International Journal of Electrical Electronics & Computer Science Engineering
Volume 4, Issue 3 (June, 2017) | E-ISSN : 2348-2273 | P-ISSN : 2454-1222
Available Online at www.ijeecse.com

A Novel Index Measured Segmentation Based Imputation Algorithm (with Cross Folds) for Missing Data Imputation

Priyadharsini[1], Dr. Antony Selvadoss Thanamani[2]
[1]Ph.D Research Scholar, [2]Head & Associate Professor Department of CSE, NGM College, Pollachi Coimbatore, India
priyadharsini.ast@gmail.com

***Abstract:** With the rapid increase in the use of databases, missing data make up an important and unavoidable problem in data management and analysis. A most important task when pre-processing the data is, to fill in missing values, smooth out noise and correct inconsistencies. This paper presents the missing value problem in data mining and evaluates some of the methods generally used for missing value imputation. The new method that uses mathematical model for impute missing data. The novel A novel Index Measured segmentation based Imputation Algorithm (with cross folds) for missing data imputation was proposed in this paper. The databases were used to demonstrate the performance of the proposed method. The proposed algorithm is evaluated by extensive experiments and comparison with KNNI, SVMI. The results showed that the proposed algorithm has better performance than the existing imputation algorithms in terms of classification accuracy.*

I. Introduction

Missing values has long been an unavoidable problem that occurs to almost data-driven solutions. There are various causes such as incomplete data collection, data entry errors, incompetent data acquisition from experiments, and unfinished responses to a questionnaire [1]. This raises a significant problem towards data analysis, especially to those learning

Models that are compatible only with a complete data set. Over the past decades, Provision of innovative research aiming to fill in missing vales is continuously developed [2]. A rich collection of data pre-processing techniques has been made available, including zero imputation, average imputation, minimum imputation, maximum imputation, expectation maximization, linear regression imputation and k-nearest neighbours. Unlike the conventional approach that excludes any record with missing values, the aforementioned statistical and machine learning methods attempt to predict those with the values close to the original data. In this research the following supervised and unsupervised learning algorithms are compared with the proposed algorithm.

