International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue III, Mar 2019- Available at www.ijraset.com

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, Dept. of Computer Science, NGM College, Tamilnadu, India*
[2]*Department of Computer Science NGM College, Tamilnadu, India*

Abstract: ***This research proposes to implement transmogrification process in the Imputation procedures to overcome the challenges in missing values. Appropriate data pre-processing methods 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 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, MSC, AHC, EM-GMM and FEM-GMM The results showed that the proposed algorithm has better performance than the existing imputation algorithms in terms of classification accuracy.***

Keywords: k-Nearest neighbor, Mean-shift Clustering (MSC), Naïve Bayesian Imputation and Expectation – Maximization Clustering, Gaussian Mixture

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 (KNNI), K-means Clustering Imputation (EM-GMM), Imputation with Fuzzy K-Means Clustering, imputation with Agglomerative Hierarchical clustering (AHC), Imputation with Mean-shift Clustering (MSC), Naïve Bayesian Imputation and Expectation – Maximization Clustering using Gaussian Mixture Models (EM-GMM) Algorithm.

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 (I S_iI)) Output I;
End if;
U_i required data size (U_i^{avg}, 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 ∪ 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;

IV. EVALUATION AND RESULTS

In this section we present An Improved novel Index Measured segmentation based Imputation Algorithm for missing data 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. An Improved novel Index Measured segmentation based Imputation
Algorithm (with cross folds) is also compared with other algorithms (KNNI, MSC, AHC, EM-GMM, NBM and FEM-GMM) on the real valued datasets and categorical data sets.

Table 1: Datasets Used For the Experiment

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

Table 1: Test accuracies of Transmogrified clusters and normal clusters

Dataset	KNNI	MSC	EM-GMM	AHC	FEM-GMM	NBM
IBM Log data set – Transmogrified cluster	60.64	64.90	66.78	70.45	74.54	78.40
Sonar data set-	80.96	81.37	84.28	87.89	90.52	93.85

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).

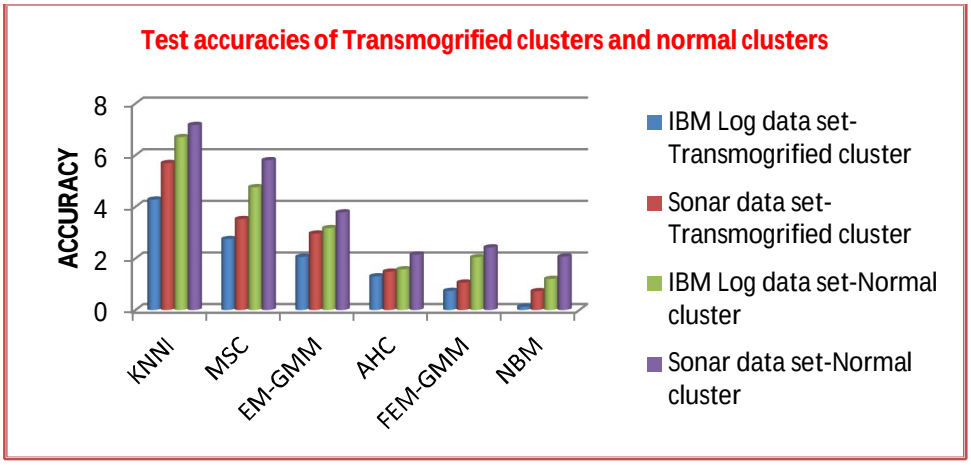

Fig.1 Accuracy on real value datasets with INMSI-Algorithm

V. CONCLUSION

Missing values are very prominent in a real world database. In this article, Transmogrified Imputation Algorithm for Clustering of Data in Missing Data 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 Transmogrified clusters and Normal clusters were compared using two different data sets IBM Log file data set and Sonar data set, 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 Transmogrified Imputation Algorithm for Clustering of Data in Missing Data improved the accuracies of the predictions on real world missing data value problems.

