iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** VI **Month of publication:** June 2019

DOI: http://doi.org/10.22214/ijraset.2019.6046

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VI, June 2019- Available at www.ijraset.com

An Efficient Imputation Approach on Scanty Data Using Bernoulli scheme based Markov Classifier

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

Abstract: ***Scanty and unruly data are pervasive and challenging issue in every information data set. It is diagnosed that using multiple imputation for missing information in multiple datasets acts as a linchpin on several unsupervised ML algorithms like standard deviation, mean, median and Supervised ML techniques for probabilistic algorithm. Such like probing is carried out employing an inclusive range of databases, for which missing fields are first filled in by various sets of tenable values to create multiple completed datasets, and then standard complete- data operations are applied to each completed dataset, and finally the multiple sets of results combine to generate a single inference. This research proposes to implement Markov Classifier using Bernoulli scheme in the Imputation procedures to overcome the challenges in scanty data. This research article offers general guidelines for selection of suitable data imputation algorithms based on characteristics of the data. Implementing Bernoulli scheme theorem in Markov chain and to produced an interesting model based on the Markov Classifier to assess the optimal output in a Random Forest algorithm, which has a monotonic subsequence and specify every sequence always has a feasible result. For estimate imputation of missing data, the standard machine learning repository dataset has been applied. Experimental analyses reveal that Markov Classifier gives better and superior classification along with Random forest for multiple imputations.***

Keywords: Stochastic process, Markov chain process, Markov Classifier, Bernoulli scheme, Maximum Likelihood, Scanty data, Random Forest, Unruly data Classification, Supervised ML, Unsupervised ML.

I. INTRODUCTION

Missing data refers to a class of problems made difficult by the absence of some portions of a familiar data structure. For example, a regression problem might have some missing values in the predictor vectors. This article concerns non-parametric approaches to assessing the accuracy of an estimator in a missing data situation. Missing data imputation techniques are classified as ignorable missing data imputation and non-ignorable missing data imputation. Earlier research 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 reveal that 5% to 8 % of the Missing data imputed were not accurate. Missing values may generate bias and affect the calibre of the supervised learning procedure. Missing value imputation is an efficient means to detect or estimate the missing values based on other data in the data sets. Most of the real world datasets are characterized by an unavoidable problem of incompleteness, in terms of dropping values. Data mining consists of the various technical approaches including machine learning, statistic and database system. The main goal of the data mining process is to discover knowledge from large database and transform into a human understandable format.

This paper focuses on several algorithms such as missing data mechanisms, multiple imputation techniques and supervised machine learning algorithm. Experimental results are separately imputed in each real datasets and checked for accuracy. A simple technique for handling with lost value is to bring forward all the values for any pattern removed one or more info items. Especially this is applicable although the database content be smaller to attain momentous outcome in the study. In parallel casing further sampling item sets can be collected. The mentioned issue hold enormous data sets that might be noticeable. As an illustration assuming that an application along 5 queries is about lost 10% of the item sets, later on moderate almost 60% of the sampling may obtain at minimum one query might be missing. These characteristics might be quite relevant to the analysis. The mechanism causing the missing data can influence the performance of both imputation and complete data methods. There are three different ways to categorize missing. Missing Completely At Random (MCAR) point into several distinct data sets being removed are separate both of noticeable scalar and of unnoticeable argument. Missing At Random (MAR) is the alternative, suggesting that what caused the data to be missing does not depend upon the missing data itself. Not Missing At Random (NMAR) is the quantities or characters or symbols removed as a precise reasoning.

II. PRELUDE OF THE STUDY

Previous studies regarding 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 Neighbour, K-means Clustering Imputation, Imputation with Fuzzy K-Means Clustering, imputation with Agglomerative Hierarchical clustering, Imputation with Mean-shift Clustering, Naïve Bayesian Imputation and Expectation – Maximization Clustering using Gaussian Mixture Models Algorithm.

III. RESEARCH THINKING

In this section, we provide a step-wise description of our methodology. Technical details are given in Section 4. Assume that we are given two suites of traces T and T_{an}. Recall that in our case a trace is simply a dataset with sequence of records that may have both normal and abnormal data. The suite T consists of traces of normal data and T_{an} consists of traces of anomalous data (presumably corresponding to some known attacks).

