Improved Class-Based Clustering Classifier for Imputation Intelligent Medical Data

Ms. P. Premalatha[1], S. Subasree[2*], N. K. Sakthivel[3]

[1]Research Scholar, Bharathiar University, Coimbatore, Tamil Nadu, India
[2]Professor and Head, Department of Computer Science & Engineering
Nehru College of Engineering and Research Center, Pampady, Thrissur, Kerala, India
[3]Department of Computer Science and Engineering
Nehru College of Engineering and Research Centre, Thrissur, Kerala, Indi

Abstract

The fast evolution in medical application yields to abundance of huge amount of data in volume and velocity. Due to this heterogeneous medical data generation from clinical trials, its typically not free from missing values. Previously introduced imputation techniques don't discourse the high spatiality problems and application of distance function that even have curse on high spatiality problem. Thus, there's a necessity an Efficient and Accurate technique to overcome this problem in Medical Data Analysis. To address the above mentioned issues, this research work proposed an efficient Class-Based Clustering Classifier for Imputation Intelligent Medical Data (C^3IMD). This work was implemented in Bio Weka and studied thoroughly. To improve the classification and prediction accuracy, missing data in Medical Data Sets were filled efficiently with the help of proposed Cluster-Classifier Model. The experiments are repeated with various datasets and results are evaluated and compared with existing classifiers WPT-DELM and SVM-DELM. From the results obtained, it was revealed that the proposed C^3IMD) is outperforming both the existing models in terms of Classification Accuracy, Sensitivity, Specificity and FScore. The Error Rate was analyzed and measured and observed that Error Rate observed in C^3IMD Classifier. Thus to improve the FScore value, some modifications are made in our Classifier C^3IMD to reduce Error Rate. The proposed Classifier is called as Improved Class-Based Clustering Classifier (ICBCC). From the results, it was noticed that the proposed ICBCC is outperforming C^3IMD in term of FScore.

Keywords*: Classification; Clustering; Hybrid classifier; Imputation; Medical Data; SVM; DELM.*

1. Introduction

The year of big data is ongoing, due to bulk-volume, complex and increasing number of data sets which are produced by numerous sources such as Internet of Things (IoT), government records, health records, multimedia, phone logs, social media and some other digital sectors [1-3]. Furthermore, big data are being used to convert medical practice, notify business decision making, and modernize public procedure[4,5]. Therefore, the production of complex data from medical and healthcare increase rapidly with huge essential information. So, big data has infinite potential in efficiently storing, processing, querying, and analyzing medical data [1,2,6-12].

Due to rapid increasing medical data, imputation is presently a lively area of research[15-18]. Imputation of medical data needs information of statistical features and data mining applications. In many times, medical data mining requires missing value handling for feature extraction by execution of imputation, so that the imputed values will leads better classification results. Now a days many imputation techniques are fail to give good classification rates. One of the simplest techniques to solve the incomplete values is that, just to eliminate the records which are having missing values. This common method is only possible on medical data; it has minimum missing values and no knowledge about the pattern of missing values. Furthermore, this removal method has a change to remove important data and information loss too. Previous existing approaches for missing value prediction are discussed in [1,2,6,13,14], some methods are longitudinal data for imputation, and imputation based on regression techniques, rough estimation of incomplete data using the concept of mean and median.

Imputation process is defines as replacement of statistical missing data. If an imputation on single data or single attribute is called unit imputation. Consecutively, item imputation is defined as the missing data or incomplete data is handling at the module level. Missing attribute is also affecting the accuracy of classification on medical data, also which is common in clinical data [1,2,7,28-32].

Imputation method needs numerous pre-processing techniques, which is common in all data mining process. Additional concern is important attribute of medical data records should be considered for data analysis. For this purpose, feature selection techniques are applied. It is to be taken care for data those are not affect final data classification will be discarded [1,2,8,21-23]. Value of data elements is important for statistical approaches and techniques in data analysis. Missing values [1,2,9] control application of the statistical approach and data analysis is thus not possible. Here application of imputation will help to do analysis and classification on statistical record.

