ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2018.0230

MISSING VALUE IMPUTATION AND NORMALIZATION TECHNIQUES IN MYOCARDIAL INFARCTION

K. Manimekalai[1] and A. Kavitha[2]

[1]Department of Computer Applications, Sri GVG Visalakshi College for Women, India
[2]Department of Computer Science, Kongunadu Arts and Science College, India

Abstract

Missing Data imputation is an important research topic in data mining. In general, real data contains missing values. The presence of the missing value in the data set has a major problem for precise prediction. The objective of this paper is to highlight possible improvement of existing algorithm for medical data. KNBP imputation method based on KNN and BPCA is proposed and evaluate MSE and RMSE estimates. Normalization is done by comparing three algorithms namely min-max normalization, Z-score and decimal scaling. The experiment is done with standard bench mark data and real time collected data. KNBP imputation method and Decimal Scaling Algorithm for Normalization got lower error rate.

Keywords:

Mean, Hot Deck, KNN, BPCA, KNBP, Min-Max Algorithm, Z-Score, Decimal Scaling

1. INTRODUCTION

Myocardial Infarction (MI) is a disorder in which cardiac myocytes undergo necrosis as a consequence of interrupted coronary blood flow. Myocardial infarction is a major cause of morbidity and mortality worldwide, with more than 7 million people in the world suffering from acute myocardial infarction each year. It is a part of the heart causing damage to the heart muscle. It is the irreversible death of heart muscle secondary to prolonged lack of oxygen supply [1]. MI is a minor stage in lifelong cardiac disease, although rarely noticed, but can lead to sudden death. Since it is the first symptom of coronary artery disease, it is very important to detect MI in an early stage. Even though MI can be detected by a number of different signs like biochemical markers, imaging or pathological characteristics, the most important initial clinical test for MI diagnosis still remains as an electrocardiogram (ECG). Heart disease is generally diagnosed by the Cardiologists based on three types of tests like patient symptoms, ECG information and enzymatic test. Since enzymatic tests are expensive and time consuming, ECG and symptoms of the patient are used to decide on the disorder in cases of emergency. So, MI can be normally identified from patient's history and ECG. Nowadays, computer-aided detection systems are used to detect MI. At present, the best practice for reducing mortality rates caused by complex diseases is to detect the symptoms at the early stages.

Data preprocessing is a technique which involves the transformation of raw data into an understandable format. Real-world data is often incomplete, inconsistent, and lacking in certain behaviors or trends. The major weakness with clinical data set is the presence of redundant records. The redundant instance causes the learning algorithm to be biased and unbiased towards frequent and infrequent records. These redundant records are removed in order to improve the detection accuracy. Mostly it contains many errors. Data preprocessing is one of the best method of resolving such issues.

This paper is organized as follows: Section 4 explains the introduction of Data preprocessing. Section 3 focuses on data cleaning and noisy data. Section 4 describes methodology and section 5 explains the results and discussion. Finally, section 6 concludes the paper.

2. DATA PREPROCESSING

The healthcare industry collects the data in huge number, which is unfortunately not "knowledgeable" to discover the hidden information for effective decision making.

Data Preprocessing includes various steps such as selection, cleaning, transformation, Interpretation/evaluation etc. The ultimate outcome of data preprocessing is the complete data set with consistent attributes.

The data set contains the collected data from both women and men under the age group ranging between 25 and 75. The data is observed from the patients who were affected with MI and few normal patients with minor symptoms. In this research, two types of data are used for experimentation i.e. (i) Standard benchmark data, it is taken from UCI machine Learning Repository and (ii) Real Time data collected in and around Coimbatore and Tiruppur District. The data obtained is distributed among various age groups and both genders. The real time data is inconsistent. So, it is necessary to preprocess for getting accurate result.

2.1 STEPS DURING KDD PROCESS

Data Mining involves few steps from raw data collection to some form of new knowledge. The Fig.1 shows the data pre-processing categories.

The iterative process consists of the following steps like data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation. [2]

- *Data Selection*: Data is selected and convert into target data. In this step, data relevant to the analysis task are retrieved from the database.
- *Preprocessing*: It is an important step in the data mining process. The target data is converted into processed data. It describes any type of process performed on raw data to prepare it for another processing procedure.
- *Data Transformation*: The processed data is converted into transformed data. In this process data are consolidated into forms appropriate for mining.
- *Data Mining*: Transformed data is converted into pattern.
- *Data Interpretation or Evaluation*: It involves the evaluation and possibly interpretation of the patterns to make the

decision of what qualifies as knowledge. It describes Interpretation/Evaluation patterns into knowledge.

