ISSN (e): 2250 – 3005 || Volume, 09 || Issue, 3 || March– 2019 ||
International Journal of Computational Engineering Research (IJCER)

An Analysis of Missing Data Imputation Methods on Microarray Gene Expression Data

K Ishthaq Ahamed, Dr.Shaheda Akthar ,
Research Scholar, Dept. of Computer Science and Engineering, Acharya Nagarjuna University, Guntur.
Registrar FAC Dr.Abdul Haq Urdu University, Kurnool.
Corresponding Author: K Ishthaq Ahamed

ABSTRACT: *Micro array Data usually used in the gene expression data analysis. Before obtaining the micro array data, it undergoes many transformations. Micro array gene expression data occupies huge amount of memory. Micro array data was analyzed statistically and mathematically to obtain the correlation among different genes. Generally micro array gene expression data consists of missing values. These missing entries in the micro array datasets were introduced by many reasons. Dealing with micro array gene expression data with missing entries is a challenging issue. Analysis of gene expression data could not be successful if the micro array data consisting of missing entries. So before any analysis is done missing entries are imputed successfully. In this paper we made the comparative study of random Forest , mean and KNN (K nearest neighbor) Imputation methods. , based on real time datasets. The performance of this technique is analyzed by varying the percentage of missing entries in the original datasets.*

Date of Submission: 28-03-2019 Date of acceptance:08-04-2019

I. INTRODUCTION

Micro array data has been generated from gene expression experiments under different conditions [1][2]. Generally this gene experiment consists of DNA sample which has to under goes series of steps and finally obtains the micro array data. Generated microarray Data further undergoes analysis process. Different types of analysis like clustering to get genes with similar properties, classification to identify the categories of genes which falls under different categories depending on their properties and also useful to identify the differential genes for human tumors to yeast sporulation [3][4]. The micro data was generally a large matrix with genes as rows and attributes represents different experimental conditions.

MCAR(Missing Completely At Random) is data missing categorization in which probability of predicting the missing data does not depends on both response and non-response data, MAR(Missing At Random) is the probability of missing and predicting the missing values depending on the response (Observed) data, not on the non-response data, and MNR(Missing not At Random) in which the probability of predicting the missing values depends on the non-response(missing) pattern of data, not on the response (Observed data) data.

Analysis which made on the micro array data with missing entries misses the accuracies. So it is observed that lots of difference in the accuracies with complete micro array data and data with missing values. These missing entries due to the negligence during microarray experiment. These negligence include dust particles stayed on the glass plate after washing the glass plate. While micro array data is used for analysis the missing entries are carefully placed with the help of missing data estimation methods. Several missing data estimation are proposed by many people in the history.

$$MicroArray\ data = D_{ij} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & . & . & a_{1n} \\ a_{21} & a_{22} & a_{23} & . & . & a_{2n} \\ a_{31} & a_{32} & ? & . & . & a_{3n} \\ . & . & . & . & ? & . \\ . & . & . & . & . & . \\ a_{m1} & ? & ? & . & . & a_{mn} \end{pmatrix} \qquad (1)$$

D_{ij} is the micro array data with m genes and n number of samples. From the equation (1) a_{ij} represents the data value belongs to the i^{th} gene with j^{th} sample. From the data matrix it is observed that some cells have entries and some does not have the values in it. Total data values can be divided as observed values ($a_{ij} = O_{ij}$ and unobserved values (Missing entries)(a_{ij} =?) which was represented as ? in the matrix. From data matrix D_{ij} through analysis is made to identify, what values can be placed in the missing entries replacing the NA with values. Olga Troyanskaya1 and Michael Cantor [6] has made the comparative study of KNN and SVD

Imputation method, found SVD imputation is more robust compared with KNN imputation method. Dempster et al. (1997)[7] proposed expectation maximization with maximum likelihood estimation for missing data estimation. [8][9][10] Some authors studied that multiple imputation algorithms require huge computations while performing the parallel computations. [11][12][13]Several authors used artificial intelligence and evolutionary computation in missing data estimation. [14][15][16] Classical techniques has been used by many authors like case wise deletion, pair wise deletion and mean mode substitution. In this paper we made the comparative study of mean, random Forest and KNN imputation algorithms on original microarray data by varying the intensity of missing entries in the data sets. Estimated values are compared with original value and the error so obtained through root mean square value.

II. MEAN, RANDOM-FOREST AND KNN IMPUTATION METHODOLOGY

2.1. Mean imputation:

In this missing values are replaced by the mean of the observed values. It is useful for univariate but less useful in multivariate analysis. [17][18]This method of imputation more biased in nature.

2.2. Random-Forest Imputation:

Random Forest Imputation: Usually works on mixed types data sets either be categorical or continuous data. Basic building block is decision tree. The fundamental idea behind a random forest is to combine many decision trees into a single model. Individually, predictions made by decision trees (or humans) may not be accurate, but combined together; the predictions will be closer to the mark on average. It can handle non linear relation and complex structure in the data sets. This algorithm is based on random forest[19] [Breiman 2001]Main advantage of this algorithm is it can run in parallel to save computational time.

