iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 Issue: III Month of publication: March 2019

DOI: http://doi.org/10.22214/ijraset.2019.3390

Transmogrified Imputation Algorithm for Clustering Data in Missing Data Imputation

A. Linda Sherin[1], Dr. S.Niraimathi[2]
[1]Ph.D. Research Scholar Full Time, [2]Associate Professor , Dept. of Computer Science, NGM College, Tamilnadu, India

Abstract: This research article proposes to implement transmogrification process in the Imputation procedures to overcome the challenges in missing values. Appropriate data pre-processing methods and clustering mechanisms in data mining plays significant role to ensure good quality of data. The data pre-processing tasks include identification of outliers, smoothening noisy data and overcoming inconsistent data. Issues related to data incompleteness, still remains a challenge to many researchers. The transmogrified method uses mathematical approach and cluster based Imputation Algorithm for missing data imputation. The IBM Log data set and Sonar data set were used to demonstrate the performance of the proposed method. The proposed algorithm is evaluated by extensive experiments and comparison with KNNI - Imputation with K-Nearest Neighbor, MSC- Imputation with Mean-shift Clustering, AHC- Agglomerative Hierarchical clustering, EM-GMM- Expectation – Maximization Clustering using Gaussian Mixture Models and Naïve Bayesian Model. The results showed that the proposed algorithm has better performance than the existing imputation algorithms in terms of classification accuracy.

Keywords: KNNI - Imputation with K-Nearest Neighbor, MSC- Imputation with Mean-shift Clustering, AHC- Agglomerative Hierarchical clustering, EM-GMM- Expectation – Maximization Clustering using Gaussian Mixture Models and Naïve Bayesian Model

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

Past Literature pertaining to Missing data imputation techniques to compute the missing value for the missing record or attribute and fill the estimated value from other reported values were surveyed. In review of literature Missing data imputation techniques are classified as ignorable missing data imputation and non-ignorable missing data imputation. In the literature many researchers have proposed missing data imputation techniques such as Cold-Deck Imputation, Imputation with K-Nearest Neighbor, K-means Clustering Imputation, Imputation with Fuzzy K-Means Clustering, imputation with Agglomerative Hierarchical clustering, Imputation with Mean-shift Clustering, Naïve Bayesian Imputation, Bolzano–Weierstrass Classifier Imputation and Expectation – Maximization Clustering using Gaussian Mixture Models Algorithm. Little and Rubin summarize the mechanism of imputation method. Also introduces mean imputation method to find out missing values. The drawbacks of mean imputation are sample size is overestimated, variance is underestimated, correlation is negatively biased. For median and standard deviation also replacing all missing records with a single value will deflate the variance and artificially inflate the significance of any statistical tests based on it. Different types of machine learning techniques are supervised and unsupervised machine learning techniques summarized.

Classification of multiple imputation and experimental analysis are described. Min Pan et al. summarize the new concept of machine learning techniques like NBI also analysis the experimental results which impute missing values. Comparisons of different unsupervised machine learning technique are referred from survey paper. To overcome the unsupervised problem Peng