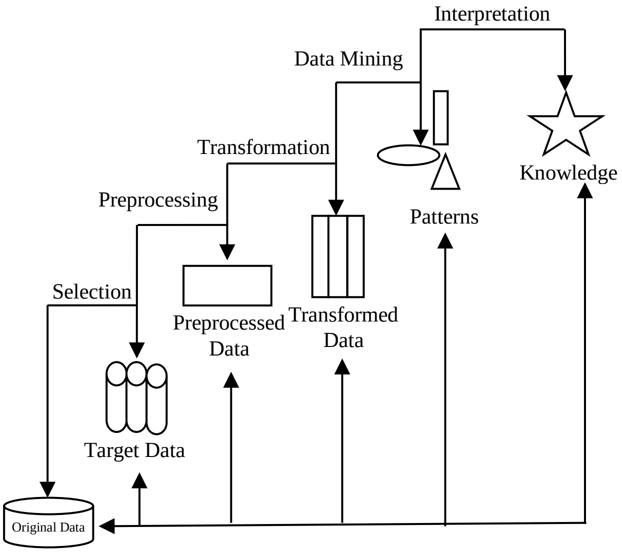

Fig.1. Data Mining- Knowledge Discovery Process

2.2 DATA PREPROCESSING METHODS

Data Preprocessing is an essential and crucial step in the data mining process and it has a massive impact on the success of data mining project. If there is inappropriate information present or noisy and unreliable data, then knowledge discovery becomes very difficult during the training process [10]. Data cleaning and transformation steps can take considerable amount of processing time but once pre-processing is done the data become more reliable and robust results are achieved. There are different of methods used for preprocessing. It includes Data Cleaning, Data Integration, Data Transformation and Data Reduction.

Data cleaning method are used to remove the noisy data, completed on uncompleted data and remove unnecessary data. Data integration method is integrated to different source of data in one place. Data transformation method change forms of data and data reduction reduce the volume of database by schema integration.

3. DATA CLEANING

Raw data may have incomplete records and noise values. Data cleaning is the first step in data preprocessing. It is the process of detecting and correcting or removing inaccurate records from a record set [5]. It is used to find the missing values. Cleaning and filtering of the data might be necessarily carried out with respect to the data and data mining algorithm employed so as to avoid the creation of deceptive or inappropriate rules or patterns [37].

3.1 MISSING VALUE

Missing value is one of the major factor from specific data set by applying data mining technique. There could be numerous reasons for missing values in a data set such as human error and hardware malfunction etc. It is to be noted that the missing values should be dealt carefully before analysis. Otherwise, the information extracted from data set containing missing values will lead to the path of wrong decision making.

In 2016, the author explained that, the simplest method is the “Deletion techniques” that are used to eliminate the attributes or cases [3]. This method is the default method for handling missing data. However, in many cases a large amount of missing data could drastically influence the result of the model.

During filling of the missing values three things should be kept in mind [11].

- Estimated values without bias.
- The relation between attributes should be maintained.
- Minimize the cost

There are several techniques available to control the issue of missing values such as replacing the missing value with: (i) closest value, (ii) mean value and (iii) median value etc.

The variables may be: (a) Missing Completely At Random (MCAR) [7] (b) Missing At Random (MAR) [8] and (c) Missing Not At Random (MNAR) [9]. Each variable should be managed freely. Imputation techniques help to credit the missing values. Pre-processing must be carried out before imputing the values by imputation techniques.

3.1.1 Different Methods for Missing Data Imputation:

Different techniques have been proposed for substituting missing values with statistical prediction. This process is known as ‘missing data imputation’.

There are several techniques based on imputation namely Mean, Mode and Median, K-nearest neighbors (KNN), fuzzy K-means (FKM), Singular Value Decomposition (SVD), Bayesian Principle Component Analysis (BPCA), Multiple Imputations by Chained Equations (MICE), Expectation Maximation (EM), Hot-Deck Imputation and C5.0.

The choice of proper imputation method is based on data types, missing data mechanisms, patterns and methods. Data type can be numerical, categorical or mixed [3]. The Table.1 shows that types of imputations already used or compared.

3.1.2 Different methods of Imputation Techniques:

Different methods of imputation techniques are shown in Table.1.