2.3. KNN (K nearest neighbor) Imputation [6]

In this imputation substituting the missing entries which have very near and similar property as observed value. Similarity between the two entries can be determined by distance and other methods. Suppose if any gene is missing value at experiment 1 then it estimate the nearest values from the K other neighboring genes. At first find the K closest genes to the missing gene with help of Pearson correlation, Euclidean distance, variance minimization methods.

III. DATASET USED IN THE EXPERIMENT [5]

Prostate cancer micro array dataset has used for this experiment. The dataset consists of 2135 genes and 102 samples. For the computational simplicity we reduced the number of sample from 102 to 24 and genes remains same. Fig 1 shows the cluster plot of missing values, which shows the missing patterns from the samples as well as genes.

IV. PERFORMANCE AND DISCUSSION

4.1 Root mean square error (RMSE)

The Root Mean Square Error (**RMSE**) (also called the root mean square deviation, RMSD) is a frequently used measure of the difference between values predicted by a model and the values actually observed from the environment that is being modeled. These individual differences are also called residuals, and the RMSE serves to aggregate them into a single measure of predictive power.

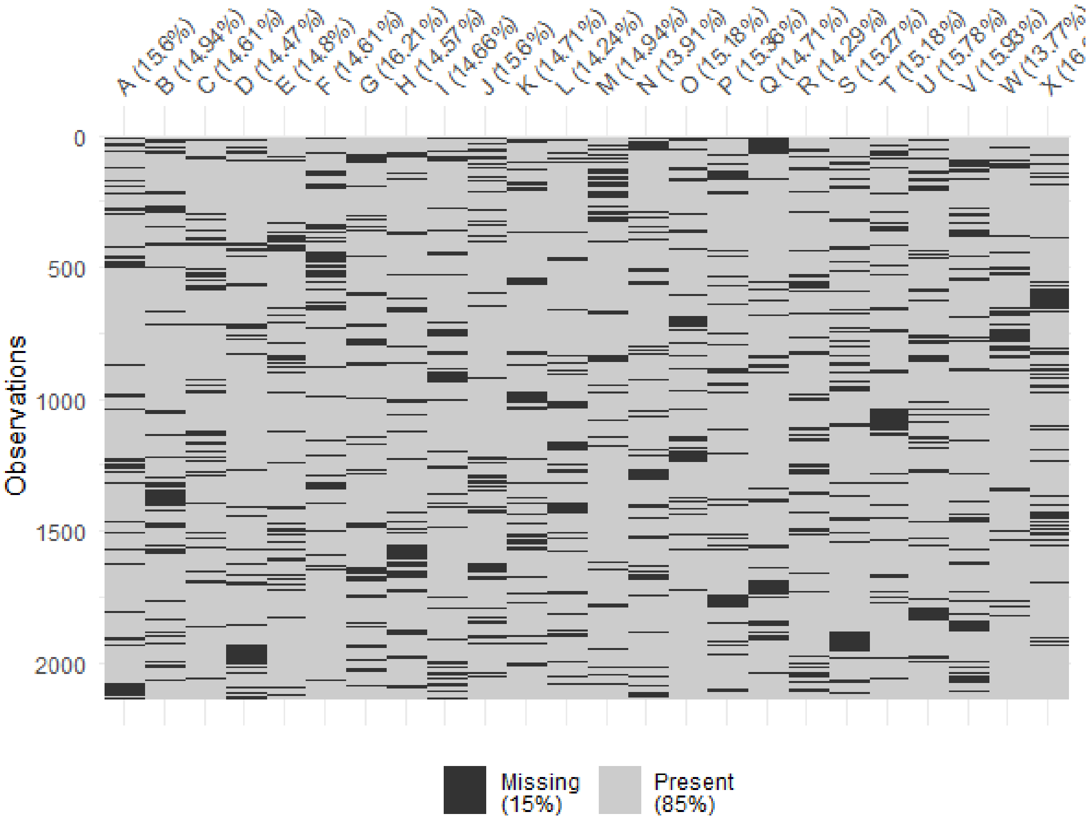

Fig 1. Missing data Clustering plot

The RMSE of a model prediction with respect to the estimated variable X_{model} is defined as the square root of the mean squared error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_{obs,i} - X_{model,i})^2}{n}} \qquad (2)$$

where X_{obs} is observed values and X_{model} is modelled values at time/place i.