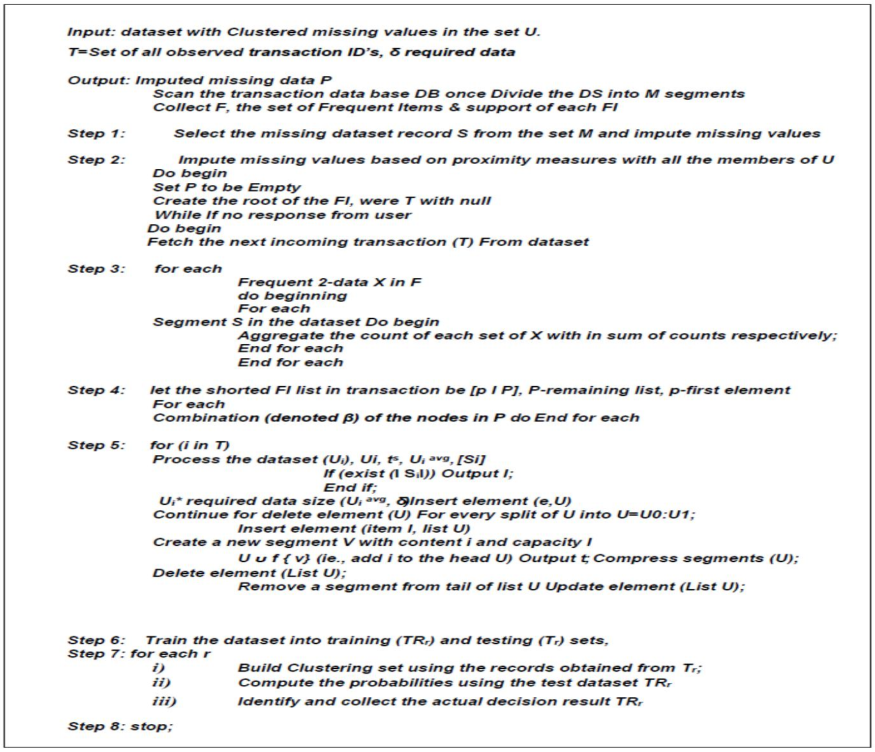
Liu, Lei Lei et al. applied the supervised machine learning techniques called Naïve Bayesian Classifier. In the data mining context, machine learning technique is generally classified as supervised and unsupervised learning technique both belong to machine learning technique. Supervised classification focus on the prediction based on known properties and the classification of unsupervised focus on commonly used classification algorithm known as Naïve Bayesian imputation techniques.

Mean Imputation is the process of replacing the missing data from the available data where the instance with missing attribute belongs. Median Imputation is calculated by grouping up of data and finding average for the data. Median can be calculated by finding difference between upper and lower class boundaries of median class. Standard Deviation calculate the scatter data concerning the mean value. It can be convenient in estimating the set of fact which can possess the identical aim but a different domain. Estimate standard deviation based on sample and entire population data. Another way of learning technique is classified as supervised learning that focus on the prediction based on known properties.

Naïve Bayes technique is one of the most useful machine learning techniques based on computing probabilities. It analyses relationship between each independent variable and the dependent variable to derive a conditional probability for each relationship. A prediction is made by combining the effects of the independent variables on the dependent variable which is the outcome that is predicted. It requires only one pass through the training set to generate a classification model, which makes it very efficient. The Naïve Bayesian generates data model which consists of set of conditional probabilities, and works only with discrete data.

III. METHODOLOGY

In this article Transmogrification of Imputation Algorithm for Clustering of Data is dealt with novel for missing data imputation, the transmogrified method uses mathematical approach and Index segmentation based Imputation Algorithm for missing data imputation. The databases were used to demonstrate the performance of the proposed method. The proposed algorithm is evaluated by extensive experiments and comparison with KNNI - Imputation with K-Nearest Neighbor, MSC- Imputation with Mean-shift Clustering, AHC- Agglomerative Hierarchical clustering, EM-GMM- Expectation – Maximization Clustering using Gaussian Mixture Models and Naïve Bayesian Model. An imputation strategy Transmogrified approach is described to compute the proximity measure in the feature missing space between the missing data to identify the nearest neighbor missing data from where the values are to be imputed.



```
Input: dataset with Clustered missing values in the set U.
T=Set of all observed transaction ID's, δ required data

Output: Imputed missing data P
            Scan the transaction data base DB once Divide the DS into M segments
            Collect F, the set of Frequent Items & support of each FI

Step 1:        Select the missing dataset record S from the set M and impute missing values

Step 2:        Impute missing values based on proximity measures with all the members of U
            Do begin
            Set P to be Empty
            Create the root of the FI, were T with null
            While If no response from user
           Do begin
           Fetch the next incoming transaction (T) From dataset

Step 3:     for each
                        Frequent 2-data X in F
                        do beginning
                        For each
            Segment S in the dataset Do begin
                        Aggregate the count of each set of X with in sum of counts respectively;
                        End for each
                        End for each

Step 4:     let the shorted FI list in transaction be [p I P], P-remaining list, p-first element
            For each
            Combination (denoted β) of the nodes in P do End for each

Step 5:     for (i in T)
            Process the dataset ($U_i$), Ui, $t^s$, $U_i^{avg}$, [Si]
                                    If (exist (l $S_i$l)) Output I;
                                    End if;
             $U_i$* required data size ($U_i^{avg}$, $\delta$)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 ∪ 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);


Step 6:    Train the dataset into training ($TR_r$) and testing ($T_r$) sets,
Step 7: for each r
            i)          Build Clustering set using the records obtained from $T_r$;
            ii)         Compute the probabilities using the test dataset $TR_r$
            iii)        Identify and collect the actual decision result $TR_r$

Step 8: stop;
```