This research work proposed an efficient Class-Based Clustering Classifier (C^3IMD) to improve classification Accuracy. However, it was noticed that our previous classifier C^3IMD has limitations to reduce Error Rate. To address this issue, we improved the C^3IMD and proposed Improved Class-Based Clustering Classifier (ICBCC).

The rest of this research article is arranged as follows. Section 2 describes the survey on various classifiers and impu-

tation techniques and our previous Classifier C^{3}IMD. The proposed Improved Class-Based Clustering Classifier (IC-BCC) Classifier is discussed in Section 3. Results were reported in Section 4 and Section 5 delivers the conclusion of this research paper.

2. Literature Survey

A decision tree based approach is discussed in [1,2,10], which have unique choice for handling missing data on medical data. Clustering is one of the mostly known techniques to handle incomplete data. One such techniques are discussed in [11] for medical data handling. In [12], Support vector regression and clustering methods are discussed for imputation process. Studies such as [13,14] discourse management of mixed features with incomplete values. [15] Describes a new framework for execution of imputation.

Incomplete record has been managed by auto regression method [2,16]. C5.0 enhanced by addition of two imputation methods called as IITMV (Intelligent imputation technique).

This methodology is proposed a tree based on C5.0 function and hot-deck application and imputation using EM approach. In [17], an enhanced classifier is proposed to improve its accuracy and its performance is compared with leading existing imputation technique of mean, median and hot deck techniques. Best matching record is used to impute incomplete data using density measure [18] and its performance is compared with some existing imputation methods like fuzzy c means, k-means and genetic algorithm based approach.

Various imputation methods are analyzed and best technique is outlined in [19], which is conducted on various synthetic datasets. All medical records are having missing value, which directly influence the classification accuracy. Both nominal and continuous missing value is imputed by class mean imputation based on the k -Nearest Neighbour Hot deck imputation approach in dataset [20]. Ratio type imputation method [21] is presented on population data for missing data estimation. Phishing attack's severity term have been detected by K-means [19,20,22] and multilayer perceptron based imputation technique in financial sector. Random forest method is presents in [23], which is machine learning based imputation method. This approach has enhanced the performance of random forest method with increasing the correction of attributes. A new novel based imputation method is demonstrated [24] for suitable model selection from a multitude of imputation method for specific attribute based on learning process on the known variable.

Moreover, there are additional issues of big data in medical data that are organized by the obtaining of details from difficult heterogonous patient sources. These tasks include procurement clinical data and understanding them in the right context, establishing medical data, observing data about biomarkers and considerate huge amounts data which can be valuable in medical settings when the patient is evaluated. Though, the modern technologies like Artificial Intelligence (AI) can support in resolving diverse complex problems.

AI in its broadest intelligence would prove the capability of a machine to do tasks alike to the human behavior. Therefore, complex task implementation in computer systems has been executed with AI, which are more problematical than normal one [24,25].With the purpose of modifying the act of AI, the computational intelligence (CI) approaches familiarize to medical data. CI usually denotes to the capacity of computer to learn a specific task from experimental observation or data, which simplify the smart behavior in difficult problems and varying settings [26].

CI methods are classified based on single and hybrid methods, where single methods denotes to those trainings which use only one of the machine learning techniques as a main method and the other classification denotes to those trainings that used hybridization of each two methods. For example, Artificial Immune Recognition System (AIRS) [1,2,27] has been used as the primary method for atherosclerosis analysis using a single classification.

In [2,28], proposed a hybrid classification in clinical datasets classification using Fuzzy Support Vector Machines (FSVM) method.

In the Class-Based Clustering Classifier (C^{3}IMD), at the subdivision, similarity measures on medical attribute is focussed to reduce error rate.

3. Class-Based Clustering Classifier (C^{3}IMD)

In this subdivision, Similarity measures on medical attribute and the imputation with hybrid classification method in CI is discussed.