Table.1. Different methods of Imputation Techniques

Paper	Methods used or Compared	Observation
[12]	RF, KNN, SVD, Mean, Median, QRILC, HM, Zero	Random Forest performed the best.
[13]	PCA, Mean, Complete case analysis, SVD, BPCA	SVD provided the best result.
[14]	KNN, Evolutionary KNN	Evolutionary KNN achieved good result.
[15]	KNN based Pearson Coefficient,	KNN based Pearson Coefficient performed best.

	KNN based Euclidean metric	
[16]	MMSD, Hot Deck, Mean, Poly Regression, FHD	Mean method by Step Digression achieved good result.
[17]	Novel KNN, Sparse KNN	Sparse KNN provided good result.
[18]	Logistic Regression, Lasso, ANN, Improved Weighted KNN	Improved Weighted KNN and Lasso provides the good result.
[19]	BPCA, F-Kmeans (FRNNI, OWANNI, WQNNI), K means, KNN, EM, SVD	WKNN Imputation is good.
[3]	Mean/Mode, K-Nearest Neighbor, Hot-Deck, Expectation Maximization and C5.0	EM and KNN effective.
[20]	Pairwise deletion single-value imputation, Mean Mode Imputation (MMI), Hot and Cold Deck Imputation (HDI, CDI) and K-Nearest Neighbour (KNN)	KNN is more efficient.

From the review of literature, it is clearly observed that, from the traditional algorithms KNN is mostly used for missing value imputation. In this experiment the imputation techniques like mean, hot deck imputation, KNN, BPCA and KNBP are implemented and the results were compared.

3.1.3 Noisy Data:

Noise is a random error or variance in a measured variable. Noise data means that there is an error in the data.

3.1.4 Noise Removal:

The noise from the dataset is removed by following steps: The pair-wise distance using Euclidean distance is calculated between pair of objects; Calculate square distance of maximum value is taken from square distance values; Threshold value is calculated; if distance is greater than threshold, then the value is considered as noise and is removed from the dataset. For this research, benchmark medical dataset is used and they do not contain any irrelevant features as the respective publishers have already processed them. However, the used noise removal method can work for any type of features such as binary, numerical and categorical etc. depending upon the type of medical data.

4. METHODOLOGY

4.1 DATA PREPARATION AND PREPROCESS

In this paper, two types of data are used for implementation i.e. (i) Standard benchmark data which is taken from UCI Repository and (ii) Real Time data which is collected from the hospital.

The purpose of Normalization techniques is to map the data to a diverse scale. Normalization is a very common technique used in data preprocessing. This method works by adjusting the data values into a specific range such as between 0 to 1 and -1 to +1 [6]. In this paper three Normalization techniques namely Min-Max Normalization, Z-Score Normalization and Decimal Scaling Normalization are compared.

4.1.1 Min-Max Normalization:

The attribute data is scaled to fit into a specific range. One of the technique is called Min-Max Normalization. Min-Max Normalization transforms a value *A* to *B*. Normalize the dataset using min-max normalization as Eq.(1):

$$B=\left(\frac{\left(A-\min\left(A\right)\right)}{\max\left(A\right)-\min\left(A\right)}\right) \tag{1}$$

4.1.2 Z-Score:

Z-score is the most commonly used method. It converts all indicators to a common scale with an average of zero and standard deviation of one. Normalize the dataset using Z-Score normalization as Eq.(2) given below:

$$Z-score=\frac{value-mean}{SD} \tag{2}$$

4.1.3 Decimal Scaling:

Decimal scaling is a data normalization technique. In this technique we move the decimal point of values of the attribute. This movement of decimal points totally depends on the maximum value among all values in the attribute.

A value *v* of attribute *A* can be normalized by the following formula using Eq.(3):

$$Normalised\ attribute\ value=\frac{v_i}{10^j} \tag{3}$$

Table.2. Materials and Methods used for Transformation

Paper	Research Objective	Techniques used	Results
[21]	To find the noisy data and correct it.	Ada Boost	Detected the mislabeled data and then correct its label and attributes.
[22]	To present an Myocardial Infarction prediction model using classification data methods	Hybrid feature selection method and cost-sensitive model.	Achieved sensitivity, F-measure and accuracy.