Transmogrified Imputation Algorithm for clustering data in Missing Data Imputation

IV. EVALUATION AND RESULTS

In this section we present the experimental evaluation for a Transmogrified Imputation Algorithm (TIA) for clustering data in Missing Data Imputation using IBM Log data set and Sonar data set. The table below reveals the test accuracies of Transmogrified clusters obtained through TIA and normal clusters. The Transmogrified Imputation Algorithm TIA is compared with other algorithms with KNNI - Imputation with K-Nearest Neighbor, MSC- Imputation with Mean-shift Clustering, AHC- Agglomerative Hierarchical clustering, EM-GMM- Expectation – Maximization Clustering using Gaussian Mixture Models and Naïve Bayesian Model.

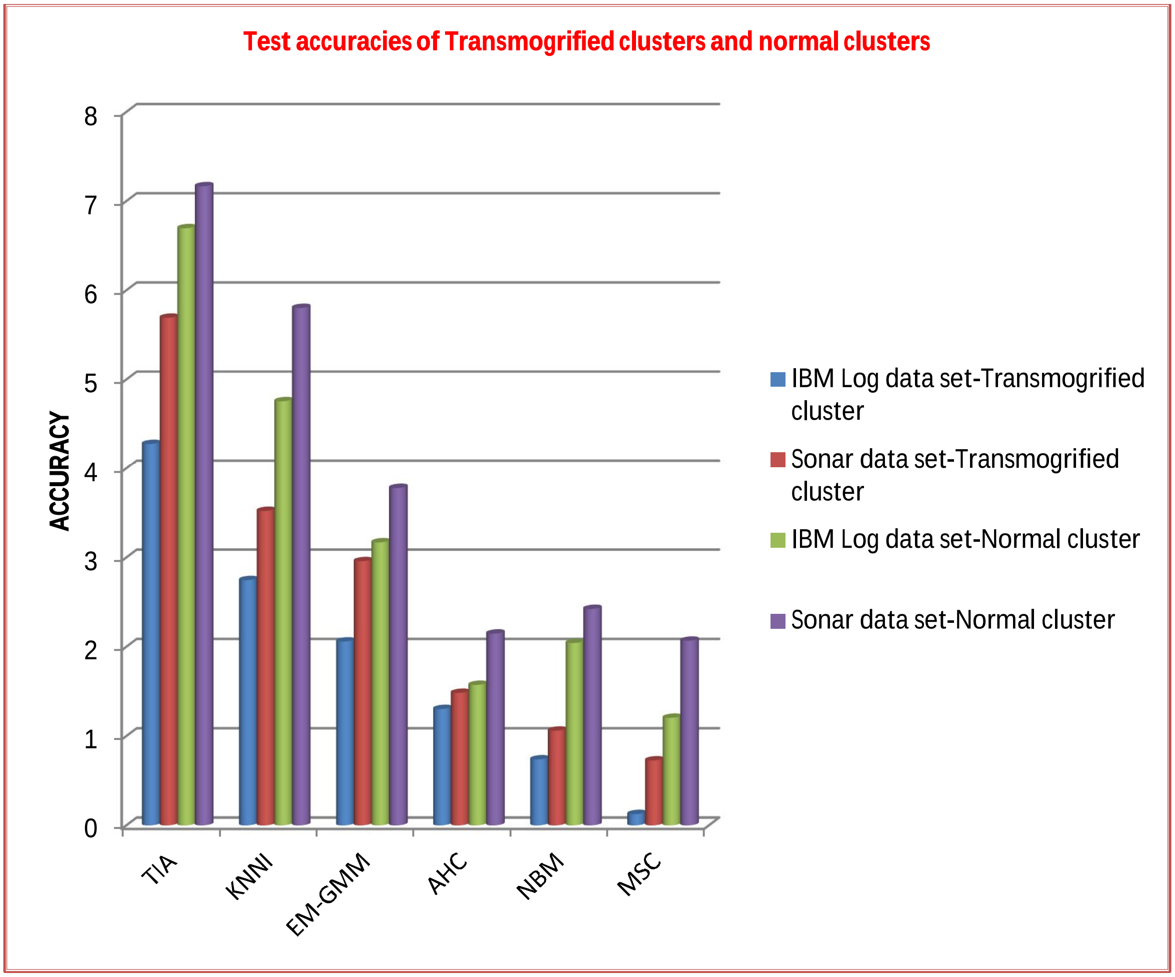
Datasets	**Records**	**Attributes**
IBM Log data set	56865	182
Sonar data set	32578	45