3.1 Imputation Method

Literature survey consists of various existing analysis method for filling incomplete attributes in different dataset. However, the classification accuracies attained using these techniques have been not so capable when studied. Due to this reason which implicitly inspired us to analyze and discourse new imputation approach. We term our approach as Class-Based Clustering Classifier for Intelligent Imputation Medical Data (C^{3}IMD). The prime objective and goal of the C^{3}IMD Model is to decreases the feature dimensionality.

Figure 1 illustrates the Class-Based Clustering Classifier Model for medical data classification.

This Model has two unique stages and objectives. The first stage is used for imputing the records of medical data by employing class-based clustering which is used for feature reduction and predicting and filling missing values. The second stage for achieving improved classification and prediction accuracy through existing SVM classifier.

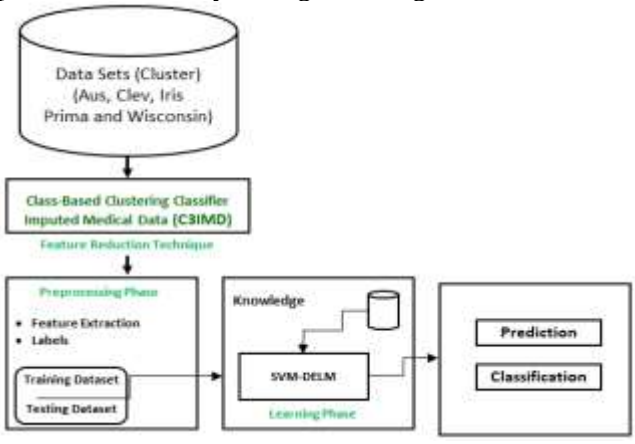


Fig. 1: Class-Based Clustering Classifier

3.1.1. Imputation measure

Let R_i and R_j be two clinical records and every record is having m attributes, which defines as $A_1, A_2 \ldots A_m$. Moreover, RA_{ik} and RA_{jk} defines the $i^{th}$and j^{th} attribute value of R_i and R_j medical records respectively. The membership value between R_i and R_j is well-defined in equation (1),

$$M^l(R_i, R_j) = e^{-\left(\frac{RA_{il}-RA_{jl}}{\sigma_l}\right)^2} \quad (1)$$

The similarity between R_i and R_j is calculated by,

$$Sim(R_i, R_j) = \begin{cases} \prod_{l=1}^{l=m} e^{-\left(\frac{RA_{il}-RA_{jl}}{\sigma_l}\right)^2}; \\ RA_{iy} \neq \emptyset \text{ and } RA_{jy} = \emptyset \\ \prod_{l=1}^{l=y-1} e^{-\left(\frac{RA_{il}-RA_{jl}}{\sigma_l}\right)^2} * \prod_{l=y+1}^{l=m} e^{-\left(\frac{RA_{il}-RA_{jl}}{\sigma_l}\right)^2} \\ RA_{iy} \neq \emptyset \text{ and } RA_{jy} = \emptyset \\ \prod_{l=1}^{l=y-1} e^{-\left(\frac{RA_{il}-RA_{jl}}{\sigma_l}\right)^2} * \prod_{l=y+1}^{l=m} e^{-\left(\frac{RA_{il}-RA_{jl}}{\sigma_l}\right)^2}; \\ RA_{iy} = \emptyset \text{ and } RA_{jy} \neq \emptyset \end{cases} \quad (2)$$

Where σ_1 represents the standard deviation of l^{th} attribute column values. The membership function is explained in the architecture of Figure 2.

Also, $\emptyset$ represents the incomplete attribute values. For all such missing values of two medical record R_i and R_j are $m^1(R_i, R_j)$ treated as 1. Attribute similarity is estimated by fuzzy similarity function.

