International Journal of Electrical and Computer Engineering (IJECE)
Vol. 6, No. 6, December 2016, pp. 3222 – 3228
ISSN: 2088-8708, DOI: 10.11591/ijece.v6i6.10632

A New Paradigm for Development of Data Imputation Approach for Missing Value Estimation

Madhu G and Nagachandrika G
Dept. of Information Technology
VNR Vignana Jyothi Institute of Engineering and Technology
Hyderabad-90, Telangana, India

Article Info

Article history:
Received Mar 26, 2016
Revised Nov 8, 2016
Accepted Nov 21, 2016

Keyword:
Centroids
Distance
Imputation
Missing value
Nearest Neighbour

ABSTRACT

Many real-world applications encountered a common issue in data analysis is the presence of missing data value and challenging task in many applications such as wireless sensor networks, medical applications and psychological domain and others. Learning and prediction in the presence of missing value can be treacherous in machine learning, data mining and statistical analysis. A missing value can signify important information about dataset in the mining process. Handling missing data value is a challenging task for the data mining process. In this paper, we propose new paradigm for the development of data imputation method for missing data value estimation based on centroids and the nearest neighbours. Firstly, identify clusters based on the k-means algorithm and calculate centroids and the nearest neighbour data records. Secondly, the nearest distances from complete dataset as well as incomplete dataset from the centroids and estimated the nearest data record which tends to be curse dimensionality. Finally, impute the missing value based nearest neighbour record using statistical measure called z-score. The experimental study demonstrates strengthen of the proposed paradigm for the imputation of the missing data value estimation in dataset. Tests have been run using different types of datasets in order to validate our approach and compare the results with other imputation methods such as KNNI, SVMI, WKNNI, KMI and FKNNI. The proposed approach is geared towards maximizing the utility of imputation with respect to missing data value estimation.

Corresponding Author:
Madhu.G
Department of Information Technology
VNR Vignana Jyothi Institute of Engineering and Technology
Telangana, Hyderabad-90, INDIA.
E-mail: madhu_g@vnrvjiet.in

1. INTRODUCTION

Many real-world applications encountered a common issue in data analysis is the presence of missing or incomplete data values. However, data mining applications are associated with industry applications, wireless sensor networks, medical applications, psychological applications and others. Modern computational techniques of data cleaning require complete dataset.These databases are highly susceptible to missing and inconsistent data due to their huge amount of data sizes [1-2]. Missing value has different causes like the data value might not be recorded, equipment malfunctions, improper measurements and deny answers to certain questions [3-4]. Missing data value in the training dataset can reduce the performance of the model or that can lead biased results to the model and it leads to misleading inferences in data analysts. Recently, many researchers have been proposed several methods to treat missing value problems for real-world applications [5-10]. Generally, missing value treatment methods classified into : i) ignoring and discarding data value or case/pairwise deletion [11]; ii) parameter estimation/Expectation-Maximization algorithm [12]; iii) Imputa-

tion, represents the process of filling the missing data values in datasets by some plausible values based on information available in the dataset [13]. First, two methods are suffering with several limitations such as: i) high sensitivity to the outliers and slow computations; ii) Imputation methods are suffering with how to select the conditioning variables and bias those results from a bad choice. To address these issues, we develop a new paradigm for data imputation of the missing data value estimation using cluster centroids and the nearest neighbours. Also, our approach uses the centroids and the nearest neighbours for the family of k-means clustering algorithm. This study demonstrates that the proposed imputation method can significantly estimate the missing value in datasets.

The rest of the paper is organized as follows: related work presented in section-2, while in section-3, proposed a new paradigm for data imputation method for the missing data value estimation step by step approach. In section-4 presents the results and analysis. Conclusions are deferred to section-5.

2. RELATED WORKS

This section represents mechanisms of different data imputation methods and treatments for the purpose of data cleaning in the data mining process. Description of the data imputation mechanisms and treatments are as follows:

2.1. Mechanism of missing data imputation

Little, R.J., Rubin, D.B [14] suggested three types of mechanisms for imputation of missing data such as: i) Missing Completely at Random (MCAR): Suppose few data records are missing on X, then these data records are said to be MCAR if the probability of X is missing with unrelated X or other variable Y then the $prob(Xismissing|Y,X) = prob(Xismissing)$. ii) Missing at random (MAR): Suppose few data records X are missing at random, if the probability of that X is missing does not depend on the value X, after controlling the observed data but not on missing data then the $prob(Xismissing|Y,X) = prob(Xismissing|X)$; iii) Not missing at random (NMAR):): In this approach assumption is violated and sharing of a sample having a missing data value for an attribute depends on the missing values.