Datasets Used For the Experiment

Dataset	TIA	KNNI	EM-GMM	AHC	NBM	MSC
IBM Log data set-Normal cluster	4.276185	2.7464234	2.058584	1.3027783	0.7388564	0.126606
Sonar data set-Normal cluster	5.691732	3.5210146	2.957915	1.4831539	1.06036	0.726867
IBM Log data set-Transmogrified cluster	6.696001	4.7558449	3.169164	1.5732506	2.042097	1.204268
Sonar data set-Transmogrified cluster	7.166716	5.8019504	3.778248	2.1474655	2.423084	2.067616

Table showing the Test accuracies of Transmogrified clusters and normal clusters

The chart below clearly depicts that Transmogrified Imputation Algorithm (TIA) for clustering data in missing data imputation using IBM Log data set and Sonar data set yields more test accuracies when compared to KNNI - Imputation with K-Nearest Neighbor, MSC- Imputation with Mean-shift Clustering, AHC- Agglomerative Hierarchical clustering, EM-GMM- Expectation – Maximization Clustering using Gaussian Mixture Models and Naïve Bayesian Model.



Test Accuracies of Transmogrified Clusters and Normal Clusters using IBM Log data set and Sonar Data set

V. CONCLUSION

In this article, Transmogrified Imputation Algorithm for Clustering of Data in Missing Data (TIA) is described. It is an Improved novel Clustering Algorithm where Transmogrification of Data based Imputation Algorithm of missing values is discussed, that aims to improve in terms of accuracy. The test accuracies of TIA were compared with KNNI - Imputation with K-Nearest Neighbor, MSC- Imputation with Mean-shift Clustering, AHC- Agglomerative Hierarchical clustering, EM-GMM- Expectation – Maximization Clustering using Gaussian Mixture Models and Naïve Bayesian Model using two different data sets IBM Log file data set and Sonar data set. We conclude that the use of our Transmogrified Imputation Algorithm for Clustering of Data in Missing Data improved the accuracies of the predictions on real world missing data value problems.

REFERENCES

[1] Alexander J Stimpson and Mary L. Cummings (2016), "Assessing Intervention Timing in Computer-Based Education Using Machine Learning Algorithms", IEEE Access, Volume 2.

[2] AlirezaFarhangfar, Lukasz A.Kurgan (2007), "A Novel Framework For Imputation Of Missing Values In Databases", IEEE Transactions On Systems Man And Cybernetics Part A: Systems And Humans, Vol. 37.

[3] Anton Akusok (2016), "Extreme Learning Machines: Novel Extensions and Application to Big Data", springer.

[4] Archana Purwar, Sandeep Kumar Singh (2015), "Hybrid Prediction Model with Missing Value Imputation for Medical Data", Expert Systems with Applications, Doi: Http://Dx.Doi.Org/10.1016/J.Eswa.2015.02.050.