3.1.2 Computation using Similarity Measure

Similarity between two clinical records R_i and R_j are estimated by below equation and detailed explanation is given in Fig 1.

$$Sim(R_i, R_j) = \prod_{l=1}^{l=4} e^{-\left(\frac{RA_{il}-RA_{jl}}{\sigma_l}\right)^2} \quad (3)$$

3.1.3 Imputation algorithm

Incomplete data or Missing medical data imputation algorithm is demonstrated in below steps.

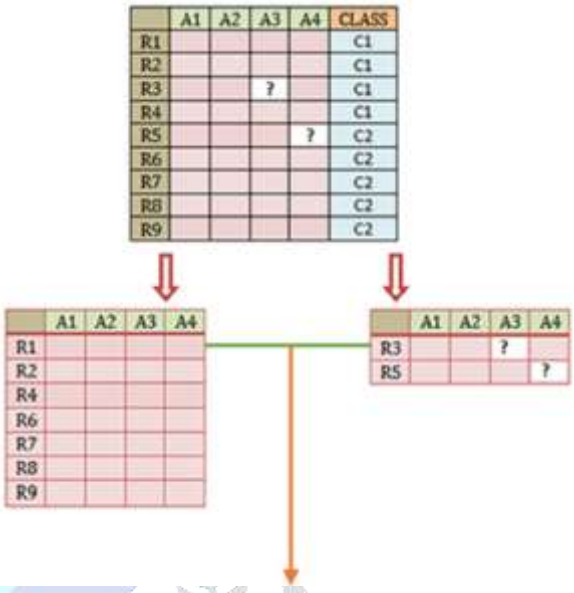
Step 1: In first Step, class based clustering imputation method is used to cluster the whole medical data into two groups, namely complete data G^1 and incomplete data G^{IM}, this needs to be imputed.

Step 2: Complete data G^1 are grouped (Cluster) by k-means algorithm. In which, mean ans standard deviation is calculated for each attributes and act as representatives for cluster forming.

Step 3: Next, fuzzy similarity is used to transform attributes into fuzzy vectors. For this, both G^1 and G^{IM} records are considered for similarity and dissimilarity calculation. This process includes the mean value of cluster by applying Fuzzy measures or distance vectors using Euclidean distance on medical data. When considering G^{IM}, mean vector values of incomplete data are removed in both similarity and dissimilarity calculation.

Furthermore, in Step 3, all attribute data are converted into new dimension, which represents the overall label count of the dataset. It is to be noted that the standard deviation is calculated for transformed data G^{IM} only.

Step 4: After data transformation, each attribute from G^{IM} is considered for similarity (dissimilarity) calculation. Each data from G^{IM} is chosen and similarity is calculated with complete data from G^1 group.



Equation (2)

Fig 2: Computation of Membership Function

Step 5: In this step, incomplete values is filled. For each missing attribute A^y in medical record R^{IM} presents in the group G^{IM} is filled by selecting R^X in complete data G^1, for which similarity is maximum.

Step 6: Step 1- Step 5 is repeated until the record has no missing values. After that the filling medical record is applied to hybrid classifier for disease prediction and classification.

Algorithm

Input: Incomplete medical dataset
Output: Imputed medical data

Variable notation

n = Total number of Medical record
i, j = Index of Medical records
k = Index of medical record attribute
P_k = k^{th}attribute variable
M_i = i^{th}Medical record
$M_i(P_k)$ = k^{th}attribute value in i^{th} medical record
$\emptyset$ = Missing attribute value
$|L|$ = Number of unique class label
$C[r]$ = r^{th}cluster
U = Medical dataset with class labels
μ_r = Meanof r^{th}cluster

