I J C T A, 9(1), 2016, pp. 79-83

Optimized Method using Pre-Handler, Clustering and SKNN Imputation (OPCSI) for handling Missing Value

S. Kavitha[*] and M. Hemalatha[**]

Abstract: Preprocessing describes any type of processing performed on raw microarray data to prepare it for another processing procedure. It transforms the data into a format that can be more easily and effectively processed. Data preprocessing is a task that should be given utmost attention for the following two reasons(i)Quality decisions are based on quality data - e.g., duplicate or missing data may cause incorrect or even misleading statistics.(ii)Data preparation, cleaning, and transformation comprise the majority of the work in a data mining application (90%).In this research work, preprocessing step is used to handle the issues raised by missing values in microarray data. Enhanced Imputation based method is proposed for handling missing values and to create a complete dataset.

Keywords: missing values, SKNN, OPSCI.

INTRODUCTION

Data pre-processing is a critical task in the knowledge discovery process in order to ensure the quality of the data to be analyzed. One widely studied problem in data pre-processing is the handling of missing values with the aim to recover its original value. A missing value can signify a number of different things. Perhaps the field was not applicable, the event did not happen, or the data was not available. It could be that the person who entered the data did not know the right value, or did not care if a field was not filled in.

Anyone who does statistical data analysis or data cleaning of any kind runs into the problems of missing data. In a characteristic dataset we always land up in some missing values for attributes. For example in surveys people generally tend to leave the field of income blank or sometimes people have no information available and cannot answer the question. Also in the process of collecting data from multiple sources some data may be inadvertently lost. For all these and many other reasons, missing data is a universal problem in both social and health sciences. A missing data is defined as an attribute or feature in a dataset which has no associated data value. Incomplete data is an unavoidable problem in dealing with microarray data. Correct treatment of these data is crucial, as they have a negative impact on the interpretation and result of gene analysis. Normally, missing rates less than one per cent are considered trivial, 1-5 per cent are considered manageable. But databases with 5-15% missing data values rate needs sophisticated methods to handle them correctly and more than 15% requires careful handling as they affect interpretation.

GENERAL APPROACH

In the analysis of gene expression data, KNNimpute is a popular method to impute missing expression values based on weighted k nearest neighbor. Existing approaches use single imputation or non-imputation algorithms to handle the problem of missingness in microarray data.Garc1aLaencina *et al.* (2009) – Termed as Imputation Method Using Weighted KNN Method (IWKNN). It Can predict both quantitative and qualitative data.Multiple missing values are easily handled by the IKNN algorithm.The major drawback of

[*] Research Scholar, Karpagam University, Coimbatore, *E-mail: kavi_shanmukavi@yahoo.co.in*

[**] Associate Professor, Karpagam University, Coimbatore, *E-mail: csresearchhema@gmail.com*