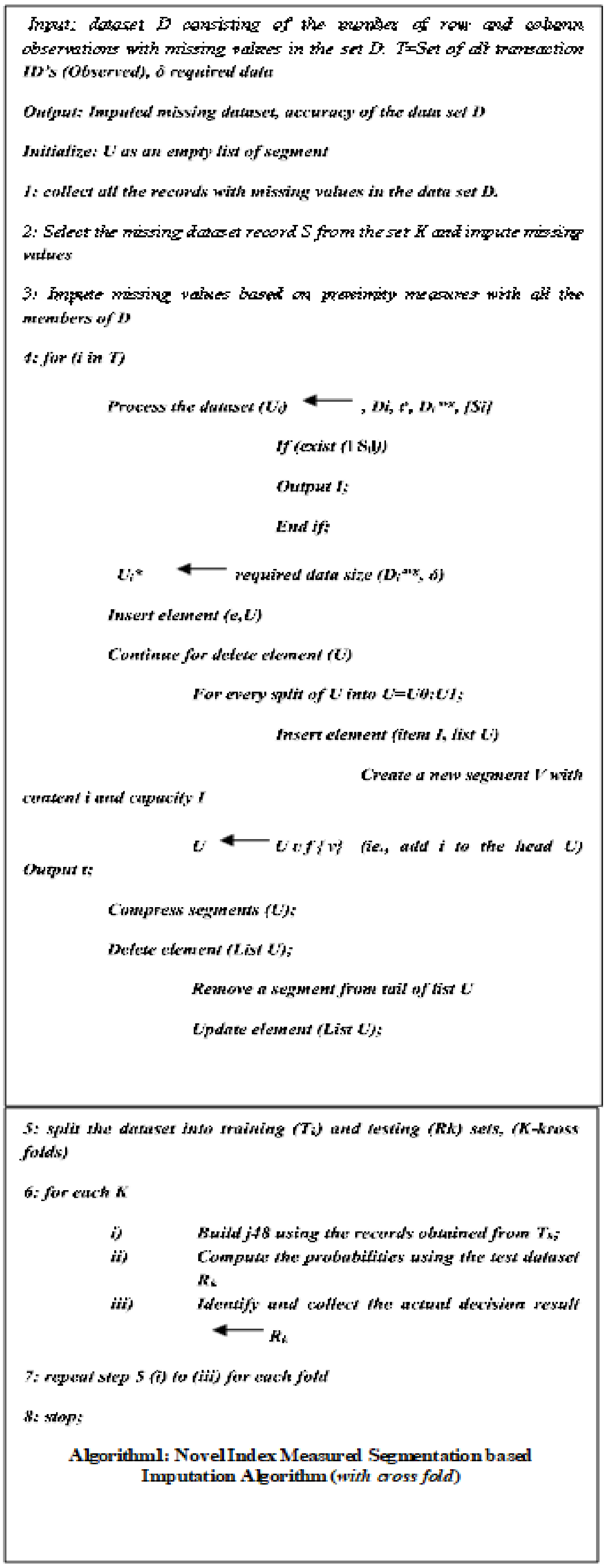
II. Literature Review

Missing data imputation techniques [4] to compute the missing value for the missing record or attribute and fill the estimated value from other reported values. Missing data imputation techniques are classified into two categories that is a) ignorable missing data imputation [6] b) non-ignorable missing data imputation [6]. In the literature many researchers have proposed missing data imputation techniques such as Regression imputation [7], Hot-Deck Imputation is a statistical method [8], Imputation with K-Nearest Neighbor (KNNI) [9], K-means Clustering Imputation (KMI) [10], Imputation with Fuzzy K-Means Clustering (FKMI)[11], Weighted imputation with K-Nearest Neighbor (WKNNI) [5], Support Vector Machines Imputation (SVMI) [12], Singular Value Decomposition Imputation (SVDI) [5], Bayesian Principal Component Analysis (BPCA) [13], Radial Basis Function Network, Event Covering technique[14] and RNI Algorithm [15]. In this research work, we proposed a non- parametric imputation strategy that can be applied to any data set be it nominal and/or categorical by employing an indexing measure for computing the similarity between the data records (n tuples) and also develop illustrative examples showcasing the application of this algorithm. A new indexing measure to the imputation algorithm for missing data values of the attributes to compute the similarity measure between any two typical elements in the dataset [3].

III. Methodology

In this section we present the novel A novel Index Measured segmentation based Imputation Algorithm (with cross folds) for missing data imputation, in which the information within the incomplete instance of the dataset, it can be applied to any type of data, be it categorical (nominal segmented data) and / or numeric (real or integer segmented data).

We describe a non-parametric imputation strategy demystified approach to compute the proximity measure in the feature space between the data record to identify the nearest neighbors from where the values are to be imported. The algorithm follows,



```
Input: dataset D consisting of the number of row and column observations with missing values in the set D. T=Set of all transaction ID's (Observed), δ required data

Output: Imputed missing dataset, accuracy of the data set D

Initialize: U as an empty list of segment

1: collect all the records with missing values in the data set D.

2: Select the missing dataset record S from the set K and impute missing values

3: Impute missing values based on proximity measures with all the members of D

4: for (i in T)

          Process the dataset (Ui) ←  , Di, e', $D_i^{mis}$, {Si}

                        If (exist (I Si|))

                        Output I;

                        End if;

           Ui*  ←  required data size ($D_i^{mis}$, δ)

          Insert element (e,U)

          Continue for delete element (U)

                  For every split of U into U=U0:U1;

                        Insert element (item I, list U)

                              Create a new segment V with content i and capacity I

                  U  ←  U ∪ f { v}  (ie., add i to the head U)
Output t;

          Compress segments (U);

          Delete element (List U);

                  Remove a segment from tail of list U

                  Update element (List U);

5: split the dataset into training (Tk) and testing (Rk) sets, (K-kross folds)

6: for each K

          i)        Build j48 using the records obtained from Tk;
          ii)       Compute the probabilities using the test dataset Rk
          iii)      Identify and collect the actual decision result
                      ←  Rk

7: repeat step 5 (i) to (iii) for each fold

8: stop;
```

Algorithm1: Novel Index Measured Segmentation based Imputation Algorithm (*with cross fold*)

IV. Evaluation And Results

In this section we present our study and the classification accuracies are presented in Table 1 describes a dataset and Table 2 describes a performance. A novel Index Measured segmentation based Imputation Algorithm (with cross folds) is also compared with other algorithms (KNNI, SVMI) on the real valued datasets and categorical data sets.

Datasets	Records	Attributes
Iris	150	5
E. coli	700	7

Table 1: Datasets Used For the Experiment

Dataset	KNNI	SVMI	NIMSI
Iris	70.60	70.90	71.40
E. coli	80.96	81.37	83.46

Table 1: Test Accuracies of J48 Decision Tree Classifiers

Finally Fig.1 shows that the real values datasets accuracy with A novel Index Measured segmentation based Imputation Algorithm (with cross folds). Thus we conclude that our algorithm is the best approach to imputing the missing values, as they led to the statistically significant improvements in prediction accuracy. Thus the present results might generalize to different types of data sets (nominal and/or numeric).