Begin of algorithm

1. Split the input medical data into two group
   Complete data
   $$Group1 = \{M_i \mid \forall\, i,k \; and \; M_i\,(P_k) \neq \emptyset\}$$
   Incomplete Data
   $$Group2 = \{M_i \mid \forall\, i,k \; and \; M_i\,(P_k) = \emptyset\}$$
2. K-means algorithm is applied on medical data in order to cluster it. Number of cluster is equal to number of class label.
   For Group 1 data cluster is defined as,
   $$C[U,r] = KMeans(Grup1, |L|)$$
3. Calculate the mean of r-Cluster by using below expression
   $$\langle \mu_r^1 = \mu_r^1, \mu_r^2, \ldots \mu_r^k \rangle$$
   $$where, \mu_r^k = \frac{\sum_{j=1}^{j=y} M_j(P_k)}{|j|} \quad ,\forall j,$$
   $$r = record$$
   $$M_j \in r^{th} cluster$$
4. The Fuzzy measures are applied to all cluster of medical data for similarity calculation using below equation.
   $$Sim\,(M_i, C[r]) = \prod_{k=1}^{k=m} exp^{-\frac{(M_i(P_k) - \mu_r^k)}{\sigma}^2}$$
   where μ_r^k = Mean value of r^{th}cluster with k^{th}attribute
5. Similarity value of each group is characterized by cluster center. In this stage, number of class label is equal to the dimensionality of medical record.
6. Calculate the similarity value of Group 1 with Group 2 for each record.
7. Imputation process is applied on incomplete data and fills the value from medical dataset accordingly.
8. If the imputation process is stop, when all the incomplete values are imputed in dataset.
9. Final imputed data is achieved.

End of algorithm

3.2 Class-Based Clustering Classifier Classification

Medical diagnosis performance has been enhanced by Class-Based Clustering Classifier which integrates Support Vector Machine (SVM) and Artificial Immune Recognition System (AIRS). For instance, the hybrid AIRS and SVM were used as classifier where the AIRS was cable for reducing the computational complexity while maintaining accuracy of results and the SVM concentrated on classifying the different disease patterns quickly and accurately.

AIRS uses k-Nearest Neighbour as a classifier for data clustering. This is to be noted that k-NN classifier doesn't require ant predefined pattern or data in machine learning. This leads to low accuracy. Therefore, in this C^3IMD method, SVM Classifier is used instead of k-NN classifier. It makes a random base called memory cell pool (M) and sustains the pool of cells, which are organized through showing the system to a one-shot iteration of the training data. Providing that the memory cell is incompetently motivated for a given input attribute, candidate memory cells are organized. Maximum stimulated memory cells undertake a procedure of cloning and mutation. Then an algorithm provides resources in the development process of a candidate (Medical Attribute) memory cell. Amount of resources in each cell and its stimulating value is used for clone formation with each other. Resource competition is needed to manage the Artificial Recognition Ball (ARB) pool's size, in addition to stimulate such ARBs that have greater similarity (stimulation) for the antigen that the model is being trained on.

Table 1. Parameters Used for Class-Based Clustering Classifier based SVM–AIRS

Parameters	C^3IMD	ICBCC
Affinity Threshold	0.2	0.2
Clonal Rate	10.0	10.0
Hyper-Mutation Rate	2.0	2.0
Seed Cell	1	-
Stimulation Value	0.5	-
Total Resources	150	150
SVM Type	V-SVC	LMNN
Kernel Function	RBF	-
γ	1	0.5
Ca	7	-
Cash Memory Size	100 MB	100
Sampling Points	-	1500
Splits	-	75/25
Input Dimensionality d	-	775
Output Dimensionality	-	50

The main object was to improve a memory cell that is maximum effective in classifying a given antigen (Training Data) with high accuracy. Then the potential candidate memory cell is presented into the set of previously recognized memory cells, for training. Memory cell candidate (Incomimg data) will be added to cell's set when it is more similar than memory match in training antigen. If the affinity between memory cell candidate and memory cell match is less than a threshold, then memory cell candidate substitutes memory cell match in the pool of memory cells. The above process repeats until all training data have been presented to the system.

This research used Waikato Environment for Knowledge Analysis (WEKA) tool to discourse the classification problem, which is the package of Lib SVM in WEKA environment. It is to be noted that the Weka LIBSVM (WLSVM) is one of the package of Lib SVM that has been used in our research. Parameter details for implementation of SVM-AIRS algorithm is defined in Table 1.