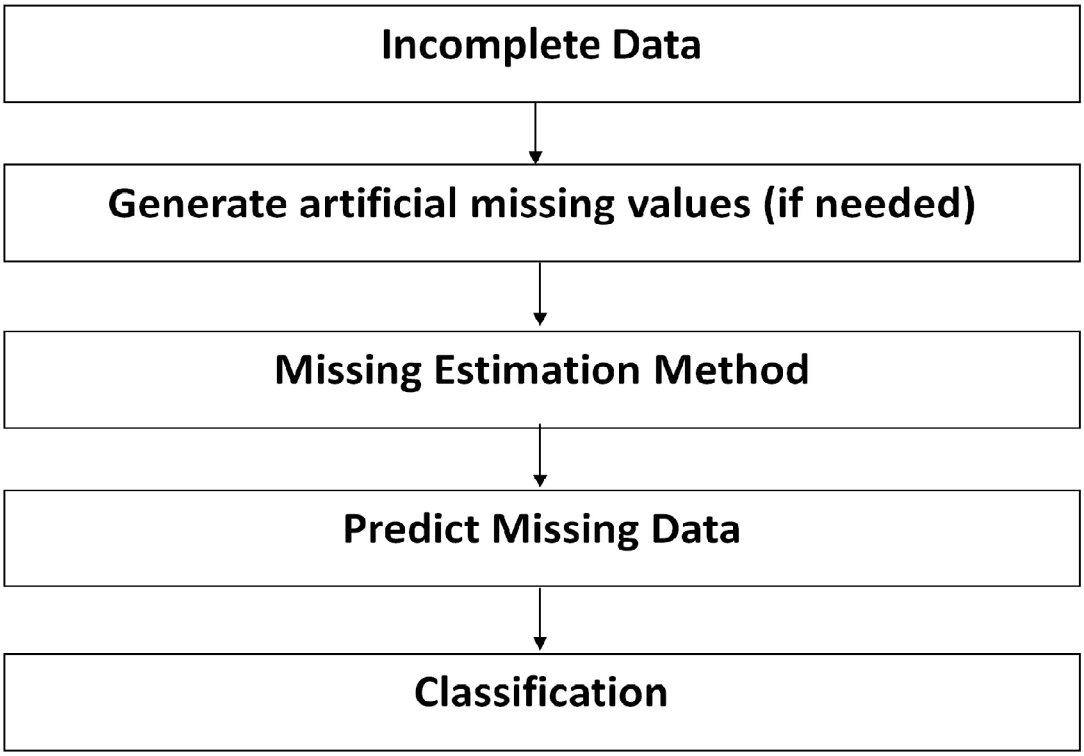

Figure 1

this approach is that whenever the algorithm looks for the most similar instances, the algorithm searches through all the dataset. Thus increasing the search time. This is solved in the present research work by including a clustering step and by replacing KNN Method with weighted KNN Imputation Method. *The second drawback is the Performance of the algorithm degrades* with *high rate of missing values.* Proposed Method is termed as Optimized Method using Pre-Handler, Clustering and SKNN Imputation (OPCSI) OPCSI.