1) *Step 1 (Set up of the test suite):* In this step we split the suite T into two. The first suite T_{tr} is called the training suite and is used for constructing classifiers. The second suite T_{te} is called the test suite and is used for testing classifiers and tuning various parameters. First, we decide a ratio γ which we call the testing ratio. We use random sampling to construct T_{tr} and T_{te}. For each trace σ in T, we generate a random number u which is uniformly distributed over the range [0, 1]. If u ≤ γ, the trace is added to T_{te}, otherwise it is added to suite T_{tr}. Roughly speaking, γ denotes the fraction of the traces that are in the test suite T_{te}.
2) *Step 2 (Formation of a classifier):* We use the training suite T_{tr}. to construct a Markov classifier we use the Bernoulli scheme on the traces. Since the classifier is constructed from a suite of normal traces, it is able to discriminate between normal and anomalous traces. Details of the construction can be found in Section 4.
3) *Step 3 (Tuning Parameters):* There are various exogenous parameters used during the construction of the classifier. First, we define various performance metrics for a classifier. These metrics are computed using suites T_{te} and T_{an}. Various exogenous parameters are tuned using these performance metrics.

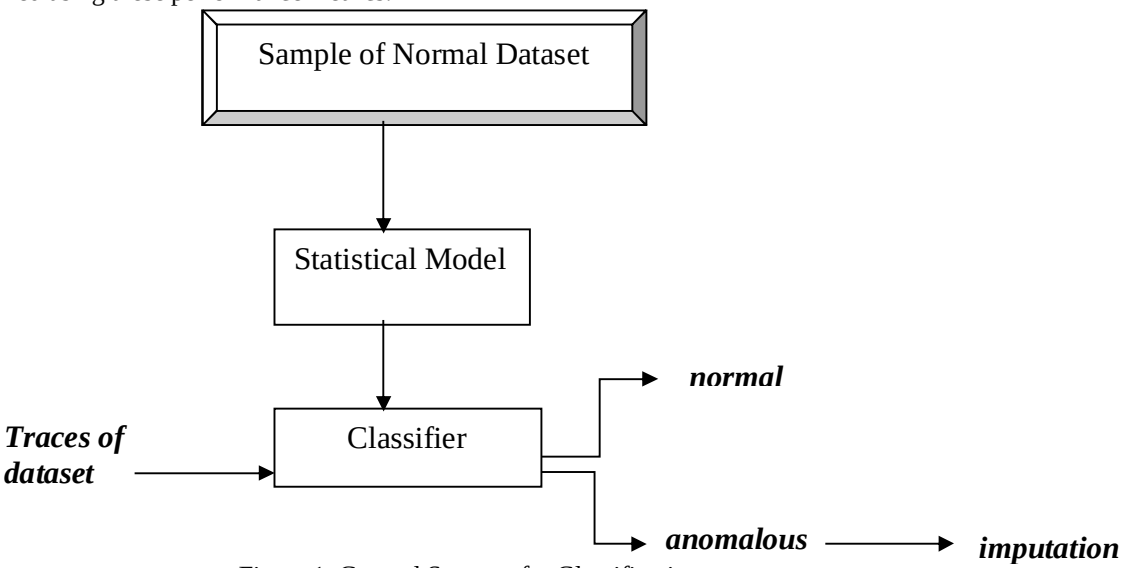


Figure 1. General Strategy for Classification

IV. DETAILED DESCRIPTION

Let Σ represent the set of alphabets.. A trace over Σ is a finite order of alphabets. This set of finite traces over Σ is denoted by Σ^*, the empty trace is denoted by ϵ, and the set of traces of length k is denoted by Σk. Given the trace $\sigma \in \Sigma$, $|\sigma|$ denotes the length of the trace. Given a trace σ and a positive integer $i \leq |\sigma|$, αi and $\alpha[i]$ having the prefix consisting of the first i alphabets and the i^{th} symbol respectively. The concatenation of two traces σ1 and σ2 is denoted by $\sigma 1 \cdot \sigma 2$. A = {0, 1} denotes the binary alphabet set.