2.2. Missing data treatment methods

In literature various approaches have been suggested to treat missing value problems. Little and Rubin [14] classified these approaches into three categories: i) Ignoring and deleting data records: this approach to discard data with the missing value into a complete case analysis and delete instances and/or attributes. These methods are applicable only missing data are MCAR otherwise it can lead bias results. ii) Parameter estimation approach: this approach is based on estimation of the parameters in given model then the presence of missing value based on Expectation-Maximization algorithm [12]. iii) Imputation Methods: to impute the missing value with probable value based on information obtainable in the dataset. The main idea of this approach is to employ known value that can be identified in the valid data values of the given dataset to assist in assessing the missing values. To impute the missing value based different type of methods such as single imputation and multiple imputation methods [15]. But multiple imputation approaches are computationally more expensive than the single imputation methods [15]. However, they accommodate for sample variability of the estimated value and uncertainty associated with a model used for computation [16]. Also, these two methods can be classified into three categories such as i) data driven approach; ii) model based approach; and iii) machine learning approach [14-17]. The above discussion leads us to propose this research work presents a novel framework for data imputation which addresses the problem of missing data values.

3. A NEW PARADIGM FOR DATA IMPUTATION: PROPOSED APPROACH

This section, presents a new paradigm for data imputation to address the problem of missing value which is based on two distance measures that are used to define a new feature between its cluster center and the nearest neighbour respectively as shown in Fig.1.

Let M be the given dataset containing of rows and columns and each row is represented by a row vector $(R_1, R_2, .R_m)$ and each column is represented by a column vector $(C_1, C_2, ..C_n)$ and the dataset M is represented as m x n matrix:

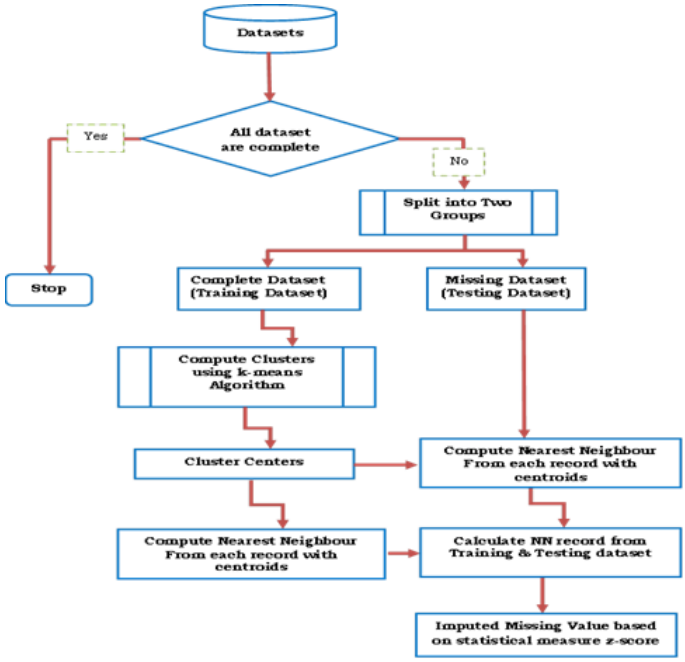

Figure 1. A New Paradigm for Data Imputation (NPDI)

$$M_{mn} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

Here matrix M represents the samples of $'m'$ observations and $'n'$ variables and each cell is denoted as ordered n-tuples of data records such as $(a_{i1}, a_{i2}, , a_{i(n-1)}, a_{in})$ for each $i = 1, 2, 3, m$ where the last column data attribute (a_{in}) for each i characterizes the target class or decision data attribute of the dataset M. Clearly, we say that all datasets are finite set. An object $'a_i'$ is known as the complete dataset, if $\{a_{ij} \neq \phi|, \forall 1 \leq j \leq R\}$ and the object $'a_i'$ is called the incomplete dataset $\{a_{ij} = \phi|, \exists 1 \leq j \leq R\}$. Now split the given dataset into two forms that is: i) first form is a complete dataset which contains records without missing values. ii) Second form is an incomplete dataset which contains missing records with one or more attributes which is known as the missing value dataset. Then consider complete dataset as the training set and missing data set as the testing set. This work is motivated by researchers Wei-Chao Lin [18], Congnan Luo [19].