[23]	To address the classification of Inferior Myocardial Infarction	Discrete Wavelet Transform (DWT) and Support Vector Machines (SVM).	Overall accuracy of SVM classifier is 97.02%
[24]	To analyze the use of K-Nearest Neighbor and Naive Bayes with different normalization techniques.	Min-max and Z-score techniques	Naïve Bayes with Z-score gives highest accuracy rate of 100%.
[25]	To select the right combination of preprocessing methods has a considerable impact on the classification potential of a dataset.	Ant Miner +	Significant improvement in classification performance measured by predictive accuracy.
[26]	To evaluate performance classification models in order to predict Myocardial Infarction.	Hybrid Feature selection method that includes Forward Selection and Genetic Algorithm.	Best results have been achieved.
[27]	To describe a novel approach for feature extraction in order to recognize trusty heart rhythm.	Baseline wander removal and Wavelet transformation	Wavelet transformation provides 96.79% accuracy.
[28]	To deal with a wavelet based method is used for detecting a Myocardial Infarction (MI) along with user identity.	Wavelet Transform and SVM.	90.42% accuracy is achieved.
[29]	To discuss the effects of change propagation resulting from using adaptive preprocessing in a multi-component Predictive System (MCPS).	Z-score normalization, PCA, and min-max normalization and one classifier GFMM neural network	Min-max normalization method was shown to reduce the classification error.
[30]	To find the risk level of MI	Min-Max algorithm	Particle Swarm Optimized Neural Network classifier can be conveniently applied to detect MI.
[31]	To maintain the large variation of prediction and forecasting	Integer scaling Normalization	This techniques works well for various data sets.
[32]	To present a method for extracting key features from each cardiac beat	Improved Bat algorithm	The performance of the classifier is improved with the help of the optimized features.
[33]	Data pre-processing is provided which utilizes a fuzzy technique in order to improve the data quality.	KCPA, MATLAB	The results demonstrate the effectiveness in classification accuracy.
[34]	To describe a preprocessing technique and analyzes the accuracy for prediction after preprocessing the noisy data.	Multilayer Perceptron and Radial Basis Function Network.	Prediction accuracy was nearly 91%. The MLP network prediction is has high accuracy with low error rates.
[35]	To predict the Heart attacks.	Normalization.	Prediction gives more accurate output.
[36]	To solve the limitation of abundant data to construct classification modeling in Data Mining.	Rough Sets.	The results become more Effective.

From the review of literature, it is clearly observed that Normalization algorithm takes an important part. It involves three techniques namely min-max algorithm, Z-score algorithm and decimal scaling algorithm. Though most of the authors use min-max algorithms, decimal scaling gives better result and minimum error rate which can improve the classification accuracy.

4.2 PROPOSED METHOD

Data should be cleaned before processing. The Fig.2 shows the steps of proposed method. Both Standard Benchmark data and Real time data are considered as MI data.

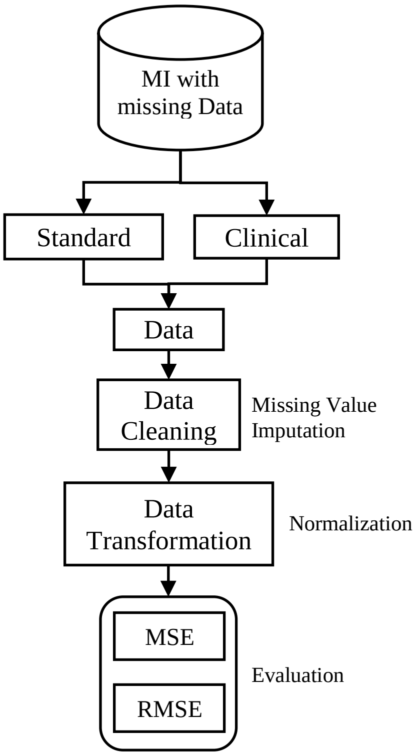

Fig.2. Proposed technique

The dataset is moved to clean. In data cleaning process missing value imputation is done by using mean method, hot deck imputation method, KNN method, BPCA method and hybrid KNBP method using Matlab. In data transformation process Normalization is done by comparing three techniques namely min-max normalization, Z-score normalization and decimal scaling normalization. The drawn output is shown in Table.4 and Table.5 respectively.

4.2.1 *Mean:*

This method is the easiest way to impute the missing values. In this all the missing values are replaced with its mean value. The mean of the attribute is computed using the non-missing values and is used to impute the missing values of that attribute. Mean is calculated by the formula using in Eq.(4).