A classifier $f : \Sigma^* \rightarrow B$ is called on-line if and only if there exists "efficiently computable" functions U, T, and $\beta : \Sigma^* \rightarrow < k$ such that following equations hold:

$\beta(\sigma) = U(\beta(\sigma n-1), \sigma[n-1], \sigma[n])$

$f(\sigma) = T(\beta(\sigma))$

In the equations given above, the length of the trace σ is denoted by n. Types for functions *U* and *T* can be easily inferred from the equations. Notice that the value of the function β on the trace σ only depends on the value of the function for σ, and the last two symbols of the trace. In some sense, β only depends on the "direct previous stage" and can be efficiently computed. It clearly states that the function β is called "Markov" because its value only depends on the immediate history and the current state).

The Bernoulli scheme, as any stochastic process, may be viewed as a dynamical system by endowing it with the shift operator T where,

$(T_x)_k = x_{k+1}$

Since the output data are independent, the shift preserves the measure, and thus *T* is a measure-preserving transformation. The quadruplet (X, A, σ, T) is a measure-preserving dynamical system, and is called a **Bernoulli scheme** or a **Bernoulli shift**. It is often denoted by

$BS(p)=BS(p_1....,p_N)$

V. EXPERIMENTAL RESULTS

A. Dataset

Experimental datasets were carried out from the IBM Log file dataset. The number of instances is 17368940 and the number of attributes is 1458, captured and recorded on 16.05.2018 at the IBM power9 series X3100 M4 2582E4-1220. This dataset is a log file in nature with a log stream of data with the missing value 5% to 12%. The main objective of the experiments conducted in this work is to analyse the classification of machine learning algorithm. Datasets without missing values are taken and few values are removed from it randomly. The rates of the missing values considerable reduced from 5% to 3.5%. In these experiments, missing values are artificially imputed in different rates in different attributes.

B. Experimental Evaluation

The Experimental evaluation features the overall structure of all the attributes and classes without any missing data. The chart listed below depict the live log data from IBM power9 series X3100 M4 2582E4-1220 server featuring various attributes like No. of Nodes, distance between sites, Connected Components, Network Diameter, Network Radius, Shortest Path, Avg.num.Neighbours, Characteristic Path Length, , Deviated Path Len, average connectivity, Multi Node Pair, , Outdegree(job depture), Stress, SelfLoops, MultiEnd NodePairs Partner, NeighbourhoodConnectivity, scalability, paging coeff with zero, Indegree(job arriving), Eccentricity, Closenesscentrality, Eucledian Distance, Segmentation with zero, X(latency time), Y(ee correction factor), Z(ee deviation), X1((processor wait state)), X2 (IDLE TIME), A1(NO OF REPLICATIONS), A2(LFA), A3(RFA), connectivity based EXECUTION TIME, HIT RATIO, EFFECTIVE NETWORK USAGE, OUTDEGREE NEW, MJET, bandwidth, average storage space, actual storage space, communication cost, Actual segmentation, and actual paging.

Impedance measurements were made at the frequencies 15.625, 31.25, 62.5 etc. KHZ. The following charts represents the classification of all attribute of original dataset using supervised machine learning techniques like Markov Classifier and unsupervised machine learning techniques like Mean, Median and STD without missing values and also describes the single instance IBM log file dataset without missing values.

