on System Modeling & Advancement in Research Trends

[8] Da-Qi Ren, Da Zheng, Guowei Huang, Shujie Zhang, Zane Wei,” Parallel Set Determination and K-means Clustering for Data Mining on Telecommunication Networks”, 2013 IEEE International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing

[9] Amirah Mohamed Shahiria, Wahidah Husaina, Nur’aini Abdul Rashida, "A Review on Predicting Student’s Performance using Data Mining Techniques", The Third Information Systems International Conference, 2015

[10] Muhammad Arif, Khubaib Amjad Alam, Mehdi Hussain, "Application of Data Mining Using Artificial Neural Network: Survey", International Journal of Database Theory and Application Vol.8, No.1 (2015), pp.245-270

[11] Mr. Chintan Shah, Dr. Anjali G. Jivani, "Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction", 4th ICCCNT

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 05, May - 2017

[12] Brijesh Kumar Bhardwaj, Saurabh Pal, "Data Mining: A prediction for performance improvement using classification", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4

[13] Da-Qi Ren, Da Zheng, Guowei Huang, Shujie Zhang, Zane Wei, "Parallel Set Determination and K-means Clustering for Data Mining on Telecommunication Networks", 2013 IEEE International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing

[14] Gurpreet Singh, Jaskaranjit Kaur, MD. Yusuf Mulge, "Performance Evaluation of Enhanced Hierarchical and Partitioning Based Clustering Algorithm (EPBCA) in Data Mining", 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)