The k-means clustering algorithm [20-21] is one of simplest and common clustering algorithms used to classify the number of clusters based on k-value. The value k is the cluster centroid of each cluster. This study step-1 we use a k-means clustering algorithm to form clusters based its decision classes (here k=2) and to extract the cluster centroids using the centroid vector $c_k$is defined as follows:

$$c_k = \frac{1}{|N|} \sum_{m=1}^{N} a_{im} \tag{1}$$

Then, calculate its nearest neighbours based on a Euclidian distance measure on without missing values of dataset. The nearest neighbour function defined as follows: Let S is a finite set of elements and a and b belongs to S. B is said to be a the nearest neighbour of A if B is closest to A among all the points in $S - A$. Then B is said to be the nearest neighbour of A if and only if

4. EXPERIMENTS AND RESULTS

In order to evaluate NPDI model, we carried out experimentions on the benchmark datasets from Knowledge Extraction based on Evolutionary Learning (KEEL) repository [22], includes Bands, Cleveland,

Hepatitis, Horse-Colic, Iris, Magic, Pima, Wine. The completely summary of datasets are used in the experiments are shown in Table.1.

Table 1. Summary of Datasets used in experiments [22].

Datasets	Instances	Classes	MVs	Databases(KB)
Bands	539	2	32.28	608
Cleveland	303	5	1.98	233
Hepatitis	155	2	48.39	129
Horse-Colic	368	2	98.1	444
Magic	1902	2	58.20	1423
Pima	768	2	50.65	440
Wine	178	3	70.22	136

First, we divided the given dataset into 10 folds of the equal size based on the stratified cross validation test and each step of the experiments 1-fold is used for the test dataset and remaining 9- folds are used for the training dataset.Also, this study presents the results obtained from the experimental evaluation of our proposed approach discussed in Section .3 and we took results for the comparison with other imputation methods such as Imputation with K-Nearest Neighbour (KNNI), Support Vector Machine Imputation (SVMI), Weighted Imputation with K-Nearest Neighbour (WKNNI), K-Means clustering Imputation (KMI) and Fuzzy-Means clustering Imputation (FKNNI) on a popular decision tree C4.5 classifier [23].

Table 2. Test classifier accuracy using C4.5 deceson tree

Methods	NPDI	KNNI	SVMI	W-KNNI	KMI	F-KNNI
Bands	71.15	70.32	69.18	69.57	70.11	68.25
Cleveland	57.32	56.09	55.41	56.09	55.43	56.09
Hepatitis	89.95	84.10	84.82	85.09	83.20	83.47
Horse-Colic	90.15	83.10	83.67	83.40	82.04	83.94
Magic	82.92	79.96	78.23	79.75	79.86	80.07
Pima	75.56	71.09	73.32	73.69	72.78	72.91
Wine	88.55	87.64	86.56	88.75	86.54	87.45

In Table 2 shows that evaluation results, from these results we clearly state that the proposed novel framework for data Imputation is higher than KNNI, SVMI, WKNNI, KMI and FKNNI. Then classifier accuracy is improved significantly compared with other imputation methods such as KNNI, SVMI, WKNNI, KMI and FKNNI on benchmark datasets using C4.5 decision tree, which proved the validity of novel framework for data Imputation. The predictions of classifier for the proposed approach is compared with other imputation algorithms are shown separately in Figure 2-6.

From the bar graphs, we can clearly state that the proposed NPDI has attained the best accuracy among all other seven benchmark datasets. Then we compared with other imputation methods such as KNNI, SVMI, WKNNI, KMI and FKNNI based on C4.5 decision tree classifiers have scored approximately 6.75 percentage improvement in terms of accuracy. .

5. CONCLUSION

This paper proposed a new paradigm for data imputation called NPDI for estimating the missing value in datasets. Firstly, split the given dataset into training and testing datasets that is complete dataset and incomplete dataset. Secondly, applied k-means algorithm on training dataset to generate clusters and its centroids, then calculated the nearest neighbours distance between centroids and other complete data records and missing data records. Finally, applied a popular statistical measure called z-score on mapping distances and then imputed plausible value for imputation. Further, we applied C4.5 decision tree classifier for test classifier accuracy. The classifier accuracy is improved significantly compared with other imputation methods such as KNNI, SVMI, WKNNI, KMI and FKNNI on benchmark datasets using C4.5 decision tree, which proved the validity of the novel framework for data Imputation. Approximately 6.75 percentage improvement in terms of classifier