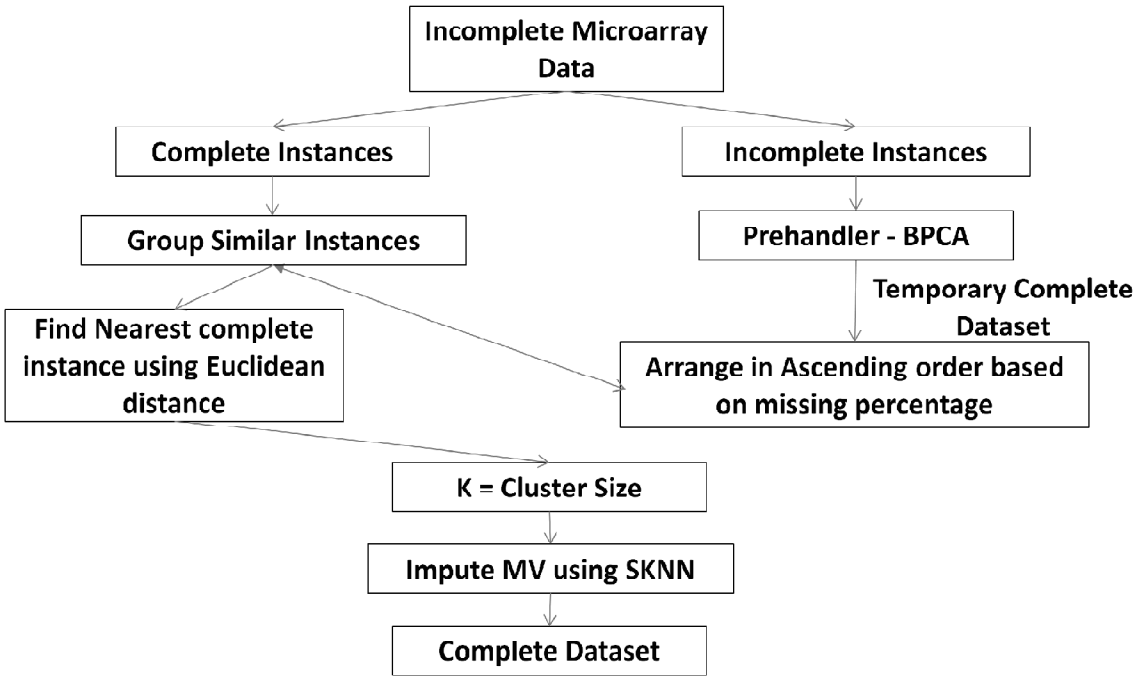

Figure 2

Grouping Similar Genes Using K-Means Clustering Algorithm

Set K = Number of classes in Micro Array Data

Step 1: Put any initial partition that classifies the data into 'k' clusters.

Take the first 'k' data as *single-element clusters*

Assign each of the remaining (N-k*)* data to the cluster with the *nearest centroid*. After each *assignment*, *compute* the *centroid* of the *gaining* cluster.

Step 3: Take each data in *sequence and compute* its *distance* from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample. The distance metric used is *Euclidean distance*.

Step 4: Repeat step 3 until convergence is achieved, that is, until a pass through the training sample causes *no new assignments*.

Pre-Handler Using Bayesian Principal Component Analysis

Bayesian model is a popular model which has been widely used. BPCA method mainly comprise three elementary processes, which are

1. Principal component regression.
2. Bayesian estimation.
3. Expectation-maximization-like repetitive algorithm.

The performance of BPCA method has been compared with that of KNN and SVDimpute and showed its advantages over them. However, when genes have dominant local similarity structures, BPCA doesn't work as well as KNN

SKNN Impute

SKNN impute was proposed as an improvement to KNNimpute, differing from the latter method in two main points:

- MVs are estimated sequentially starting with the gene having the smallest missing rate
- SKNN impute uses the estimated values for estimating the MVs of the remaining genes.

In SKNN impute, X is split into two sets by considering the genes with no MVs ($X_{complete}$) and the genes comprising MVs ($X_{incomplete}$). The former matrix is used as the candidate set, while the latter contains the target genes to be estimated following the order of their missing rate. Applying the KNN principle, the target gene's missing entries are filled according to the weighted average of all its neighbours. Once the estimation of a given target gene is completed, the candidate set is updated with that gene, so that it can be used for the next estimation round.In SKNNimpute, all MVs in a given target gene are estimated simultaneously using the selected neighbour genes, since the latter genes were taken from Xcomplete. Moreover, a simple Euclidean distance can be used. Consequently, SKNNimpute offers an advantage over KNNimpute in terms of speed.

Experimental Results

Several experiments were conducted using five datasets. The experiments were designed to study the impact and importance of handling missing values. Two parameters, namely, NRMSE and speed, were used during performance evaluation. The results pertaining to NRMSE and speed are presented in Figures 3 and 4 respectively.

Datasets Used

Table 1

Cancer Dataset	*No. of Features*	*No. of Samples*	*No. of Normal Cases*	*No. of Cancer Cases*
Breast	24481	97	51	46
Colon	1909	62	22	40
Ovarian	15154	253	91	162
CNS (Central Nervous System)	7129	60	39	21
Leukemia	7129	72	25	47

Performance Metrics Used

Normalized Root Mean Square Error (NRMSE)

Its Helps to evaluate the effectiveness of missing values imputation using the following equation.

$$NRMSE = \frac{\sqrt{\mu[(y_{true} - y_{imp})^2]}}{\sigma(y_{true})}$$

where μ is the mean and σ is the variance and are calculated over missing entries in the whole matrix. y_{true} is the original dataset and y_{imp} is the imputed dataset. Its Helps to analyze the speed complexity of the imputation algorithm. Estimated as the total execution time taken by the algorithm to convert the incomplete dataset into a complete dataset.