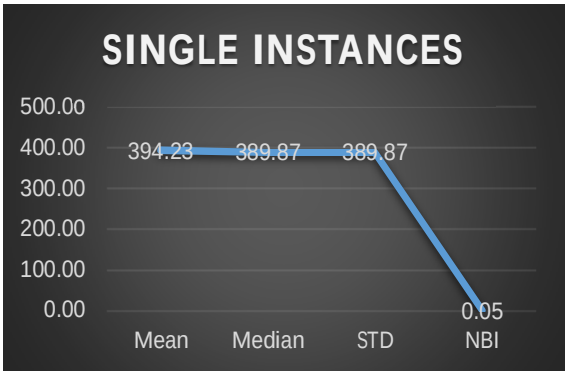


Figure 2. Original Datasets without Missing Values

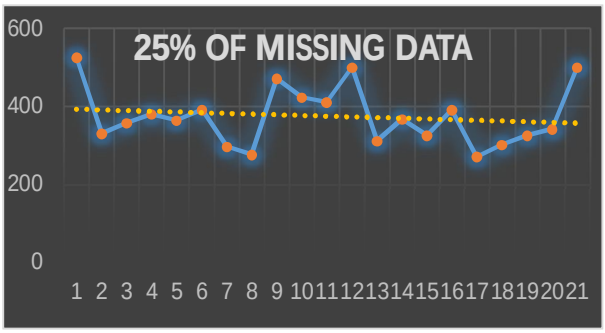


Figure 3. Original Datasets with Missing Values

The below figure 4 represents the percentage rates of missing values using both the techniques like supervised and unsupervised using missing values with the rate of 25%, 50%, 75%, 100% respectively. Figure 5 specifies the different percentage rates of missing values for experimental analysis of unsupervised techniques like Mean, Median and STD with the missing rate of different percentage.

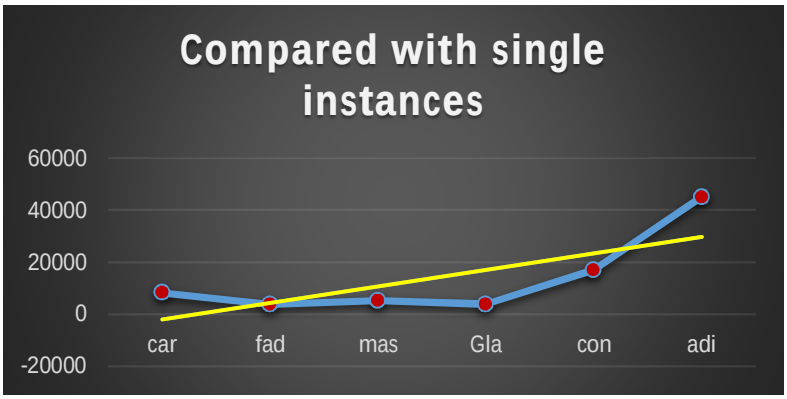


Figure 4. Missing Value Rates for Experimental Analysis

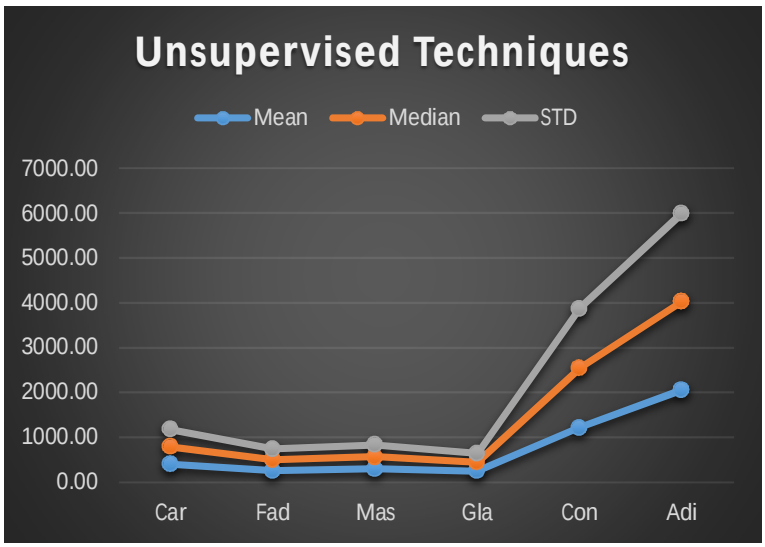


Figure 5. Experimental Results for Mean, Median and STD

Figure 6 represent the experimental results of both supervised machine learning techniques like Markov Classifier Imputation using missing value with the rate of 25%, 50%, 75%, 100% respectively. Figure 7 represents the comparison of both supervised Markov Classifier Imputation and unsupervised techniques Mean, Median and Standard Deviation using missing values for all the attributes contains different rate of percentage.