4.2 Normalized root mean square error (NRMSE)

Non-dimensional forms of the RMSE are useful because often one wants to compare RMSE with different units. There are two approaches: normalize the RMSE to the range of the observed data, or normalize to the mean of the observed data.

$$NRMSE = \frac{RMSE}{X_{obs,\max} - X_{obs,\min}} \qquad (3)$$

S.No	% of Missing Data	miss-Forest Imputation NRMSE	Mean Imputation NRMSE	KNN Imputation NRMSE K=5	KNN Imputation NRMSE K=10	KNN Imputation NRMSE K=20
1	2	0.2556	0.9503	1.977	1.963	1.988
2	5	0.2629	0.9495	1.951	1.944	1.939
3	10	0.2731	0.9578	1.942	1.968	1.966
4	15	0.2747	0.9472	1.967	1.962	1.976
5	20	0.2773	0.9441	1.949	1.965	1.936

Table 1. Performance comparisons for different imputation methods.

From the Table-1 column 2 shows the percentages of missing values during the experiment. Each experiment the missing values are varies and performance of each imputation method is recorded. From Table-1 shows in each case missing values in the dataset, miss-Forest Imputation with low normalized root mean square value, gives the better performance than other imputation methods.

V. CONCLUSION:

Micro array datasets with missing values gives worst performance during the analysis. Before measuring the performance and analysis of gene expression data we must ensure that it does not have missing values. In this paper we made a comparative study of three different missing data imputation methods (mean, miss-Forest and KNN Imputation). Performance of these imputation methods are compared based on normalized root mean square measure. In this experiment we used a real time Prostate cancer microarray dataset. Performance of these missing data imputation methods has been recorded by introducing the percentage of missing values in the dataset. Among the three imputation methods miss-Forest imputation method gives the best performance and low normalized root mean square value.

REFERENCES:

[1]. DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science, 278, 680–686.

[2]. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell, 9, 3273–3297

[3]. Chu,S., DeRisi,J., Eisen,M., Mulholland,J., Botstein,D., Brown,P.O. and Herskowitz,I. (1998) The transcriptional program of sporulation in budding yeast. Science, 282, 699–705.

[4]. Perou,C.M., Sorlie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S., Rees,C.A., Pollack,J.R., Ross,D.T., Johnsen,H., Akslen,L.A., Fluge,O., Pergamenschikov,A., Williams,C., Zhu,S.X., Lonning,P.E., Borresen-Dale,A.L., Brown,P.O. and Botstein,D. (2000) Molecular portraits of human breast tumours. Nature, 406, 747–752

[5]. Dinesh Singh,. "Gene Expression Correlates of Clinical Prostate Cancer Behavior". Cancer Cell, 1:203-209, March, 2002

[6]. Olga Troyanskaya and Michael Cantor "Missing value estimation methods for DNA microarrays" Vol. 17 no. 6 2001 Pages 520–525

[7]. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1997). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistics Society, 39(1), 1–38.

[8]. Faris, P. D., Ghali, W. A., Brant, R., Norris, C. M., Galbraith, P. D., & Knudtson, M. L. (2002). Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. Journal of Clinical Epidemiology, 55(2), 184–191.

[9]. Hui, D.,Wan, S., Su, B., Katul, G., Monson, R., & Luo, Y. (2004). Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations. Agricultural and Forest Meteorology, 121(1–2), 93–111.

[10]. Sartori, N., Salvan, A., & Thomaseth, K. (2005). Multiple imputation of missing values in a cancer mortality analysis with estimated exposure dose. Computational Statistics&DataAnalysis, 49(3), 937–953.

[11]. Dhlamini, S.M., Nelwamondo, F. V., & Marwala, T. (2006). Condition monitoring of HV bushings in the presence of missing data using evolutionary computing. Transactions on Power Systems,1(2), 280–287.

[12]. Nelwamondo, F. V., Mohamed, S., & Marwala, T. (2007a). Missing data: A comparison of neural network and expectation maximization techniques. Current Science, 93(11), 1514–1521.

[13]. Nelwamondo, F. V., Mohamed, S., & Marwala, T. (2007b). Missing data: A comparison of neural network and expectation maximisation techniques. Current Science, 93(12), 1514–1521.

[14]. Yansaneh, I. S., Wallace, L. S., & Marker, D. A. (1998). Imputation methods for large complex datasets: An application to the Nehis. In: Proceedings of the Survey Research Methods Section, pp. 314–319.

[15]. Allison, P. D. (2000). Multiple imputation for missing data. Sociological Methods & Research,28(3), 301–309.

[16]. Pérez, A., Dennis, R. J., Gil, J. F. A., Róndon, M. A., & López, A. (2002). Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia. Journal of Statistics in Medicine, 21(24), 3885–3896.

[17]. Enders, C. K. (2010). Applied Missing Data Analysis. New York, NY, The Guilford Press.

[18]. Eekhout, I., R. M. de Boer, et al. (2012). Missing data: a systematic review of how they are reported and handled. Epidemiology 23(5): 729-732.

[19]. L. Breiman. Random forests. Machine learning , 45(1):5{32, 2001. ISSN 0885-6125.

K Ishthaq Ahamed" An Analysis of Missing Data Imputation Methods on Microarray Gene Expression Data" International Journal of Computational Engineering Research (IJCER), vol. 09, no. 3, 2019, pp 39-42