Speed

Speed Helps to analyze the speed complexity of the imputation algorithm And also Estimated as the total execution time taken by the algorithm to convert the incomplete dataset into a complete dataset.

NRMSE

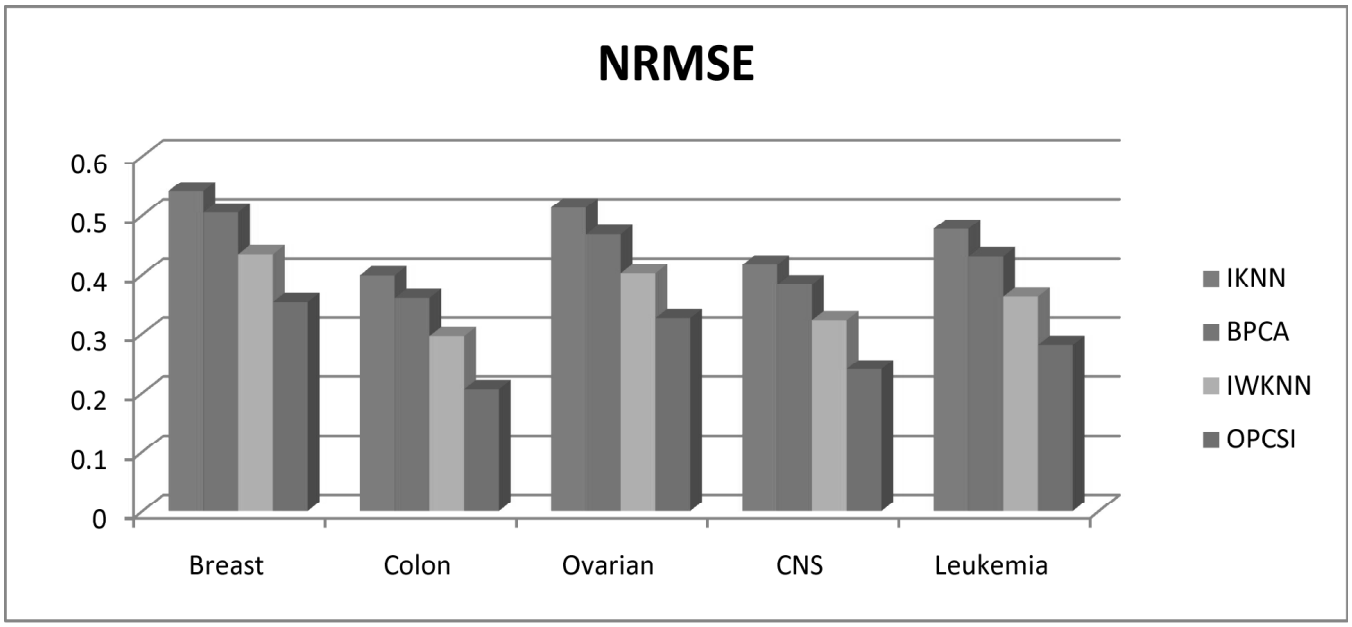

Figure 3

EXECUTION SPEED (SECONDS)