Figure 6. Experimental Results for Supervised Techniques

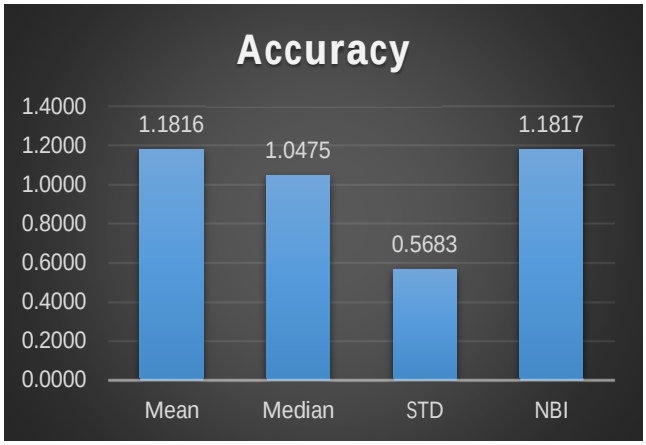


Figure 7. Comparative Results using both ML Techniques

VI. DISCUSSION

According to the previous discussion, Markov Classifier consists of 2 process. Process 1. State the imputation of element and the imputation sequence. Process 2. Apply Markov Classifier Imputation to assign missing values. For the sequence suitable approach, as stated above the imputation of element and the imputation sequence, Markov Classifier assign the missing value in the first imputation element of the sequence and then assign the later on the altered new database. Markov classifier construct classification model, however it can't be improved systemically also it can't automatically select suitable features like ADABOOST tree as the performance of Markov classifier lies on the rightness of the element selection in database. The most important drawbacks of Markov classifier is that it has strong feature independence assumptions. Following figure 8 & 9 show the performance of MCI is improved by the improvement of MCI

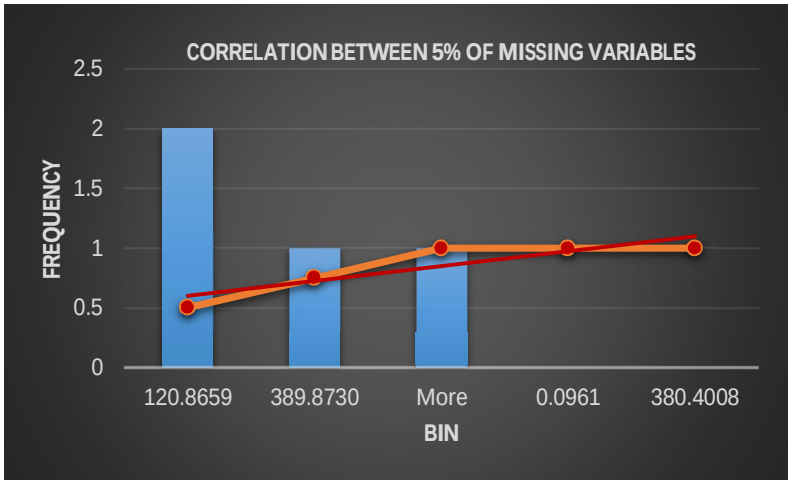


Figure 8. Correlation between less missing values

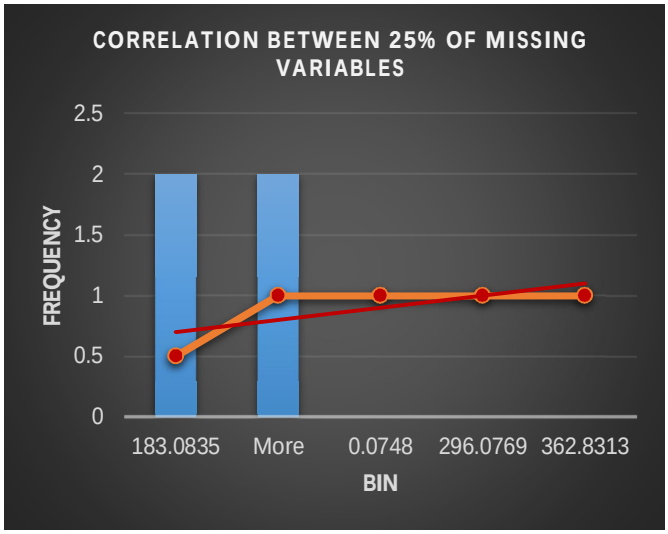


Figure 9. Correlation between more missing values

VII. CONCLUSION

In this paper, presents the implementation and evaluation of independence classifier gives the complete view about the multiple imputation of missing values in large dataset for abnormal data detection and imputation on scanty data based on Markov chains. We presented a general framework for constructing classifiers from Markov chains and presented three specific classifiers based on this framework. Experimental results clearly demonstrated the effectiveness of our approach.