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N} \tag{4}$$

where, $\bar{X}$ = mean, X_1 = first value, X_2 = second value, X_3 = third value, X_N = last value, N = number of values

4.2.2 *Hot Deck Imputation:*

This method is used for handling missing data in which each missing value is replaced with an observed response from a “similar” unit. This is very simple yet effective imputation method. Hot deck methods impute missing values within a data matrix by using available values from the same matrix. This is done by a correlation matrix that is used to determine the most highly correlated variables. The observation unit that contains the missing values are known as the recipient unit. The observation unit that provides the value for imputation is known as the donor unit.

4.2.3 *KNN:*

KNN imputation uses K-Nearest Neighbors approach to impute missing values. It is a neighbor based method. It replaces the missing values with the corresponding value from the nearest-neighbor. The nearest neighbor is the closest value based on Euclidean distance [38]. The missing values are imputed considering a given number of instances that are mostly similar to the instance of interest. By using Euclidean distance, the similarity of two instances are determined [39]. Euclidean distance *d* is calculated using Eq.(5),

$$d = \sqrt{\sum_{i=1}^{N} \left(X_i - Y_i \right)^2} . \tag{5}$$

4.2.4 *BPCA:*

This method uses the method of Bayesian estimation of PCA. It is a global based imputation method based on Eigen values. In this method, some continuous hyper parameters are introduced to determine the latent space [40]. BPCA required parameter optimization. A variational Bayes algorithm is used to iteratively estimate the posterior distribution of the model parameters. And the missing values until convergence is reached. The key feature of this approach is that principal axes with small signal to noise ratios are shrunk toward zero, so that the algorithm automatically screens for those axes that are the most relevant. MVs are initially imputed by row wise average.

4.2.5 *KNBP:*

This method uses the hybridization of KNN and BPCA. The procedure for this Algorithm is explained below:

Step 1: Load the dataset.

Step 2: Set the random values for *u*

Step 3: Get 100 random locations that can make into nan’s

Step 4: Find the *nan*

Step 5: Get the *x*, *y*, *z* of all other locations that are non *nan*.

Step 6: Get the *x*, *y*, *z* location

Step 7: Get distances of this location to all the other locations

Step 8: The closest non-nan value will be located at sorted Indexes

Step 9: Get the *u* value there

Step 10: Replace the bad *nan* value in u with the good value.

Step 11: The value *u* should be fixed now no *nan* in it.

Step 12: Double check: Sum of nans should be zero now.

4.2.6 *Evaluation Criteria:*

Mean Square Error and Root Mean Square Error are used to evaluate. By using the Eq.(6) and Eq.(7), MSE and RMSE are calculated.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \bar{y}_i \right)^2 \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \bar{y}_i \right)^2} \tag{7}$$

5. RESULTS AND DISCUSSION

In this experiment 2 types of datasets are used. The first data set is taken from the standard benchmark dataset. (i.e) UCI repository. The second data set is collected from real time. Data observed from both the patients i.e. women as well as men but their age ranges from 25 to 90. Standard benchmark dataset contains 209 records and 8 attributes. The real time data contains 500 records and 21 attributes. The detailed description of benchmark dataset is shown in the Table.3.

Table.3. Description of the Selected Attribute Set

S.No.	Name of the attribute	Description
1	Age	Age of the Patient
2	Chest Pain	Chest Pain Type
3	Rest BP	Resting Blood Pressure
4	Blood sugar	Fasting Blood Sugar
5	Rest_Elctro	Resting Electrocardiographic Result
6	Max-Heart Rate	Maximum Heart Rate
7	Exercise Angina	Exercise Induced Angina
8	Disease	Diagnosis of Heart Disease

Preprocessing reconstructs the data into a format that would be very easy and effective for further processing. For missing value imputation, it can be observed that ‘Deletion’ is the worst performing method. The best one is ‘Imputation by Predictive Model’ followed by ‘Imputation by Average’. Imputations of missing values are difficult. The comparison of algorithms and missing value imputation is done by using Matlab.

The Table.4 shows the comparison of three normalization techniques. In this mean square error is high in real time data when compared to bench mark data. When compared to three algorithms, decimal scale point algorithm got the lower error rate. The comparison of three Normalization algorithms is shown in Fig.3.

The Table.5 shows the comparison of missing value imputation. The proposed method is compared with the following algorithms namely mean, hot deck imputation, KNN and BPCA. The imputing performance is evaluated by MSE and RMSE between the estimated missing points and the original points.