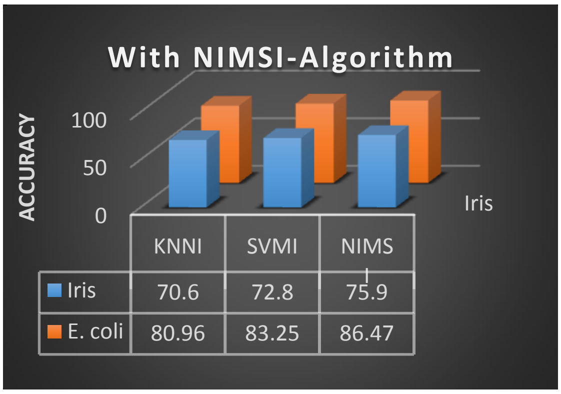


Fig.1 Accuracy on Real Value Datasets with NIMSI-Algorithm

V. Conclusion

Missing values are very prominent in a real world database. In this paper, A novel Index Measured segmentation based Imputation Algorithm (with cross folds) of missing values is discussed, that aims to improve in terms of classification accuracy. We compared A novel Index Measured segmentation based Imputation Algorithm (with cross folds) with the state-of-the art methodologies of real world imputation algorithms on categorical and real values of benchmark datasets. We conclude that the use of our A novel Index Measured segmentation based Imputation Algorithm (with cross folds) strategy improved the accuracies of the predictions on real world missing data value problems.

VI. REFERENCE

[1] C. Gautam and V. Ravi. Evolving clustering based data imputation. In Proceeding of International Conference on Circuit, Power and Computing Technologies, pages 1763–1769, 2014.

[2] Brock G, Shaffer J, Blakesley R, Lotz M, Tseng G (2008) Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. BMC Bioinformatics 9: 1-12.

[3] G.Madhu1, T.V.Rajinikanth2 "A Novel Index Measure Imputation Algorithm for Missing Data Values: A Machine Learning Approach "©2012 IEEE

[4] Friedman, N. "The Bayesian structural EM algorithm", In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. pp. 129–138, 1998.

[5] Troyanskaya, O., Cantor, M., Sherlock, G., et al. "Missing value estimation methods for DNA microarrays", Bioinformatics, vol.17, no.6, pp.520–525, Jun 2001.

[6] Rubin, D.B., "The design of a general and flexible system for handling non-response in sample surveys", Report prepared for the U.S Social Security Administration, 1997.

[7] Little, R. J. A., "Modeling the Drop-Out Mechanism in Repeated- Measures Studies". Journal of the American Statistical Association, vol.90, pp.1112-1121, 1995.

[8] Frane, J.W., "Some simple procedures for handling missing data in multivariate analysis", Psychometrika, vol. 41, pp.409–415, 1976.

[9] J.N. K. Rao J. Shao. "Jackknife variance estimation with survey data under hot deck imputation", Biometrika, vol.79, no.4, pp. 811 – 822, 1992.

[10] G.E.A.P.A. Batista M.C. Monard, "An analysis of four Missing Data treatment methods for supervised learning", Applied Artificial Intelligence, vol.17, pp. 519-533, 2003.

[11] Li D, Deogun J, SpauldingW, Shuart B., "Towards missing data imputation: a study of fuzzy k-means clustering method", In Proceedings of 4th international conference of rough sets and current trends in computing (RSCTC), pp.573–579, 2004.

[12] Acuna E, Rodriguez C., "The treatment of missing values and its effect in the classifier accuracy", Classification, Clustering and Data Mining Applications, Springer, Berlin, pp.639–648, 2004.

[13] H. Feng, C. Guoshun, Y. Cheng, B. Yang, Y. Chen, "A SVM Regression Based Approach to Filling in Missing Values", in Proc. KES ,vol.3, pp.581-587. 2005.

[14] Shigeyuki Oba, Masa-aki Sato, et al., "A Bayesian missing value estimation method for gene expression profile data", Bioinformatics, vol.19, no.16, pp.20882096, 2003.

[15] Luengo J, García S, Herrera F., "A study on the use of imputation methods for experimentation with Radial Basis Function Network classifiers handling missing attribute values: the good synergy between RBFNs and Event Covering method", Neural Nets,vol.23, no.3, pp. 406–418, 2010.

[16] Sree Hari Rao. V, Naresh Kumar. M, "A new intelligence based approach for computer-aided diagnosis of dengue Fever", IEEE Transactions on Information Technology in Biomedicine , vol.16, no.1, pp.112-118, 2012.