Single imputation technique generates bias result and affects the caliber of the execution. The carrying out of the missing value imputation algorithms was measured with regard to different portions of missing information in large dataset.

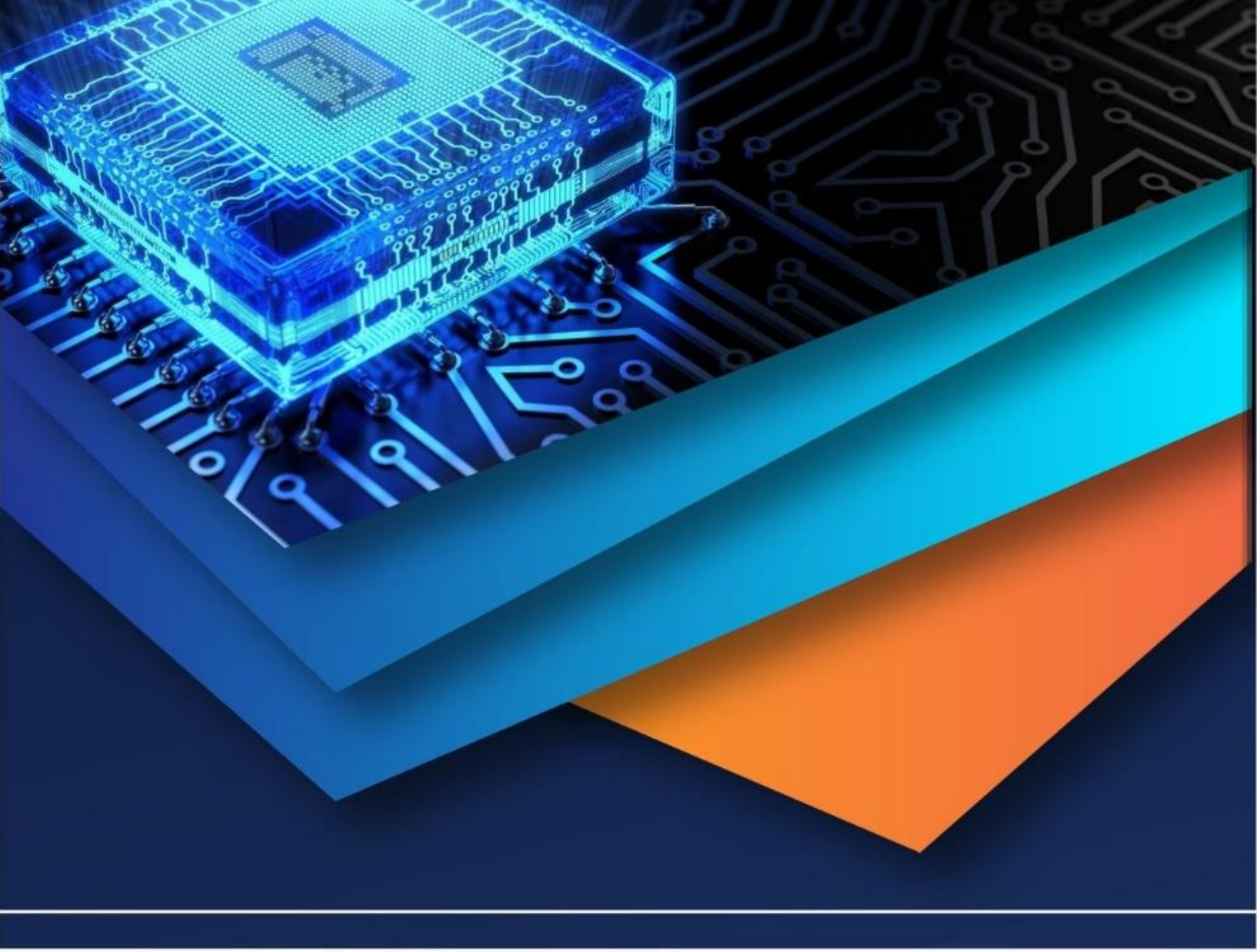
Also this paper shows the experimental result of standard deviation and Markov Classifier using limited parameter for their analysis and the performance evaluation stated, among the other missing value imputation techniques, the proposed method produce accurate result. In future, it can be expanded to handle categorical attributes and it can be substituted by other supervised machine learning techniques.