Table.4. Comparison of Normalization Techniques

Approach	Min-Max Normalization		Z-Score Normalization		Decimal Point Normalization	
	MSE	Time (sec)	MSE	Time (sec)	MSE	Time (sec)
Real Time Data	3.5371	0.8987	3.5395	1.0813	3.3028	0.3544
Benchmark Data	1.5098	1.0530	1.5088	0.7029	1.5071	0.8397

When compared to other algorithms, both real time data and benchmark data error rate is lower in the proposed method. It is shown in Table.5.

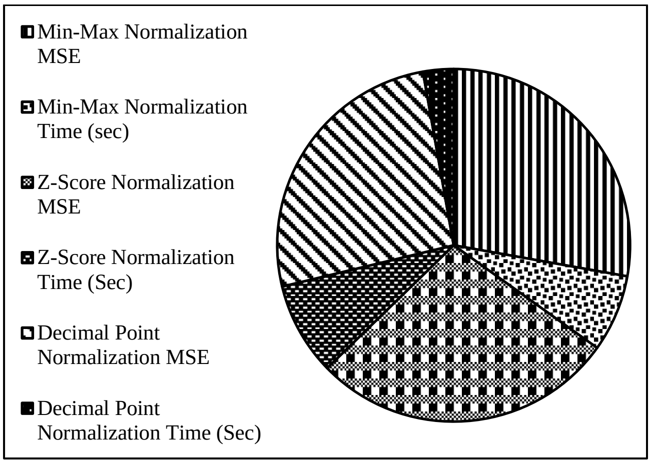

Fig.3. Comparison of Normalization algorithms

Table.5. Comparison of Missing Value Imputation Algorithms

Algorithms	Real Time Data		Bench Mark Data	
	MSE	RMSE	MSE	RMSE
Mean	7.3654	2.7139	3.7581	1.9386
Hot Deck	7.5520	2.7481	3.9564	1.9891
KNN	7.2591	2.6943	3.8922	1.9729
BPCA	7.2586	2.6942	3.8912	1.9726
KNBP	7.0414	2.5910	3.5325	1.8795

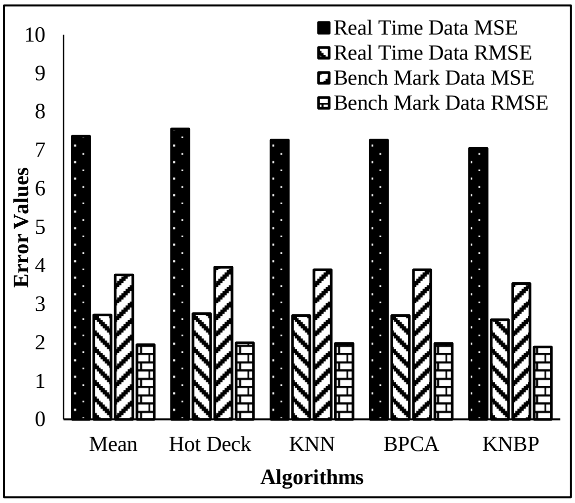

Fig.4. Comparison of Missing Value Imputation Algorithms

Comparison of Missing value imputation is shown in Fig.4. In this experiment 5 algorithms are used. The above graph shows that the proposed algorithm KNBP got the lower error rate for both Benchmark data and Real time data when compared to other algorithms.

6. CONCLUSION AND FUTURE WORK

Experiment shows that the proposed KNBP imputing method is able to estimate the missing values with lower error rate than other traditional methods. In normalization, the error rate and time is calculated for three algorithms. While comparing the algorithms decimal scaling algorithm produced better result. From the graphs it is clearly understood that real time data achieved higher error when compared to standard data. Due to standard data are already cleaned. In this experiment 500 real data with 21 attributes were taken. It can cause a lot of work for applying data mining algorithm. In future, dimensionality reduction is applied to reduce the number of attributes with the use of 1000 real time datasets.

REFERENCES

[1] A. Sudha, P. Gayathri and N. Jaishankar, "Utilization of Data Mining Approaches for Prediction of Life Threatening Disease Survivability", *International Journal of Computer Applications*, Vol. 14, No. 17, pp. 51-56, 2012.

[2] M. Durairaj and S. Sivagowry, "A Pragmatic Approach of Preprocessing the Data Set for Heart Disease Prediction", *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 2, No. 11, pp. 23-29, 2014.

[3] Tahani Aljuaid and Sreela Sasi, "Proper Imputation Techniques for Missing Values in Data sets", *Proceedings of International Conference on Data Science and Engineering*, pp. 168-176, 2016.