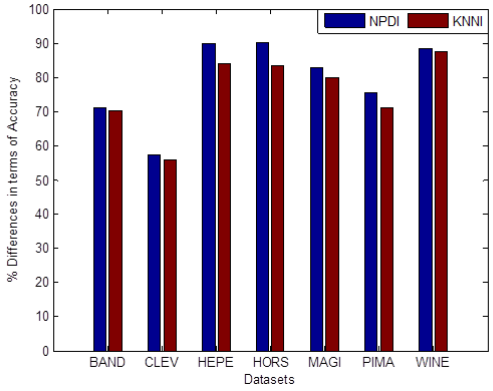

Figure 2. NPDI vs KNNI

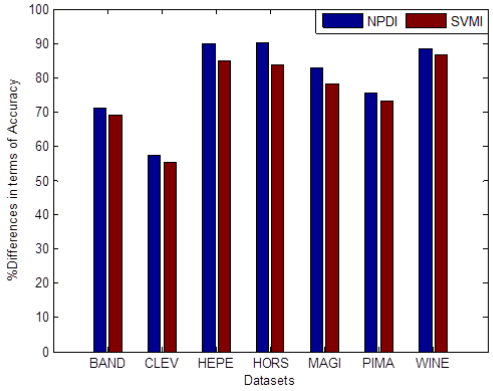

Figure 3. NPDI vs SVMI

accuracy with compared to KNNI algorithm, WKNNI algorithm for Hepatitis dataset and KMI algorithm for Horse-Colic dataset. Major strength of this approach is not lost the any attribute information during the dimensionality reduction process. Limitation of this research work is before imputations need suitable normalization technique to smooth dataset.

REFERENCES

[1] Jiawei Han, Micheline Kamber."Data Mining: Concepts and Techniques", 2nd edition. Morgan Kaufmann, San Francisco, CA, 2006.
[2] Mehmed Kantardzic et al., Data Mining: Concepts, Models, Methods, and Algorithms, 2nd edition, IEEE Press, 2011.
[3] G.Madhu et al., A Novel Index Measure Mimputation Algorithm for Missing Data Values: A Machine Learning Approach, IEEE International Conference on Computational Intelligence and Computing Research, pp.1-7, 2012.
[4] Dan Li et al., Towards Missing Data Imputation: A Study of Fuzzy k-Means Clustering Method, RSCTC 2004, LNAI 3066, pp. 573579, 2004.
[5] Rahman, M.M and Davis, D.N. "Machine Learning-based Missing Value Imputation Methods for Clinical datasets", IAENG Transactions on Engineering Technologies, Springer Netherlands, pp.245-257, 2013.
[6] Rupam Deb et al., "Missing Value Imputation for the Analysis of Incomplete Traffic Accident Data", Information Sciences, vol.339, pp.274-289, 2016.
[7] Bashir.F, et al., "Parametric and Non-Parametric Methods to Enhance Prediction Performance in the Presence of Missing Data", 19th International conference on system theory, control and computing (ICSTCC), pp.337-342, 2015.