[5] Biao Qin, Yuni Xia, Fang Li (2010), "A Bayesian Classifier for Uncertain Data", Proceedings of the ACM Symposium on Applied Computing, Pages 1010-1014, March 22 – 26.

[6] Bing Zhu, Changzheng He, PanosLiatsis (2012), "A Robust Missing Value Imputation Method for Noisy Data", Published Online: 28 July 2010 Springer Science Business Media, Llc, ApplIntell 36 Pg no: 61–74 Doi 10.1007/S10489-010-0244-1.

[7] Fabio Lobato, Claudomiro Sales, Igor Araujo, Vincent Tadaiesky, Lilian Dias, Leonardo Ramos, Adamo Santana (2015), “Multi-Objective Genetic Algorithm for Missing Data Imputation”, Pg no: S0167-8655(15) 00288-Doi: 10.1016/J.Patrec.2015.08.023 Reference: PATREC 6335, Received Date: 10 February 2015,

[8] G.Madhu, T.V.Rajinikanth (2012), “A Novel Index Measure Imputation Algorithm For Missing Data Values: A Machine Learning Approach”, Computational Intelligence & Computing Research (ICCIC), IEEE International Conference, Doi: 10.1109/Iccic.2012.6510198.

[9] Kuen-Fang, Jea, Chao-Wei Li, Chih-Wei Hsu, Ru-Ping Lin (2010), “A Load -Controllable Mining System for Frequent-Pattern Discovery in Dynamic Data Streams”, Machine Learning and Cybernetics (ICMLC), International Conference On (Volume: 5), ISBN No. 978-1- 4244-6526-2.

[10] Liqiang Pan, Jianzhong Li (2010), “K-Nearest Neighbor Based Missing Data Estimation Algorithm In Wireless Sensor Networks”, Wireless Sensor Network, Vol- 2, Pg.no 115-122 Doi:10.4236/Wsn.2010.22016.

[11] Matthew Eric Otey, Chao Wang, Srinivasan Parthasarathy, Adriano Veloso, Wagner Meira Jr (2003)”, Mining Frequent Itemsets In Distributed and Dynamic Databases,” IEEE Int'l Conf. On Data Mining.

[12] Mehran Amiria, Richard Jensen (2016), “Missing Data Imputation Using Fuzzy-Rough Methods”, Article in Neuro computing 205, Doi: 10.1016/J.Neucom.2016.04.015.

[13] Nan Jiang (2007), “Selective Integration of Linguistic Knowledge in Adult Second Language Learning”, Selective Integration of Linguistic Knowledge.

[14] P. Vasudevan (2014), “Iterative Dichotomiser-3 Algorithm in Data Mining Applied to Diabetes Database” Journal of Computer Science Vol-10 (7) Pg.no: 1151-1155, ISSN: 1549-3636 © Science Publications Doi:10.3844/Jcssp.2014.1151.1155 (Http://Www.Thescipub.Com/Jcs.Toc).

[15] P.Saravanan, P.Sailakshmi (2015), “Missing Value Imputation Using Fuzzy Possibilistic C Means Optimized With Support Vector Regression And Genetic Algorithm”, Journal Of Theoretical And Applied Information Technology. Vol.72.

[16] Priyadharsini.C, Dr. Antony SelvadossThanamani (2014), “Prediction of Missing Values in Blood Cancer & Occurrence of Cancer Using Improved Id3 Algorithm” International Journal Of Innovative Research In Computer And Communication Engineering (An Iso 3297, Certified Organization) Vol. 2, Issue 8.

[17] Vincent, S.Tseng, Chun-Jung Chu, Tyne Liang (2006), “Efficient Mining Of Temporal High Utility Itemsets From Data Streams”, Copyright ACM 1-59593-440-5/06/0008.

[18] W.G. Teng, M.S. Chen, P.S. Yu (2003)”, A Regression-Based Temporal Pattern Mining Scheme for Data Streams”, Proceedings of the 29th VLDB Conference Pg no. 93–104.