Pseudo Code for Hybrid SVM-AIRS

I. Initialization and normalization of input dataset.
II. Kernel the memory cell pool (M), if preferred.
III. For each training attribute (antigen) do the following:
   1. If M is void, add antigen (Training) to M.
   2. Choose the memory cell (mc) in M of the same classification having the highest similarity to antigen.

3. Clone mc in fraction to its similarity to antigen.
4. Modify every clone and add to the B-cell pool (ARB).
5. Distribute resources to ARB. Eliminate the weak cells (population manage of ARB).
6. Compute the average stimulation of ARB to antigen and check for elimination. If the elimination condition is satisfied, go to step 9.
7. Random selection of B-cells in ARB is applied to cloning and mutation based on their stimulation.
8. Go back to step 5.
9. Choose the B-cell in ARB with the maximum similarity to antigen (training data). If training data has a higher similarity to antigen than mc, add training data to M. If mc and candidate are appropriately similar, then eliminate mc from M. Return M:make content of M for SVM

IV. Make M to format: run SVM as classifier
V. Implement SVM classification using M.

4. Improved Class-Based Clustering Classifier (ICBCC)

The prime objective of the proposed Classifier, Improved Class-Based Clustering Classifier (ICBCC) is to reduce Error Rate for better classification accuracy and F-Score.
The Architecture of the proposed machine learning algorithm ICBCC is shown in the Fig. 4. This is the enhanced version of our previous work Class-Based Clustering Classifier (C^3IMD).
We updated the second stage of Class-Based Clustering Classifier (C^3IMD) through existing Large Margin Nearest Neighbor (LMNN).
The proposed machine learning algorithm ICBCC optimizes the patterns to maximize classification performance. That is the proposed Classifier has two sampling approaches namely i. Target Neighbors and Imposters.

1. **Target Neighbors:** Let us consider Class Label Yi, Cluster Set S, Targetted Neighbors k and input Xi. The minimal distance can be calculated as follows.

$$D(\overrightarrow{x_i}, \overrightarrow{y_i}) = \|L(\overrightarrow{x_i} - \overrightarrow{y_i})\|^2$$

2. **Imposters:** The Learning Technique will minimize all Imposters that do not similar to Targeted Pattern which for all Input Data in the Training Set.

Initially, training input will have i. Impostors and ii. Target Neighbours. These two may be inside a same local Neighbourhood. During learning, the impostors will be moved to outside Cluster. The detailed process of the proposed Classifier is shown in the Fig. 3.

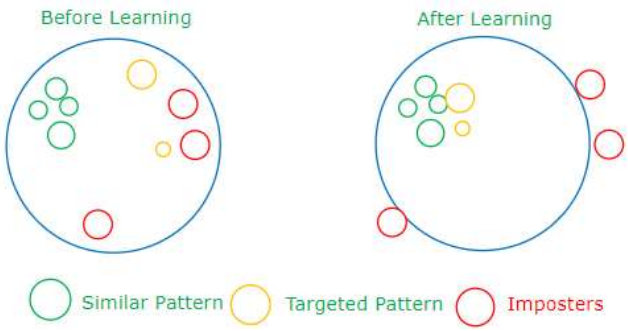


Fig. 3: Process of Proposed ICBCC - Schematic Illustration

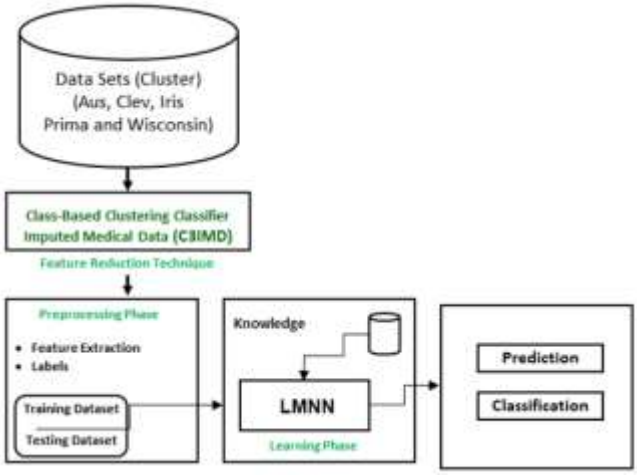


Fig. 4: Proposed Improved Class-Based Clustering Classifier (ICBCC)

