IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

Overview on Microarray using Advanced Regression for Lost Data

K. Lakshmipriya[1], Dr. R. Manickachezian[2]

Research Scholar, Dr. Mahalingam Centre for Research and Development, NGM College, Pollachi, India[1]

Department of Computer Science, NGM College (Autonomous), Pollachi, Coimbatore, India[2]

Abstract: This overview is based on dealing with lost data and Missing data during the time of data transaction and bulk data transmission. Missing data are characterized as a portion of the qualities in the data set are either lost or not watched or not accessible because of natural or non natural reasons. Data with missing qualities confuses both the data examination and the accommodation of an answer for new data. Numerous specialists are working on this issue to present more modern techniques. Despite the fact that numerous strategies are available, investigators are confronting trouble in seeking an appropriate technique because of absence of information about the strategies and their applicability. This research paper additionally directs a formal review of the missing data strategy. It talks about the strategies that are analyzed in the written works and perceptions that the authors have made. This survey is based on microarray data processing with regression method.

Keywords: Missing data, Data transmission, Micro array, Regression method, Data fixing.

I. INTRODUCTION

A large portion of this present datasets experience suffers from the issue of missing data. It might lead information mining examiners to end with wrong inferences about data under review. Data mining is the process which provides a concept to attract attention of users due to high availability of huge amount of data and need to convert such data into useful information. Data preparation is a principal phase of data investigation [5][8]. Three types of mean imputation methods presented on missing data [20]. Rubin investigated about inference and missing data and multiple imputation for non-reaction in the overview [1]. Allison explored estimates of linear models with incomplete data and on missing data [2]. Myrtveit et al. connected missing data strategies to a software project data set, and assessed four missing data procedure are list wise deletion (LD), mean imputation (MI), similar response pattern imputation (SRPI) and full information maximum likelihood (FIML) [4]. Junninen et al. evaluated and compared with univariate and multivariate methods for missing data imputation in air quality data sets[10]. Types of Incomplete Data are little and Rubin characterize a list of missing mechanisms, which are generally acknowledged by the community. There are three mechanisms under which missing data can happen 1) Missing completely at random (MCAR): MCAR is the probability that a observation (Xi) is missing, is unrelated to the estimation of Xi or to the estimation of some other variable and the explanation behind missing is completely random. This circumstance is uncommon in real world and is generally talked about in statistical theory. 2) Missing at random (MAR): MAR is the probability of the observed missingness design, given the observed and unobserved data, does not depend on the values of the unobserved data. This component is normal in practice and is normally considered as the default kind of missing data. 3) Not missing at random (NMAR). If the probability that an observation is missing depends on information that is not observed, this kind of missing data is called not missing at random [1]. This circumstance is generally confused and there is no universal solution. 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. This is because every standard statistical method works on the fact that every problem has information on all the variables an it needs to be analyzed.

II. BACKGROUND STUDY – MICRO ARRAY DATA

Microarray technology is a powerful tool for modern biomedical research. It could monitor relative expression of thousands of data under a variety of experimental conditions. Therefore, it has been used widely in numerous studies over a broad range of biological disciplines, such as Missing data and lost data. Although microarray technology has been used for several years, expression data still contain missing values due to various reasons.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

Basically, microarray data contain 1-10% missing values that could affect up to 95% of data [3]. The occurrence of missing values in microarray data disadvantageously influences downstream analyses, such as discovery of differentially expressed data [4][6], construction of gene regulatory networks [18][12], supervised classification of clinical samples [19], gene cluster analysis [3][12], and biomarker detection.

One straightforward solution to solve the missing value problem is to repeat the microarray experiments, but that is very costly and inefficient. Another solution is to remove data (rows) with one or more missing values before downstream analysis, but it is easily seen that part of important information would be lost. Hence, advanced algorithms must be developed to accurately impute the missing values. The most common and simple solution to this problem is if any case has missing data for any of the attribute to be analyzed we can simply ignore it. This will give us a dataset which will not contain any missing value and we can then use any standard methods to process it further. But this method has a major drawback which is deleting missing values sometimes might lead to ignoring a large section of the original sample. This paper first illustrates different types of missing values and analyzes their consequences on datasets.

