International Journal of Research in Advent Technology, Vol.7, No.5S, May 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Comparative Analysis of Handling Missing Values and Imputation Techniques on Gene Expression Data

Alagukumar.S [1], Lawrance. R [2]
[1]Assistant Professor, [2]Director,
Department of Computer Applications,
Ayya Nadar Janaki Ammal College,Sivakasi – 626 124, Tamil Nadu, India
alagukumarmca@gmail.com,lawrancer@yahoo.com

Abstract: Data Mining is emerging research field in bioinformatics. In the present scenario data mining has become the eminent methodology for accessing huge volume of information from the data set. In practice, a data analyst spends much of time on preparing the data before doing any statistical operation. Data Cleaning is the process of transforming raw data into consistent data that can be analyzed. It is aimed at improving the content of statistical statements based on the data as well as their reliability like handling missing data, imputation or outlier handling for analysis. In this paper it has been analyzed various imputation techniques to handle the missing values on gene expression data.

Keywords - Data mining, Imputation, Missing Values, and KNN imputation, Gene Expression Data.

1.INTRODUCTION

Data Mining is the process of discovering the interesting patterns or information from the data in large databases. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically. Data mining is defined as knowledge Discovery in Databases, knowledge extraction, pattern analysis, data archeology, business intelligence [1,2,3].

Many Data Mining algorithms are used for preprocessing of data which removes noise from data sets, redundancy in data sets, which makes data sets useful for processing of knowledge from data sets[4].

D.T. Larose has examined the methods for imputing missing values for continuous variables, and categorical variables. Missing data may arise from any of several different causes. The method is simply to construct a flag variable and another method is for dealing with missing data to reduce the weight that the case wields in the analysis[5]. The structure of the paper is organized as imputation process, the various methods of handling missing values and imputation, comparative analysis, conclusion and references.

2.IMPUTATION PROCESS

Missing values becomes one of the problems that frequently occur in the data observation or data recording process. The needs of data completeness of the observation data for the uses of advanced analysis becomes important to be solved [13].

Many real-world datasets may contain missing values for various reasons. They are often encoded as NaNs, blanks or any other placeholders. Training a model with a dataset that has a lot of missing values can drastically impact the machine learning model's quality. One way to handle this problem is to analyze the missing data without risk losing data points with valuable information. A better strategy is to impute the missing values [6,7,8,10].

The process towards handling missing values or imputation always involves the following three steps.

1. Detect the inconsistency of the values or missing values.
2. Selection of the field or fields causing the inconsistency.
3. Correction of the fields or imputation to replace the missing values and inconsistent data [7].

Figure 1, depicts the process of handling missing values and imputations.

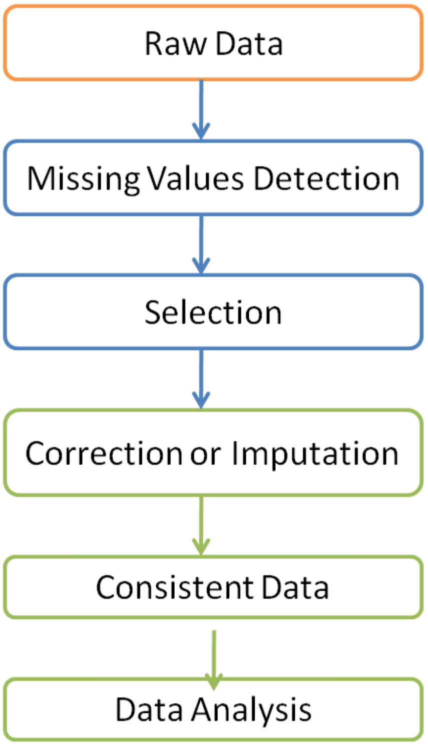

Fig. 1. Process of handling missing values and imputations

A. *Data formats*

The microarray gene expression dataset[9] can be form of an M x N matrix D of expression values, where the row represents samples X={ $x_1,x_2,x_3\ldots$ x_n }and column represents genes G= {$g_1,g_2,g_3\ldots, g_n$}, The gene expression data shown in Table 1.