After learning, the cluster Region will be generated needed samples clustering and impostors. That is, the proposed model will fix the target neighbours and it will be referred during learning processes. The similarities labelled inputs are clustered together. All the Imposters (Non-Similarities labelled) are removed and those Imposters will be minimized in the Cluster. That is similar labelled inputs will be pulled together to form Cluster and a Imposters will be pushed away further apart. Thus, the proposed Classifier maximizes Classification Accuracy and FSore as well.

5. Result and Discussion

The proposed model was implemented in Bio-Weka and studied thoroughly. The simulation parameters have given in the Table. 1. The experiments were repeated and consolidated the reports in terms of Classification Accuracy, Sensitivity, Specificity and F-Score. This research work used different data sets such as AUS, CLEV, IRIS, PRIMA and WISCONSIN for Evaluation and Validation.
From the experimental results, it was noticed that the proposed model Improved Class-Based Clustering Classifier (ICBCC) achieves the best Classification Accuracy, Sensitivity and Specificity and FScore as shown in the Table 2 that is the consolidated report if the repeated experiments.

Table – 2 Performance Evaluation Summary of classifiers

Method	Accuracy	Sensitivity	Specificity	F-Score
ICBCC	100%	100%	100%	**0.94**
C^3IMD	100%	100%	100%	**0.92**

Fig. 6: F-score of proposed C^3IMD Classifier

6. CONCLUSION

This research work is proposed an efficient Classifier, called Improved Class-Based Clustering Classifier (ICBCC) to improve Classification and Prediction Accuracy in Medical Data Sets. The proposed ICBCC is implemented in Bio Weka and studied thoroughly. Missing data were filled in Medical Data Sets efficiently with the help of proposed Cluster-Classifier Model to improve prediction accuracy. The experiments are repeated with various datasets and results are evaluated and compared with existing our previous classifier C^3IMD. From the experimental results, it was revealed that the proposed Classifier is outperforming FScore.

References

[1] Mohammad Mahmudur Rahman Khan, Rezoana Bente Arif, Md. Abu Bakr Siddique, and Mahjabin Rahman Oishe, "Study and Observation of the Variation of Accuracies of KNN, SVM, LMNN, ENN Algorithms on Eleven Different Datasets from UCI Machine Learning Repository," IEEE 4th International Conference on Electrical Engineering and Information and Communication Technology, 2018.

[2] Kilian Q.Weinberger, Lawrence K. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification," Journal of Machine Learning Research, Vol. 10, Pp. 207-244, 2009

[3] Ali Kalantari and et. al, "Computational intelligence approaches for classification of medical data: State-of-the-art, future challenges and research directions," International Journal of Neuro Computing, Pp. 1- 21, (2017).

[4] UshaRani Yelipe, Sammulal Porika, Madhu Golla, "An efficient approach for imputation and classification of medical data values using class-based clustering of medical records," International Journal Computers and Electrical Engineering, Pp. 1-18, (2017).

[5] Zhang S , Qin Z , Ling C , Sheng S, "Missing is useful: missing values in cost-sensitive decision trees," IEEE Transaction on Knowledge Data Engineering, 17(12): Pp. 1689–93, (2005).

[6] Zhang C , Qin Y , Zhu X , Zhang J , Zhang S, "Clustering-based missing value imputation for data preprocessing," IEEE International Conference On Industrial Informatics, Pp. 1081–6, (2006).