[4] Peter Schmitt, Jonas Mandel and Mickael Guedj, "A Comparison of Six Methods for Missing Data Imputation", *Journal of Biometrics and Biostatistics*, Vol. 6, No. 1, pp. 1-6, 2015.

[5] Vinod Bharat, Balaji Shelale, K. Khandelwal and Sushant Navsare, "A Review Paper on Data Mining Techniques", *International Journal of Engineering Science and Computing*, Vol. 4, No. 5, pp. 1976-1979, 2016.

[6] Suad A. Alasadi and Wesam S. Bhaya, "Review of Data Preprocessing Techniques in Data Mining", *Journal of Engineering and Applied Sciences*, Vol. 12, No. 16, pp. 4102-4107, 2017.

[7] J.F. Mac Gregor and T. Kourti, "Statistical Process Control of Multivariate Processes", *Control Engineering Practice*, Vol. 3, No. 3, pp. 403-414, 1995.

[8] R. Dunia, S.J. Qin and T.F. Edgar, "Identification of Faulty Sensors using Principal Component Analysis", *AICHE Journal*, Vol. 42, No. 10, pp. 2797-2812, 1996.

[9] R. Little, "*Statistical Analysis with Missing Data*", 2nd Edition, Wiley Press, 2002.

[10] Nazri Mohd Nawi, Walid Hasen Atomi and M. Z. Rehman, "The Effect of Data Pre-Processing on Optimized Training of Artificial Neural Networks", *Procedia Technology*, Vol. 11, pp. 32-39, 2013.

[11] Bhavisha Suthar, Hemant Patel and Ankur Goswami, "A Survey: Classification of Imputation Methods in Data Mining", *International Journal of Emerging Technology and Advanced Engineering*, Vol. 2, No. 1, pp. 309-312, 2012.

[12] Runmin Wei, Jingye Wang, Mingming Su, Erik Jia, Tianlu Chen and Yan Ni, "Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data", *Scientific Reports*, Vol. 8, pp. 663-674, 2017.

[13] Kristen A. Seversaon, Mark C. Molaro and Richard D. Braatz, "Principal component Analysis of process Datasets with Missing Values", *Processes*, Vol. 5, No. 3, pp. 38-49, 2017.

[14] Hiroshi De Silva and A. Shehan Perera, "Missing Data Imputation using Evolutionary K-Nearest Neighbor Algorithm for Gene Expression Data", *Proceedings of 6th International Conference on Advances in ICT for Emerging Regions*, 2017.

[15] Dan Zeng, Dan Xie, Ran Liu and Xiaodong Li, "Missing Value Imputation Methods for TCM Medical Data and its Effect in the Classifier Accuracy", *Proceedings of 16th International Conference on Advances in ICT for Emerging Regions*, pp. 339-354, 2017.

[16] S. Thirukumaran and A. Sumathi, "Improving Accuracy Rate of Imputation of Missing Data using Classifier Methods", *Proceedings of 10th International Conference on Intelligent Systems and Control*, pp. 1243-1251, 2017.

[17] Shichao Zhang, Debo Cheng, Zhenyun Deng, Ming Zong and Xuelian Deng, "A Novel KNN Algorithm with Data Driven K Parameter Computation", *Pattern Recognition Letters*, 2017.

[18] Xianglin Yang, Yunhai Tong, Xiang Shuai Zhao, Zhi xu, Yanjunli, Xin Jia and Shaohna Tan, "Adaptive Logistic Group Lasso Method for Predicting the No-Reflow among the Multiple Types of High Dimensional Variables with Missing Data", *Proceedings of 7th International Conference on Software Engineering and Service Science*, pp. 461-469, 2017.

[19] Mehran Amiri and Richard Jensen, "Missing Data Imputation using Fuzzy-Rough Methods", *Neurocomputing*, Vol. 205, pp. 152-164, 2016.

[20] Asma Saleem, Khadim Hussain Asif, Ahmad Ali and Shahid Mahmood Awan, "Pre-Processing Methods of Data Mining" *Proceedings of 7th International Conference on Utility and Cloud Computing*, pp. 651-663, 2014.

[21] Xiangyang Liu, "A Preprocessing Method of AdaBoost for Mislabeled Data Classification", *Proceedings of 29th International Conference on Control and Decision*, pp. 23-32, 2017.