Table 1 : Microarray Data

Samples	**Attributes(genes)**			
	Gene1	**Gene2**	**…**	**Gene m**
1	G(1,1)	G(1,2)	…	G(1,m)
2	G(2,1)	G(2,2)	…	G(2,m)
3	G(3,1)	G(3,2)	…	G(3,m)
…	…	…	…	…
…	…	…	…	…
n	G(n,1)	G(n,2)	…	G(n ,m)

3.HANDLING MISSING VALUES AND IMPUTATION METHODS

In real world data, the data are collected from the there are some instances where a particular element is absent because of various reasons, such as, corrupt data, failure to load the information, or incomplete extraction. Handling the missing values is one of the greatest challenges faced by analysts, because making the right decision on how to handle it generates robust data models [6,7,8,10]. Table 2 depicts the sample data with missing values where the NA represents the missing data or Not Available in the data set. There are three types of missing data, Missing Completely, Missing at Random and Missing Not at Random. The missing values are handled by two types, first one is List wise Deletion, where, delete all data from any participant with missing values, if the data is large enough, and then drop data second type is recover the values or replace the values by imputation methods.

Table 2. Sample data with missing values

X	S1	S2	S3	S4
G1	8.09	4.63	5.86	4.91
G2	-2.78	-2.34	NA	1.22
G3	NA	-0.28	1.59	-0.63
G4	0.54	0.39	NA	-0.1
G5	-1.52	-2.02	0.84	-1.82

There several methods are used to handle the missing values, such as mean, median, mode imputation and deletion.

B. Deleting Rows

This method commonly used to handle the null values. Here, delete a particular row if it has a null value for a particular feature and a particular column if it has more than 70-75% of missing values. This method is applied only when there are enough samples in the data set. Removing the data will lead to loss of information which will not give the expected results while predicting the output. Table 3 depicts the removed rows of missing values [6,7,8,10]. Where the G2, G3, G4 rows are removed from the data set.

Table 3. Removed Rows of Missed Values

X	S1	S2	S3	S4
G1	8.09	4.63	5.86	4.91
G5	-1.52	-2.02	0.84	-1.82

C. Mean and Median methods

The mean and median methods [6,7,8,10] can be applied on a feature which has numeric data like the gene expression data. It can only be used with numeric data. In these methods, the mean or median of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others. The missing values are imputed using mean method by using given formula $Mean_A = \frac{S}{N}$. Where $Mean_A$ is average represent the arithmetic mean , N represents the number of terms and S represents the sum of the numbers in the set.

The loss of the data can be negated by these methods which yields better results compared to removal of rows and columns. Table 4 depicts the result of mean imputation method where the missing values are replaced by the mean imputation method. Table 5 depicts the result of median imputation method, where the missing values are replaced by median imputation method.

Table 4. Imputation using Mean Method

X	S1	S2	S3	S4
G1	8.09	4.63	5.86	4.91
G2	-2.78	-2.34	**2.76**	1.22
G3	**1.08**	-0.28	**1.59**	-0.63
G4	0.54	0.39	2.76	-0.1
G5	-1.52	-2.02	0.84	-1.82

Table 5. Imputation using Median Method

X	S1	S2	S3	S4
G1	8.09	4.63	5.86	4.91
G2	-2.78	-2.34	1.59	1.22
G3	-0.49	-0.28	1.59	-0.63
G4	0.54	0.39	1.59	-0.1
G5	-1.52	-2.02	0.84	-1.82

D. K-nearest neighbor(KNN) Method

The k nearest neighbors is an algorithm that is used for simple classification. This algorithm can be very useful in making predictions about the missing values by finding the $k's$ closest neighbors to the observation with missing data and then imputing them based on the non-missing values in the neighborhood.

In k nearest neighbor imputation one defines a distance function $\mathrm{d(i, j)}$ that computes a measure of dissimilarity between records. A missing value is then imputed by finding first the k records nearest to the record with one or more missing values. Next, a value is chosen from or computed out of the k nearest neighbors.

In this paper, the VIM [8,11,12] package (R programming language) is used to impute the missing values, where the VIM package is contains a function called kNN that uses *Gowers* distance [9] to determine the k nearest neighbors.