Using modern mathematical and computational techniques can effectively impute missing values. Early approaches included replacing missing values by zero, row average or row median [11]. Recently, many studies found that merging information from various biological data can significantly improve the missing values estimation. The categorized existing algorithms into four different classes:
(1) local algorithms,
(2) global algorithms,
(3) hybrid algorithms, and
(4) knowledge assisted algorithms [10][7].

The first category includes k nearest neighbors (KNN) [11], least squares adaptive (LSA). The second category includes Bayesian principal component analysis (BPCA), partial least squares (PLS) belong to the knowledge assisted approach algorithms. In this study, we did not use the hybrid algorithms and the knowledge assisted algorithms because their programs are not freely available or cannot be easily modified.
In the past few years, several papers have preliminary and objective analyses for the systematic evaluation of different imputation algorithms. The weaknesses of these studies are as follows. First, few microarray datasets were used. Second, few independent rounds of the imputed procedure were performed (usually 10 times). Third, single performance measure was used. Here, we present a fair and comprehensive evaluation to assess the performances of different imputation algorithms on different datasets using different performance measures.

III. METHODOLOGIES

a) KNN: K – Nearest Neighbour

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. To evaluate any technique we generally look at 3 important aspects:
1. Ease to interpret output
2. Calculation time
3. Predictive Power
For the statistic index, we used the normalized root mean squared error (NRMSE) to evaluate the performance of the imputation algorithms. Lower the value of the statistic index, better the algorithm performs. Normalized root mean squared error (NRMSE): NRMSE is a popular index used to evaluate the similarity between the true values and the imputed values

b) LSA: Least Square Adaptive

An important data analysis in the microarray data Least Square Adaptive. In this study, k-means was used to do data clustering for the complete datasets and the imputed datasets. We used cluster pair proportions (CPP) as a clustering index to evaluate the performance of the algorithms. Higher the value of the clustering index, better the algorithm performs. CPP. G_1,G_2, ..., G_{16} are genes in the microarray data. CC1,…, CC4 are the four clusters for the complete dataset. CI1,…,CI4 are the four clusters for the imputed dataset.
A new family of algorithms is presented which solve exactly the underdetermined Least square adaptive (LSA) criterion. The order-recursive solutions of the least-squares problem, in particular the QR decomposition, are used to obtain the order recursive URLS algorithms. Such algorithms have better numerical properties than the Fast Transversal Filters (FTF) counterparts and they provide the flexibility to alter the order of the URLS algorithm without adding extra variables. Also, they are amenable to parallel implementations and data predictions.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

c) BPCA: Bayesian Principal Component Analysis

A high BPCA value indicates that the list of the significantly differentially expressed genes of the complete data is similar to that of the imputed data. And it also means that the imputed data does not significantly change the result of downstream analysis, so the algorithm has excellent performance. We expect that a good algorithm has a high BPCA value. The BPCA is defined as follows:

$$BLCI(B_{CD}, B_{ID}) = \frac{n(B_{CD} \cap B_{ID})}{n(B_{CD})} + \frac{n(B_{CD}^{C} \cap B_{ID}^{C})}{n(B_{CD}^{C})} - 1,$$

The Bayesian solution provides two notable results in relation to PCA. The first are uncertainty bounds on principal components (PCs), and the second is an explicit distribution on the number of relevant PCs. Bayesian Principal Component Analysis (BPCA) is one of the classical data analysis tools for dimensionality reduction. It is used in many application areas including data compression, de-noising, pattern recognition, shape analysis and spectral analysis. Bayesian solutions of rank-restricted models have usually been based on the Factor Analysis model. The simulated data study in the last section suggests that the proposed posterior rank distribution is capable of correct inference, if the data comply with the model. Furthermore, all but one of the true simulated parameters are within the estimated uncertainty bounds.

d) PLS: Partial Least Squares.