[7] Wang L , Fu D , Li Q , Mu Z, "Modelling method with missing values based on clustering and support vector regression," Journal of Systems Engineering and Electronics, 21(1), Pp.142–7, (2010).

[8] Kirkpatrick B , Stevens K, "Perfect phylogeny problems with missing values," IEEE/ACM Transaction on Computational Biology Bioinformatics, 11(5), Pp. 928–41, (2014).

[9] Choong M K , Charbit M , Yan H, "Autoregressive-model-based missing value estimation for DNA microarray time series data," . IEEE Transactions on Information Technology in Biomedicine, 13(1), Pp. 131–7, (2009).

[10] Zhang Z, "Missing data imputation: focusing on single imputation," . Annals of Translational Medicine, 4(1), (2016).

[11] D. Boyd , K. Crawford, "Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon," Information, Communication & Society, 15 , Pp. 662–679, (2012) .

[12] X. Wu , X. Zhu , G.-Q. Wu , W. Ding, "Data mining with big data," IEEE Transactions on Knowledge and Data Engineering, 26, Pp. 97–107, (2014).

[13] I.A.T. Hashem , I. Yaqoob , N.B. Anuar , S. Mokhtar , A. Gani , S.U. Khan , "The rise of "big data" on cloud computing: review and open research issues," Journal of Information System, 47, Pp. 98–115, (2015) .

[14] V. Mayer-Schönberger , K. Cukier, "Big Data: A Revolution That will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt," 2013 .

[15] A . Gani , A . Siddiqa , S. Shamshirband , F. Hanum , "A survey on indexing techniques for big data: taxonomy and performance evaluation," Knowledge and Information Systems, 46, Pp.241–284, (2016).

[16] Lewis HD, "Missing data in clinical trials," New England Journal of Medicine, Pp. 2557–8, (2012) .

[17] 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 RBFs and event covering method," Neural Networks. 23 406–418,(2009).

[18] Wang L , Fu D , Li Q , Mu Z, "Modelling method with missing values based on clustering and support vector regression," Journal of Systems Engineering and Electronics, 21(1), Pp.142–7, (2012).

[19] Zhu X , Zhang S , Zhi J , Zhang Z , Xu Z, "Missing value estimation for mixed-attribute data sets," IEEE Transactions on Knowledge and Data Engineering, 23(1), Pp.110–21, 2013.

[20] Farhangfar A , Kurgan L , Pedrycz, "A novel framework for imputation of missing values in databases," IEEE transactions on systems, man, and cybernetics, 37(5), Pp.692–709, (2017). .

[21] Choong MK , Charbit M , Yan H, "Autoregressive-model-based missing value estimation for DNA microarray time series data," IEEE Transactions on Information Technology in Biomedicine, 13(1), Pp.131–7, (2009). .

[22] Thirukumaran S, Sumathi A. "Improving accuracy rate of imputation of missing data using classifier methods," International Conference On Intelligent Systems And Control (ISCO), Coimbatore; Pp.1–7, (2016).

[23] Razavi-Far R, Saif M, "Imputation of missing data using fuzzy neighborhood density-based clustering," IEEE international conference on fuzzy systems (FUZZ-IEEE), Pp. 1834–41, (2016).

[24] Aljuaid T, Sasi S., "Proper imputation techniques for missing values in data sets,"International Conference On Data Science And Engineering (ICDSE), Pp. 1–5, (2016)

[25] Gira A, "Estimation of population mean with a new imputation method," Applied Mathematical Sciences, 9(34), Pp.1663–72, 2015.

[26] Nishanth KJ , et al, "Soft computing based imputation and hybrid data and text mining: the case of predicting the severity of phishing alerts," Journal of Expert Systems With Applications, 39(12), Pp. 10583–9, (2012).

[27] Tang F , Ishwaran H, "Random forest missing data algorithms," Journal of Statistical Analysis and Data Mining (2017).

[28] Petrozziello A , Jordanov I , "Column-wise guided data imputation,". Procedia Computer Science, 108, Pp. 2282–6, (2017).