Gower's distance between two records labeled i and j is defined as $\mathrm{d_{g(i,j)}} = \frac{\sum_{\mathrm{w}} \mathrm{d_k(i,j)}}{\sum_{\mathrm{k}} \mathrm{w_{ijk}}}$ where the sum runs over all variables in the record and $\mathrm{d_k(i, j)}$ is the distance between the value of variable k in record i and record j.

For numerical variables the distance is calculated by $1-(x_i-x_j)/(\max(x)-\min(x)$. The weight $w_{ijk}=0$ when the k variable is missing in record i or record j and otherwise 1.

The table 6 depicts the imputation result using kNN method, where k value is 2, which is given by the user. The missing values are replaced by the k values.

Table 6. Imputation using k Nearest Neighbor Method

X	S1	S2	S3	S4
G1	8.09	4.63	5.86	4.91
G2	-2.78	-2.34	1.215	1.22
G3	-0.49	-0.28	1.59	-0.63
G4	0.54	0.39	1.215	-0.1
G5	-1.52	-2.02	0.84	-1.82

4. COMPARATIVE ANALYSIS OF IMPUTATION METHODS

A number of imputation methods exist, each having its own characteristic and doing well in different situations. Every imputation methods have its-own strengths. A comparative analysis of common imputation methods based on different imputations and their advantages and Disadvantages has displayed in Table 7.

Table 6. Comparative Analysis of Imputation methods

Methods	Handling Missing Values by Removing / Imputations	Advantages	Disadvantages
Delete Rows	Removing	Complete removal of data with missing values results in robust	Loss of information and data Works poorly
Mean and Median Method	Imputation	Easy and fast. Works well with small numerical datasets. It can prevent data loss which results in removal of the rows and columns	Works poorly compared to other multiple-imputations method. It only works on the column level. It gives poor results on categorical features.
k nearest neighbor Method	Imputation	It can be much more accurate than the mean, median imputation methods	Computationally expensive. It can be critical in data mining where large databases are being extracted

5. CONCLUSION

In this paper, it has been analyzed various methods of handling missing values and imputations methods on gene expression data. The imputation methods are plays an important role in data pre-processing before applying a number of machine learning and data mining algorithms on the real valued data sets. Lastly a comparative analysis has been given based on different issues of handling missing values and imputations. The preprocessed data and imputed data give the better accuracy in the data analysis, rule generation and classification model. The imputed data is used to predict data in efficient and produce the better result.

REFERENCES

[1] Hen J. and Kamber M., “Data Mining: Concepts and Techniques”,Second Edition, ELSEVIER Publications, ISBN: 978-81-312-0535-81, 2005.

[2] Alagukumar, S., and R. Lawrance. "A Selective Analysis of Microarray Data Using Association Rule Mining." Procedia Computer Science 47 (2015): 3-12.

[3] Alagukumar, S., and R. Lawrance. “Algorithm for Microarray Cancer Data Analysis using Frequent Pattern Mining and Gene Intervals”, International Journal of Computer Applications.pp.1-6, 2015.

[4] Srinivasan Parthasarathy and Charu C. Aggarwal, On the use of conceptual Reconstruction for Mining Massively Incomplete Data Sets, IEEE PP.1512-1521,2003.

[5] D.T. Larose andC.D. Larose, “Imputation of Missing Data”, Wiely Online Library,2014.DOI:10.1002/9781118874059.ch13.

[6] Buuren, S. V., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software

[7] De Jonge, Edwin, and Mark Van Der Loo. *An introduction to data cleaning with R*. Heerlen: Statistics Netherlands, 2013.

[8] Kowarik, A., & Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software*, *74*(7), 1-16.

[9] J.C. Gower. A general coefficient of similarity and some of its properties. Biometrics,27:857--874, 1971.

[10] Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P., & Botstein, D. (1999). Imputing missing data for gene expression arrays.

[11] https://www.r-project.org/

[12] Crookston, Nicholas L., and Andrew O. Finley. "yaImpute: an R package for kNN imputation." Journal of Statistical Software. 23 (10). 16 p. (2008).

[13] Pratama, Irfan, et al. "A review of missing values handling methods on time-series data." 2016 International Conference on Information Technology Systems and Innovation (ICITSI). IEEE, 2016.