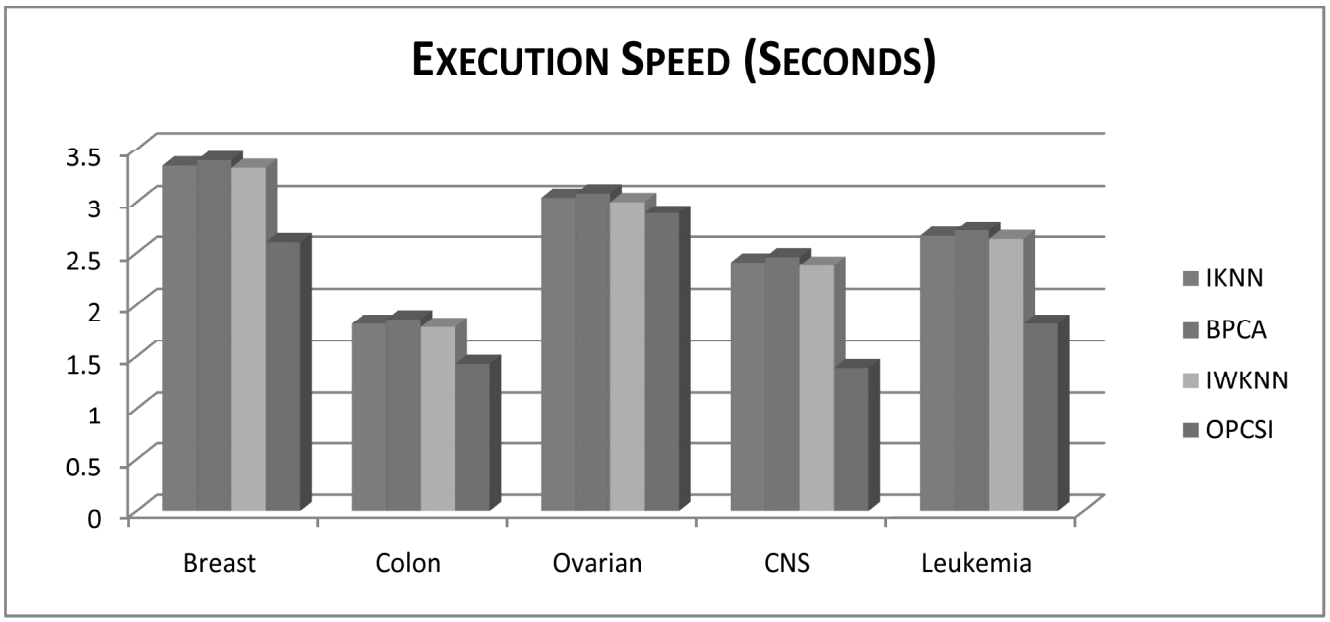

Figure 4

CONCLUSION

Microarray datasets tend to be small in sample size due to the cost associated with the arrays. There are many more methods for handling missing values. This huge dataset makes missing value handling a crucial step while designing and building a classifier. This research work has been proposed to use Optimized Method using Pre-Handler, Clustering and SKNN Imputation (OPSCI) for handling missing value and has been analyzed. The proposed method is designed with the help of BPCA and SKNN for handling huge number of missing values. The BPCA creates temporary complete dataset that is given as input to SKNN Algorithm. Experimental results proved that the proposed method is efficient and enhances the process of handling missing values which can safely be used by gene selection and classification.

References

[1] Shin-mu Tsengand Kuo-howang, A Pre-processing Method to Deal with Missing Values by Integrating Clustering and Regression Techniques, Applied Artificial Intelligence 17:535–544, 2003.

[2] Alshamlan, H.M., Badr, G.H. and Alohali, Y. (2013), A study of cancer microarray gene expression profile: Objectives and approaches, Proceedings of the World Congress on Engineering, Vol. II, Pp. 1-6.

[3] Bhavik Doshi , Handling Missing Values in Data Mining [IV] Chen, C.K. (2012) The classification of cancer stage microarray data, Computer Methods and Programs in Biomedicine, Vol. 108, Issue 3, Pp. 1070- 1077.

[4] JiYí Kaiser Faculty of Civil Engineering, Czech Technical University Dealing with Missing Values in Data *Journal of Systems Integration* 2014/1.

[5] S Greenland, WD Finkle, A critical look at methods for handling missing covariates in epidemiologic regression analyses - American journal of epidemiology, 1995 - Oxford Univ Press.

[6] JW Graham, PE Cumsille, Methods for handling missing data, Handbook of psychology, 2003 - Wiley Online Library.