A distinct illustration that can point out the optimal method for the microarray datasets used. The x-axis means the algorithms used and the y-axis means the average rank of each algorithm. For example, if we perform an experiment with 5 independent rounds, in which ranks of an algorithm are 1, 2, 2, 1 and 2 respectively. The average rank of the algorithm in this experiment is (1 + 2 + 2 + 1 + 2)/5 = 1.6.

$$NRMSE = \sqrt{\frac{mean\left[\left(\mathbf{y}_{\mathbf{guess}} - \mathbf{y}_{\mathbf{answer}}\right)^{2}\right]}{variance\left[\mathbf{y}_{\mathbf{answer}}\right]}}$$

IV. RESULT

The below mentioned table 1.1 and figure 1.1 shows the comparative study of four major algorithms involved in missing data.

Table 1.1 Accuracy Measurement

Algorithm	Year	Result in missing data (Accuracy)
K – Nearest Neighbour	2007	84.9 %
Least Square Adaptive	2010	88.7 %
Bayesian Principal component analysis	2014	89.9 %
Partial Least Squares	2017	90. 3 %

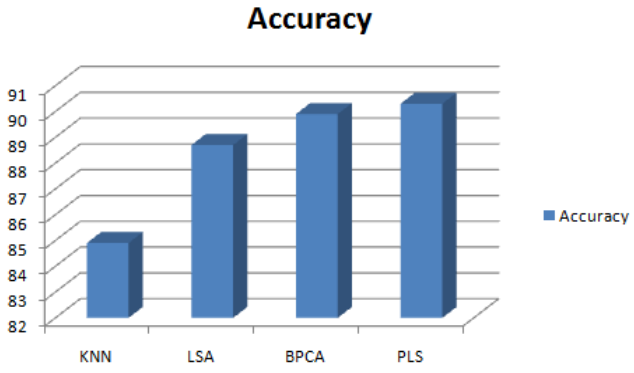

Figure 1.1 Machine Learning Algorithms

V. CONCLUSION

Data Analysis and preparation is the primary step in data mining process. We first identify different types of missing data and then discuss four approaches (KNN,LSA,BPCA,PLS) to deal with missing data in different scenarios. This

paper addresses the issues of handling missing values in datasets and methods in which missing values can be tackled. We first discuss the different types of missing data and analyze their impact on the dataset. We now look into the problem of missing values in monotonous datasets. We suggest a simple pre-processing method which when used with other techniques help in eliminating missing values and help in maintaining the dataset monotonous. The experiments to test the algorithm and find that taking the most frequent value and replacing it in place of missing values give better results. Missing data sometimes also disguise themselves as valid data and are difficult to identify. We therefore propose a heuristic approach to tackle a practical and challenging issue of cleaning disguised missing data. With the help of this approach we identify suspicious sample of data and then develop an unbiased sample heuristic approach to discover missing values.