[22] A. Daraei, H. Hamaidi, "An Efficient Predictive Model for Myocardial Infarction using Cost-sensitive J48 Model", *Iran Journal of Public Health*, Vol. 46, No. 5, pp. 682-692, 2017.

[23] Thripurna Thatipelli and Padmavathi Kora, "Classification of Myocardial Infarction using Discrete Wavelet Transform and Support Vector Machine", *International Research Journal of Engineering and Technology*, Vol. 4, No. 7, pp. 429-432, 2017.

[24] V. Hemalatha and C. Usha Nandhini, "An Efficient Approach for Constructing a Model for Diagnosing Heart Disease Dataset", *International Journal of Contemporary Research in Computer Science and Technology*, Vol. 3, No. 3, pp. 41-44, 2017.

[25] Sarab AlMuhaideb, "An Individualized Preprocessing for Medical Data Classification", *Procedia Computer Science*, Vol. 82, pp. 35-42, 2016.

[26] Hojat Hamidi and Atefeh Daraci, "A New Hybrid Method for Improving the Performance of Myocardial Infarction Prediction", *Journal of Community Health Research*, Vol. 5, No. 2, pp. 110-120, 2016.

[27] Muhammad Sheikh Sadi, et al., "A New Approach to Extract Features from ECG Signals", *Proceedings of 2nd*

International Conference on Electrical Information and Communication Technology, pp. 189-194, 2015.

[28] S. Selva Nithyananthan, S. Saranya and R. Santha Selva Kumari, “Myocardial Infarction Detection and Heart Patient Identity Verification”, *Proceedings of International Conference on Wireless Communications, Signal Processing and Networking*, pp. 1107-1111, 2016.

[29] Manuel Martin Salvador, Marcin Budka and Bogdan Gabrys, “Effects of Change Propagation Resulting from Adaptive Preprocessing in Multicomponent Predictive Systems”, *Procedia Computer Science*, Vol. 96, pp. 713-722, 2016.

[30] V. Seenivasagam and R. Chitra, “Myocardial Infarction Detection using Intelligent Algorithms”, *Neural Network World*, Vol. 1, pp. 91-110, 2016.

[31] S. Gopal Krishna Patro, Kishore Kumar sahu, “Normalization: A Preprocessing Stage”, Available at: https://arxiv.org/ftp/arxiv/papers/1503/1503.06462.pdf.

[32] Padmavathi Kora and Sri Ramakrishna Kalva, “Improved Bat Algorithm for the Detection of Myocardial Infarction”, *Springerplus*, Vol. 3, No. 4, pp. 666-678, 2015.

[33] Thripurna Thatipelli and Padmavathi Kora, “Classification of Myocardial Infarction using Discrete Wavelet Transform and Support Vector Machine”, *International Research Journal of Engineering and Technology*, Vol. 4, No. 7, pp. 429-432, 2014.

[34] M. Durairaj and S. Sivagowry, “A Pragmatic Approach of Preprocessing the Data Set for Heart Disease Prediction”, *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 2, No. 11, pp. 6457-6465, 2014.

[35] S. Florence, N.G. Bhuvaneswari Amma, G. Annapoorani and K. Malathi, “Predicting the Risk of Heart Attacks using Neural Network and Decision Tree”, *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 2, No. 11, pp. 7025-7030, 2014.

[36] Li Xiang-Wei and Qi Yian-Fang, “A Data Preprocessing Algorithm for Classification Model based on Rough Sets”, *Physics Procedia*, Vol. 25, pp. 2025-2029, 2012.

[37] V.V. Jaya Rama Krishniah, D.V. Chandra Sekar and K. Ramchand H Rao, “Predicting the Heart Attack Symptoms using Biomedical Data Mining Techniques”, *The International Journal of Computer Science and Applications*, Vol. 1, No. 3, pp. 10-18, 2012 .

[38] V. Kumutha and S. Palaniammal, “An Enchanced Approach on Handling Missing Values using Bagging K-NN Imputation”, *Proceedings of International Conference on Computer Communication and Informatics*, pp. 123-128, 2013.

[39] Tahani Aljuaid and Sreela Sasi, “Proper Imputation Techniques for Missing Values in Data sets”, *Proceedings of International Conference on Data Science and Engineering*, pp. 108-116, 2016.

[40] S. Oba, I. Takemasa, M. Monden and K. Matsubara, “A Bayesian Missing Value Estimation Method for Gene Expression Profile Data”, *Bioinformatics*, Vol. 19, No. 16, pp. 2088-2096, 2003.