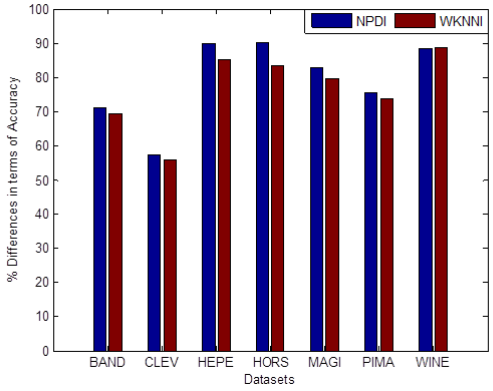

Figure 4. NPDI vs WKKNI

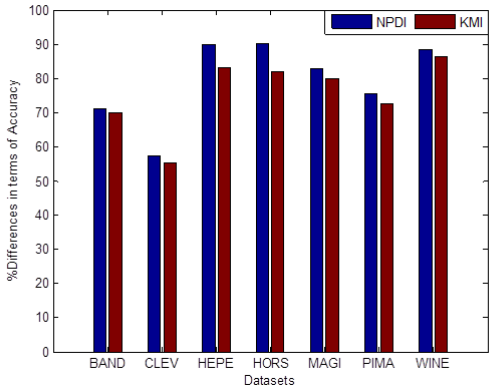

Figure 5. NPDI vs KMI

[8] Fabio Lobato, et al., "Multi-Objective Genetic Algorithm for Missing Data Imputation", pattern recognition letters, vol.68, part-1, pp.126-131, 2015.
[9] G. Madhu, et al., "A Non-Parametric Discretization Based Imputation Algorithm for Continuous Attributes with Missing Data Values", International Journal of Information Processing, vol.8, no.1, pp.64- 72, 2014.
[10] Chandan Gautam and Vadlamani Ravi, "Data Imputation via Evolutionary Computation, Clustering and a Neural Network", Neurocomput, 156, 134-142, 2015.
[11] Gary, K., et al., "Listwise Deletion is Evil: What to do about Missing Data in Political Science", (2000) (avaible http://GKing.Harvard.edu)
[12] Dempster, A.P., Laird, N.M., Rubin, D.B., "Maximum likelihood from incomplete data via the EM algorithm", J. of Royal Statistical Society Series, vol. 39,pp, 138, 1977.
[13] Myrtveit, I., et al., "Analyzing data sets with missing data: an empirical evaluation of imputation methods and likelihood-based methods", IEEE Transactions on Software Engineering, vol.27, pp. 9991013, 2001.
[14] Little, R.J., Rubin, D.B.,"Statistical Analysis with Missing Data", Wiley, New York, 1987.
[15] Alireza Farhangfar et al., "A Novel Framework for Imputation of Missing values in Databases", IEEE Transaction on systems, man, and cybernetics-part-a: systems and humans, vol.37, no.5, sept 2007.
[16] Lakshminarayan K, et al., "Imputation of Missing data in industrial databases", Appl, Intell, vol.11, no.3, pp. 259-275, 1999.
[17] H.L.Oh and F.L Scheuren, "Weighting adjustments for unit nonresponse, incomplete data in sample survey", Theory and Bibliographies, vol.2, pp.143-183, 1983.
[18] Wei-Chao Lin, Shih-Wen Ke, and Chih-Fong Tsai "CANN. An Intrusion Detection System based on Combining Cluster Centers and Nearest Neighbors, Know.-Based Syst", vol. 78, pp. 13-21, 2015..
[19] Congnan Luo, Yanjun Li, and Soon M. Chung "Text document clustering based on neighbors", Data

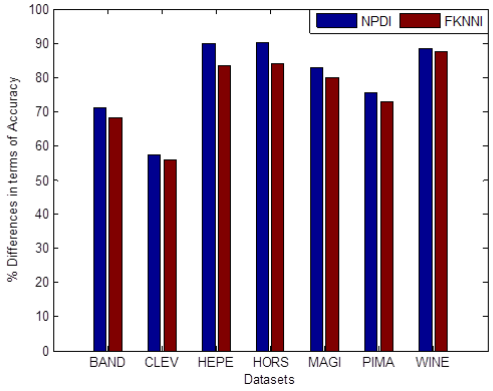

Figure 6. NPDI vs FKNNI

Knowl. Eng, vol.68, issue. 11, pp. 1271-1288, 2009.
[20] J.A. Hartigan, "Clustering Algorithms", John Wiley and Sons, 1975.
[21] A.K. Jain, M.N. Murty, P.J. Flynn, "Data clustering: a review", ACM Comput.Survey, vol.31 (3), pp. 264323, 1999.
[22] J. Alcal-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. Garca, L. Snchez, F. Herrera. "KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework", Journal of Multiple-Valued Logic and Soft Computing 17:2-3 (2011) 255-287.
[23] J.R. Quinlan, "C4.5: Programs for Machine Learning". Morgan Kaufmann Publishers, Inc., 1993.

BIOGRAPHIES OF AUTHORS

Madhu.G received Ph.D Degree in Compter Science and Engineering from Jawaharlal Nehru Technological University, Hyderabad in 2015. He is currently working as Assocaite Professor, Dept of Information Technology, VNRVJIET, JNTU Hyderabad, T.S, INDIA. He has 27 research publications at International/National Journals and Conferences. His research interest includes Data Mining, Machine Learning and Methematical Modeling and he is also a reviewer of research papers of various Journals like JCIT,JETAI and Journal of Soft Computing. He is a Member of IEEE, IEEE Computational Intelligence Society, Life Member of CSI,Member of IRSS, and Member of IAENG. Recently, machine learning based algorithms on malaria daignosis has been tackled.

G.Naga Chandrika received M.Tech Degree in Compter Science and Information Technology from Jawaharlal Nehru Technological University, Hyderabad. He is currently working as Assistant Professor, Dept of Information Technology, VNRVJIET, JNTU Hyderabad, T.S, INDIA. Her research interest includes Data Mining, Text Mining and Theory of Computation.