REFERENCE

[1] 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.
[2] Alireza Farhangfar, Lukasz Kurgan and Witold Pedrycz, "Experimental Analysis of Methods for Imputation of Missing Values in Databases.
[3] Apostol, Tom M.: Mathematical Analysis: A Modern Approach to Advanced Calculus, 2nd edition, Addison-Wesley Longman, Inc. (1974), page 112.
[4] Blessie, C.E., Karthikeyan, E, Selvaraj.B. (2010): NAD – A Discretization approach for improving interdependency, Journal of Advanced Research in Computer Science, ,pp.9-17.
[5] 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.
[6] Ciesielski, Krzysztof (1997). Set theory for the working mathematician. London Mathematical Society Student Texts. **39**. Cambridge: Cambridge University Press. pp. 106–111. ISBN 0-521-59441-3. Zbl 0938.03067.
[7] Damian Dechev, Pierre Laborde, and Steven D. Feldman, "LC-DC: Lockless Containers and Data Concurrency A Novel Nonblocking Container Library for Multicore Applications" IEEE Access Practical Innovations: Open Solutions Vol. 1, 2013
[8] E.Chandra Blessie, DR.E.Karthikeyan and DR.V.Thavavel, "Improving Classifier Performance by Imputing Missing Values using Discretization Method", International Journal of Engineering Science and Technology.
[9] G.Madhu1, T.V.Rajinikanth2 "A Novel Index Measure Imputation Algorithm for Missing Data Values: A Machine Learning Approach "©2012 IEEE
[10] 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.
[11] Ingunn Myrtveit, Erik Stensrud, "IEEE Transactions on Software Engineering", Vol. 27, No 11, November 2001.
[12] 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.
[13] Jeffrey C.Wayman, "Multiple Imputation for Missing Data: What is it and How Can I Use It?" Paper presented at the 2003 Annual Meeting of the American Educational Research Association, Chicago, IL, pp.2-16, 2003.
[14] Kamakshi Lakshminarayan, Steven A. Harp, Robert Goldman and Tariq Samad, "Imputation of Missing Data Using Machine Learning Techinques", from KDD-96 Proceedings.
[15] 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.
[16] Lim Eng Aik and Zarita Zainuddin, "A Comparative Study of Missing Value Estimation Methods: Which Method Performs Better?" 2008 International Conference on Electronic Design.
[17] 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.
[18] 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.
[19] Peng Liu, Lei Lei, "Missing Data Treatment Methods and NBI Model", Sixth International Conference on Intelligent Systems Design and Applications, 0-7695-2528-8/06.
[20] R. Kavitha Kumar and Dr. R. M. Chandrasekar, "Missing Data Imputation in Cardiac data set".
[21] R. Malarvizhi, Dr. Antony Selvadoss Thanamani, "K-Nearest Neighbor in Missing Data Imputation", International Journal of Engineering Research and Development, Volume 5 Issue 1-November-2015.
[22] R.J. Little and D. B. Rubin. Statistical Analysis with missing Data, John Wiley and Sons, New York, 2007.
[23] R.S. Somasundaram, R. Nedunchezhian, "Evaluation on Three simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values", International Journal of Computer Applications, Vol21-No. 10, May 2017, pp14-19.
[24] 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.
[25] S.Hichao Zhang, Jilian Zhang, Xiaofeng Zhu, Yongsong Qin, Chengqi Zhang, "Missing Value Imputation Based on Data Clustering", Springer-Verlag Berlin, Heidelberg,2016.
[26] Shen Qiping, GuoJianfeng, Zhang Jianping, and Liu Guiwen, "Using Data Mining Techniques to Support Value Management Workshops in Construction" TSINGHUA Science and Technology ISSN 1007-0214 13/20 pp 191-201 Vol. 13, No. 2, April 2016
[27] Shichao Zhang, Xindong Wu, Manlong Zhu, "Efficient Missing Data Imputation for Supervised Learning" Proc, 9th IEEE conference on Cognitive informatics, 2016 IEEE.
[28] 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.
[29] 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.
[30] 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.