REFERENCES

[1] Rubin, D.B., (1976). Inference and missing data.Biometrika, 63(3):581-592.
[2] Allison, P.D., (1987). Estimation of linear models with incomplete data.Sociological methodology, 71-103.
[3] Schafer, J.L., (1997). Analysis of incomplete multivariate data.Monographs on Statistics and Applied Probability No. 72.
[4] Myrtveit, I., Stensrud, E., & Olsson, U. H. (2001). Analyzing data sets with missing fata: An empirical evaluation of imputation methods and likelihood-based methods. IEEE Transactions on Software Engineering, 27:999–1013.
[5] Smyth, P. (2001). Data mining at the interface of computer science and statistics. In Data mining for scientific and engineering applications 35-61. Springer US.
[6] Strike, K., El Emam, K., Madhavji, N., (2001). Software cost estimation with incomplete data. IEEE Transactions on Software Engineering ,27 (10):890–908.
[7] Briggs, A., Clark, T., Wolstenholme, J., Clarke, P., (2003). Missing.... presumed at random: cost-analysis of incomplete data. Health Economics,12:377–392.
[8] Zhang, S., Zhang, C., & Yang, Q., (2003). Data preparation for data mining.Applied Artificial Intelligence, 17(5-6):375-381.
[9] Grzymala-Busse, J. W. (2004). Data with missing attribute values: Generalization of indiscernibility relation and rule induction. Transactions on Rough Sets 1:78-95.Springer Berlin Heidelberg.
[10] HeikkiJunninen, HarriNiska, Kari Tuppurainen, JuhaniRuuskanen, MikkoKolehmainen, (2004). Methods for imputation of missing values in air quality data sets, Atmospheric Environment 38:2895–2907
[11] Kin Wagstaff, (2004). Clustering with Missing Values: No Imputation Required,NSF grant IIS- 0325329:1-10.
[12] A.Rogier, T.Donders, Geert J.M.G Vander Heljden, Theo St ijnen, Kernel G.M Moons (2006). Review: A gentle introduction to imputation of missing values, Journal of Clinical Epidemiology,59:1087-1091.
[13] Y. Kou, C.-T.Lu, and D. Chen (2006). Spatial weighted outlier detection. In Proceedings of the Sixth SIAM International Conference on Data Mining,614–618,Bethesda, Maryland, USA.
[14] Song, Q., &Shepperd, M. (2007). A new imputation method for small software project data sets. Journal of Systems and Software, 80(1): 51-62.
[15] Graham, J. W. (2009). Missing data analysis: Making it work in the real world. Annual review of psychology, 60:549-576.
[16] Han, J., Pei, J., &Kamber, M. (2011). Data mining: concepts and techniques. Morgan Kaufmann Publishers, 225 Wyman Street, Waltham, USA pp. 83-91.
[17] A Case Study of Heart Failure Datasetl (2012). 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012).
[18] Dr.A.Sumathi, (2012). Missing Value Imputation Techniques Depth Survey And an Imputation Algorithm To Improve The Efficiency Of Imputation.IEEE- Fourth International Conference on Advanced Computing, ICoAC.
[19] N. Poolsawad L. Moore C. Kambhampati and J. G. F. Cleland (2012). Handling Missing Values in Data Mining - A Case Study of Heart Failure Datasetl.9th International Conference on Fuzzy Systems and Knowledge Discovery.
[20] Noor, M. N., Yahaya, A. S., Ramli, N. A., & Al Bakri, A. M. M. (2014). Mean imputation techniques for filling the missing observations in air pollution dataset Key Engineering Materials 594-599:902-908 Trans Tech Publications

BIOGRAPHIES

K. Lakshmipriya received her Bsc (Computer science) from NGM College, Pollachi, India. She completed her Master of Computer Application (MCA) from NGM College, Pollachi, India. Currently, she is a Research Scholar at Department of Computer Science, NGM College, Pollachi, India. She participated in a International Conference. Her area of interest includes Data mining, Missing data.

Dr. R. ManickaChezian received his M.Sc., degree in Applied Science from P.S.G College of Technology, Coimbatore, India in 1987. He completed his M.S. degree in Software Systems from Birla Institute of Technology and Science, Pilani, Rajasthan, India and Ph.D degree in Computer Science from School of Computer Science and Engineering, Bharathiar University, Coimbatore, India. He served as a Faculty of Maths and Computer Applications at P.S.G College of Technology, Coimbatore from 1987 to 1989. Presently, he has been working as an Associate Professor of Computer Science in N G M College (Autonomous), Pollachi under Bharathiar University, Coimbatore, India since 1989. He has published one-fifty papers in international/national journal and conferences: He is a recipient of many awards like Desha Mithra Award and Best Paper Award. Recently he received the award "Best Computer Science Faculty of the Year 2015" from Association of Scientists, Developers and Faculties. His research focuses on Network Databases, Data Mining, Distributed Computing, Data Compression, Mobile Computing, Real Time Systems and Bio-Informatics.