This report represents an efficient and effective missing data handling method, Markov Classifier model. The evaluation results indicate that MC is superior to multiple imputations. The performance of MC is improved based on the attribute selection. Granting to the common imputation techniques, Markov classifier is an effective missing data treatment model.

REFERENCES

[1] Liu P., Lei L., and Wu N., A Quantitative Study of the Effect of Missing Data in Classifiers, proceedings of CIT2005 by IEEE Computer Society Press, September 21-23,2005.
[2] 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.
[3] R.J. Little and D. B. Rubin. Statistical Analysis with missing Data, John Wiley and Sons, New York, 1997.
[4] R. Kavitha Kumar and Dr. R. M. Chandrasekar, “Missing Data Imputation in Cardiac data set”.
[5] 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-2012,
[6] 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 2011, pp14-19.
[7] Shichao Zhang, Xindong Wu, Manlong Zhu, “Efficient Missing Data Imputation for Supervised Learning” Proc, 9th IEEE conference on Cognitive informatics, 2010 IEEE.
[8] S.Hichao Zhang, Jilian Zhang, Xiaofeng Zhu, Yongsong Qin, Chengqi Zhang, “Missing Value Imputation Based on Data Clustering”, Springer-Verlag Berlin, Heidelberg,2008.
[9] S. Kanchana, Dr. Antony Selvadoss Thanamani, “Classification of Efficient Imputation Method for Analyzing Missing values”, International Journal of Computer Trends and Technology, Volume-12 Part-I, P-ISSN: 2349-0829.
[10] S. Kanchana, Dr. Antony Selvadoss Thanamani, “Multiple Imputation of Missing Data Using Efficient Machine Learning Approach”, Internation Journal of Applied Engineering Research, ISSN 0973-4562 Volume 10, Number 1 (2015) pp.1473-1482.
[11] Alireza Farhangfar, Lukasz Kurgan and Witold Pedrycz, “Experimental Analysis of Methods for Imputation of Missing Values in Databases.

[12] Blessie, C.E., Karthikeyan, E, Selvaraj.B. (2010): NAD – A Discretization approach for improving interdependency, Journal of Advanced Research in Computer Science, 2910,pp.9-17.
[13] 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.
[14] Han J. and Kamber M., Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers, 2001
[15] Ingunn Myrtveit, Erik Stensrud, "IEEE Transactions on Software Engineering", Vol. 27, No 11, November 2001.
[16] 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.
[17] Kamakshi Lakshminarayan, Steven A. Harp, Robert Goldman and Tariq Samad, "Imputation of Missing Data Using Machine Learning Techinques", from KDD-96 Proceedings.
[18] K. Lakshminarayan, S. A. Harp, and T. Samad, "Imputation of Missing Data in Industrial Databases", Applied Intelligence, vol 11, pp., 259-275, 1999.
[19] K.Raja, G.Tholkappia Arasu, Chitra S.Nair, "Imputation Framework for missing value" International Journal of Computer Trends and Technology-Volume3 Issue2-2012.
[20] Lim Eng Aik and Zarita Zainuddin, "A Comparative Study of Missing Value Estimation Methods: Which Method Performs Better?", 2008 International Conference on Electronic Design.
[21] R. Elliott, L. Aggoun, and J. Moore. Hidden Markov Models: Estimation and Control. Springer Verlag, 1995.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY