IADIS International Journal on Computer Science and Information Systems
Vol. 12, No. 1, pp. 61-74
ISSN: 1646-3692

A CLASS BASED CLUSTERING APPROACH FOR IMPUTATION AND MINING OF MEDICAL RECORDS (CBC-IM)

Porika Sammulal[1] ,Yelipe UshaRani[2], Anurag Yepuri[2]
[1]Computer Science and Engineering, JNTUH College of Engineering, Jagityal, India
[2]Department of Information Technology, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India

ABSTRACT

Disease prediction and classification using medical record datasets is a challenging data mining research problem that also requires attribute value imputation to be carried implicitly. Medical datasets that are available in public databases are not free from missing values and this is also true when data is collected and sampled through various clinical trials. In this context, there is always a need to turn up with new approaches and methods for accurate and efficient analysis of medical records. Several imputation strategies are proposed in the literature and each of them have reported accuracies achieved on benchmark datasets. However, a better imputation approach always helps in improving classification accuracies and this in turn helps to more accurate disease prediction. An approach for imputing medical records is proposed in this paper. We name the approach as Class-Based-Clustering-Imputation (CBC-IM). Experiments are carried out on several benchmark datasets. Results achieved using our imputation approach is compared to existing imputation approaches using classifiers such as KNN, SVM and C4.5. The results show improved performance on most of the datasets and are almost nearer to remaining approaches discussed in this paper.

KEYWORDS

Classifier, Medical Record, Prediction, Outliers, Imputation

1. INTRODUCTION

Analyzing medical data for disease prediction and classification is most challenging task because of hidden challenges that generate under various practical conditions and requires addressing such implicit challenges. In many cases, when there is a need to study medical records for disease prediction, the most obvious and widely accepted fact is that attributes of medical records are not free from missing values. In this scenario, the process of imputation received wide attention from data mining and data analysts.

Preprocessing of medical data records is the initial task that may be carried to eliminate biasing. Also, medical records must be normalized and scaled to make them suitable for performing effective analysis. Normalization must be done before imputation so that the imputation process yields correct results. This also affects classification accuracies. Dimension of medical records is another important challenge and concern. We must see that dimensions which are not affecting the final accuracies be only eliminated or discarded.

In Zhang S., Zhenxing Qin, Ling C. X, & Sheng S. (2005) the researchers debate whether to consider missing values or simply eliminate them from consideration for analysis in the context of decision trees. Clustering data records is a known problem and is also applied for handling medical datasets (Zhenxing Qin, Shichao Zhang, & Chengqi Zhang, 2006) to handle missing values applying clustering. Imputation using support vector regression and clustering is discussed in Wang Ling, Fu Dongmei, Li Qing, & Mu Zhichun (2010). An approach for arabic text classification is proposed in Bilal Hawashin, Ayman Mansour, & Shadi Aljawarneh (2013). Aljawarneh, S., Yassein, M. B., & Talafha, W. A (2017); Aljawarneh, S., (2011); Aljawarneh, S. A., Moftah, R. A., & Maatuk, A. M. (2016) propose an algorithm for encrypting big data which can be applied for privacy preserving of medical data, cloud computing applications and detecting worm signatures. Rajesh Kumar Gunupudi, Mangathayaru Nimmala, Narsimha Gugulothu, & Suresh Reddy, Gali (2017) propose a method for feature reduction and transformation. In Xiaofeng Zhu, Zhang S, Zhi Jin, Zili Zhang, & Zhuoming Xu (2011) authors discuss various problems arising from missing values. Handling mixed attributes is studied in Xiaofeng Zhu, Zhang S, Zhi Jin, Zili Zhang, & Zhuoming Xu (2011). In Farhangfar A, Kurgan L.A, & Pedrycz (2007), a novel framework for handling missing values is proposed. A discussion on using auto regression method to handle missing values is done in Miew Keen Choong, Charbit M, & Hong Yan, (2009). The works in Qiang Yang, Ling C, Xiaoyong Chai, & Rong Pan, (2006); Atif Khan, John A Doucette, & Robin Cohen (2013) outlines challenges associated with medical records w.r.t cost sensitiveness and decision support systems.

In our previous works, (UshaRani Y, & Sammulal P, 2015; UshaRani Y, & Sammulal P, 2016; Yelipe UshaRani, & Sammulal P., 2016) a new approach to impute missing values is proposed. Some of the other related works on medical datasets which motivated the current work are Jau-Huei Lin, & Peter J. Haug (2008); Karla L, Caballero Barajas, & Ram Akella (2015); Wei-Chao Lin, Shih-Wen Ke, & Chih-Fong Tsai (2015). New distance measures for temporal pattern mining are proposed in Vangipuram Radhakrishna, P. V. Kumar, V. Janaki & S. Aljawarneh, (2016) ; Shadi Aljawarneh et al. (2017); Vangipuram Radhakrishna, Shadi A. Aljawarneh, Puligadda Veereswara Kumar, Kim-Kwang Raymond Choo (2016).

2. CLASS BASED CLUSTERING IMPUTATION (CBC-IM)

Our class based clustering imputation approach **(CBC-IM)** targets imputing missing attribute values and performing disease prediction and classification after performing the imputation. Initially all medical records which do not have missing values are considered and clustered by applying k-means clustering technique. The number of clusters is equal to number of labels in the dataset considered. All other medical records which have missing values are separated from the dataset and placed in a group. Now, the distance from each of these records (both having missing values and not having missing values) to the Centre of every generated cluster is obtained. This is done to achieve dimensionality reduction of records to a dimension equal to number of class labels. The result is that each transformed record now represents a vector of values. This is then followed by finding distance between these transformed records and missing attribute records. The imputation is done considering record to which the test record distance is minimum.

2.1 Research Objective

To impute missing attribute values in a given dataset of medical records and then perform disease prediction and classification. This is achieved by targeting dimensionality reduction of initial dimension of medical records. i.e we map initial m-dimensional records to p-dimensional records where p is the number of decision classes in original dataset. Missing values are imputed using these transformed low dimensional representations of medical records.

2.2 Threshold and Deviation

Given, a dataset of medical records having both missing and non-missing attribute values, we aim to fill all missing attribute values of medical records and achieve high classification accuracies on these datasets, so as prove proposed method is feasible and adoptable for classification and imputation.

2.3 Imputation Approach

The proposed imputation approach is discussed in this subsection. Initially, medical records are categorized into two different groups, those without missing attribute values (G1) and another group (G2) having missing attribute values. The objective is to achieve effective dimensionality reduction and imputation which can hence improve classification accuracies of classifiers. The idea behind CBC-IM approach is to consider all records in group G1 (having no missing values) and first obtain the number of clusters equal to the number of decision labels in the original dataset. For each of these clusters which are generated by considering records in first group (G1),their corresponding cluster mean is computed. Each value in p-dimension vector is the distance of medical record from cluster center. This will be the dimension of medical records (p-dimension) in group, (G1) for performing medical records imputation, disease prediction and classification. Secondly, for imputation of attribute value of records in group (G2), each medical record in G2 is now transformed to its equivalent p-dimensional representation but discarding those attributes having missing values. At the end of these two steps, we have high dimensional medical records expressed as equivalent low dimensional vectors. These transformed record representations are actually considered for imputation. The distance between each transformed record to be imputed in G2is computed to every transformed record in G1. The imputation is done considering record in G1 which is having minimum distance. The best approach for imputation is to consider decision class of medical record to be imputed and then perform imputation considering medical record to which this record has minimum distance w.r.t that class. In the case of numerical attribute, we can fill the mean of attribute value. After imputation is done, we have final set of medical records, with no missing values. This record set can be then used for finding classification accuracies. For classification, we use same procedure adopted for imputation but determine class labels, instead of imputing missing values. The importance of the present approach is that we can impute and classify medical records by reducing dimensionality.

3. IMPUTATION ALGORITHM

Step-1: Divide dataset into two groups

Divide medical records into two groups. The first group, $\mathbf{G_1}$ includes all medical records which do not have missing values and second group, $\mathbf{G_2}$contains medical records which have missing values.

Step-2: Cluster medical records in group, $\mathbf{G_1}$

Let, g = $|D_d|$ be the total number of decision classes. Cluster medical records in the group,$\mathbf{G_1}$to the number of clusters equal to $|D_d|$.

Step-3: Find cluster mean for each generated cluster

Consider each generated cluster of medical records obtained from $\mathbf{G_1}$and obtain their respective mean vectors. For example, Let C^g denote g^{th} cluster with four medical records R_3, R_4, R_6, and R_8defined over two attributes, $A_1$and A_2. The cluster mean is hence equal to

$$\mu_g = \left(\frac{R_3(A_1) + R_4(A_1) + R_6(A_1) + R_8(A_1)}{4}, \frac{R_3(A_2) + R_4(A_2) + R_6(A_2) + R_8(A_2)}{4}\right) = \left(\mu_g^1, \mu_g^2\right)$$

Step-4: Compute distance between each record,$\boldsymbol{R_i}$and cluster center

Let,$\boldsymbol{R_i}$and$\boldsymbol{\mu_g}$ be 'n' dimensional vectors. Find the distance from each,$\boldsymbol{R_i}$in $\mathbf{G_1}$to the mean,$\boldsymbol{\mu_g}$ of every generated cluster. The distance can be obtained by finding Euclidean distance between each medical record,$\boldsymbol{R_i}$ and mean vector,$\boldsymbol{\mu_g}$ of every cluster. Similarly, find the distance from each,$\boldsymbol{R_i}$ in $\mathbf{G_2}$to the mean,$\boldsymbol{\mu_g}$ of every generated cluster by considering only those attributes having no missing values.

Step-5: Express records as g-dimensional vectors

Distance values computed from each medical record to the mean of every generated cluster must be expressed as g-dimensional vectors. This g-dimensional vector representation of record is now used to perform imputation (and may also be used for classification).

Step-6: Find similarity of record to be imputed using proposed measure

For each g-dimensional medical record, $\mathbf{R_m}$ in group G_2consisting of missing attribute value(s), find the similarity of this record $\mathbf{R_m}$to each record,$\mathbf{R_i}$in G_1using proposed measure in section-4.Let,$\mathbf{R_i}$be the medical record in G_1 to which the record, $\mathbf{R_m}$has maximum similarity. In such a case record, R_i(in G_1) is best choice for imputation.

Step-7: Perform Imputation

For categorical attributes, we may impute the corresponding attribute value. For numeric attributes, we may impute either frequency or same element value as that of the corresponding attribute w.r.t missing record. For similarity estimation, we use proposed measure instead of traditional Euclidean distance measure widely adopted.

4. SIMILARITY MEASURE

Let, R_m and R_n be two medical records expressed as p-dimensional vectors denoted by $R_m = (R_{m_1}, R_{m_2}, R_{m_3}, \dots\dots\dots\dots\dots.., R_{m_p})$ and $R_n = (R_{n_1}, R_{n_2}, R_{n_3}, \dots\dots\dots\dots., R_{n_p})$. These record representations are obtained from step-5 of CBC-IM algorithm.

The similarity between two medical records is obtained using the similarity function defined by equation (1)

$$IM.Sim(R_m, R_n) = \boldsymbol{e}^{-(\frac{R_{m_1} - R_{n_1}}{\sigma_1})^2} * \boldsymbol{e}^{-(\frac{R_{m_2} - R_{n_2}}{\sigma_2})^2} * \dots\dots\dots\dots\dots\dots * \boldsymbol{e}^{-(\frac{R_{m_p} - R_{n_p}}{\sigma_p})^2} \qquad (1)$$

where, σ_p is the standard deviation of all corresponding attributes column values of p-dimensional vector.

5. CASE STUDY

In this section, we discuss how to impute missing values considering medical dataset records in Table 1. We first normalize the medical record dataset, shown in Table 2. This is achieved by first replacing the categorical data attributes Z1 and Z3. The column Z1 has 3 distinct values and Z3 has 2 distinct values. We assign d_{11}=1, d_{12}=2, d_{13}=3 and h_{31}=1, h_{32} = 2 for attribute values in Table 2. In Table 2, NZ1 and NZ3 denote normalized data attribute values.

Table 3 shows medical records with missing and non-missing attribute values. Table 3 is split into two tables Table 4 and Table 5. Table 4 consist records with missing attribute values. Medical records in Table 5 are clustered into two clusters say C1 and C2. This is because in our case, medical records are of only two classes.

Table 1. Medical Dataset

Records	Attributes				Disease Level
	Z_1	Z_2	Z_3	Z_4	
R_1	d_{11}	5	h_{31}	10	L1
R_2	d_{13}	7	h_{31}	5	L1
R_3	d_{11}	7	h_{32}	7	L1
R_4	d_{12}	5	h_{31}	10	L1
R_5	d_{13}	3	h_{32}	7	L2
R_6	d_{12}	9	h_{31}	10	L2
R_7	d_{11}	5	h_{32}	3	L2
R_8	d_{13}	6	h_{32}	7	L2
R_9	d_{12}	6	h_{32}	10	L2

Records from R1 to R4 belong to Class, C1 and R5 to R9 belong to class, C2. Clusters obtained are shown in Table 6 consisting records R2, R7, R8 in cluster-1 and R9, R6, R4, R1 in cluster-2. Table 7 represents mean of clusters computed for two clusters C1 and C2. Table 8 and Table 9 represents records in both clusters.

Table 2.Normalized Records

Normalized Records	Attributes			
	NZ_1	Z_2	NZ_3	Z_4
NR_1	1	5	1	10
NR_2	3	7	1	5
NR_3	1	7	2	7
NR_4	2	5	1	10
NR_5	3	3	2	7
NR_6	2	9	1	10
NR_7	1	5	2	3
NR_8	3	6	2	7
NR_9	2	6	2	10

Table 3. Normalized Records With and Without MVs

Record	NZ_1	Z_2	NZ_3	Z_4
NR_1	1	5	1	10
NR_2	3	7	1	5
NR_3	**1**	**7**	**?**	**7**
NR_4	2	5	1	10
NR_5	**3**	**3**	**2**	**?**
NR_6	2	9	1	10
NR_7	1	5	2	3
NR_8	3	6	2	7
NR_9	2	6	2	10

Table 4. Records to be Imputed

Record	Z_1	Z_2	Z_3	Z_4
NR_3	1	7	?	7
NR_5	3	3	2	?

Table 5. Records Free from Missing Values

Record	Z_1	Z_2	Z_3	Z_4
NR_1	1	5	1	10
NR_2	3	7	1	5
NR_4	2	5	1	10
NR_6	2	9	1	10
NR_7	1	5	2	3
NR_8	3	6	2	7
NR_9	2	6	2	10

Table 6. Clusters

Cluster	Medical Records
Cluster-1	NR7;NR8;NR2
Cluster-2	NR4;NR6;NR1;NR9

Table 7. Generated Clusters with Mean

Cluster	Mean $_1$	Mean$_2$	Mean$_3$	Mean$_4$
Cluster-1	2.33	6	1.67	5
Cluster-2	1.75	6.25	1.25	10

Table 8. Records in First Cluster

Record	Z_1	Z_2	Z_3	Z_4
NR_2	3	7	1	5
NR_7	1	5	2	3
NR_8	3	6	2	7

Table 9. Records in Second Cluster

Record	Z_1	Z_2	Z_3	Z_4
NR_1	1	5	1	10
NR_4	2	5	1	10
NR_6	2	9	1	10
NR_9	2	6	2	10

Table 10. Distance of records in Table.5 to Clusters

Record	Cluster-1	Cluster-2
NR_1	5.312459	1.47902
NR_2	1.374369	5.214163
NR_4	5.153208	1.299038
NR_6	5.878397	2.772634
NR_7	2.624669	7.189402
NR_8	2.134375	3.344772
NR_9	5.022173	0.829156

Table 11. Distance of Records in Table.4 to Clusters

Record	Cluster-1	Cluster-2
NR_3	2.603417	3.181981
NR_5	3.091206	3.561952

Table 10 represents distance of records free from missing values to the two clusters generated. In a Similar way, distance of records with missing attribute values to generated clusters is recorded in Table 11. Each record is represented as 2-D vector. Table 12 and Table 13 records similarity of R3 and R5 to all other records expressed as 2D vectors. The last column of Table 12 and Table 13 denotes the similarity score obtained using proposed fuzzy similarity measure. Imputed attribute values are recorded in Table 14 and Table 15. The values are imputed by choosing those records to which similarity is the maximum. In our case, the similarity score of both R3 and R5 is the maximum to record R8. Hence, record R8 is the better choice for carrying out the imputation.

Table 12. Similarity of R3 w.r.t Cluster-1 Records

Record	Similarity to c1	Similarity to c2	Final Similarity
NR_1	0.000000	0.000000	0.000000
NR_2	0.022234	0.000000	0.000000
NR_4	0.000000	0.000000	0.000000
NR_6	0.000000	0.307716	0.000000
NR_7	0.998863	0.000000	0.000000
NR_8	**0.574456**	**0.829944**	**0.476766**
NR_9	0.000000	0.000000	0.000000

Table 13. Similarity of R5 w.r.t Cluster-1 Records

Record	Similarity to c1	Similarity to c2	Final Similarity
NR_1	0.000004	0.000000	0.000000
NR_2	0.000595	0.000000	0.000000
NR_4	0.000022	0.000000	0.000000
NR_6	0.000000	0.012499	0.000000
NR_7	0.577859	0.000000	0.000000
NR_8	**0.099576**	**0.717665**	**0.071462**
NR_9	0.000083	0.176880	0.000015

Table 14. Imputed Record, R3

Record	NZ_1	Z_2	NZ_3	Z_4
NR_3	1	7	2	7

Table 15. Imputed Record, R5

Record	NZ_1	Z_2	NZ_3	Z_4
NR_9	3	3	2	**7**

6. RESULTS AND DISCUSSIONS

This section outlines experimentation results obtained from various experiments performed using benchmark datasets. Table 16 shows classification accuracies achieved using classifiers C4.5, SVM and KNN. Classification is performed after performing imputation using both the proposed imputation approach and some of the existing imputation techniques in the literature. The following insights can be seen from the table 16 below.

1. The accuracy achieved on AUS dataset with SVM classifier using CBC-IM approach is comparatively better than MV-BPCA, MV-EM, MV-FKMeans, MV-Ignore, MV-Kmeans, MV-KNN, MV-Most Common, MV-SVDimpute, and MV-WKNNimpute approaches.
2. SVM Classifier accuracies achieved on CLEVELAND dataset with CBC-IM approach is comparatively better than MV-BPCA, MV-EM, MV-FKMeans, MV-Ignore, MV-KNN, MV-Most Common, MV-SVDimpute, MV-SVMImpute and MV-WKNNimpute approaches.
3. The accuracy achieved on ECOLI dataset with CBC-IM approach and SVM Classifier is comparatively better than MV-BPCA, MV-EM, MV-FKMeans, MV-Kmeans, MV-Ignore, MV-KNN, MV-Most Common, MV-SVDimpute, and MV-WKNNimpute approaches.
4. Also, accuracies achieved for GER, HEP , WINE datasets are also comparatively better to MV-BPCA,MV-EM,MV-FKMeans,MV-Ignore, MV-KNN, MV-SVDimpute approaches
5. Accuracies on WISCON dataset is 96.4% which is better compared to all other imputation approaches and has almost same accuracies w.r.t remaining imputation approaches.

Table 16. SVM Classifier Accuracies on Benchmark Datasets Using CBCIM-Fuzzy Measure and Various Imputation Strategies

Approaches	AUS	CLE	ECOLI	GER	HEP	WINE	IRS	PIM	NEWTHY	WISCON
CBC-IM-FUZZY	85.4	58.8	70.2	77.2	85.1	84.9	92.6	69.6	84.1	96.4
MV-BPCA	76.1	50.2	53.8	67.9	79.3	59.6	81.3	75.9	87.4	59.5
MV-EM	77.8	52.8	52.9	71.4	79.3	65.2	82	69.0	85.8	95.1
MV-FKMeans	81.0	52.4	78.5	70.5	80	71.9	94.5	76.1	83.2	96.5
MV-Ignore	81.9	53.8	53.4	70.5	83.7	54.7	93.1	67.0	80.5	96.0
MV-Kmeans	82.5	54.4	62.2	72	82.5	78.1	78.1	63.5	86.5	96.1
MV-KNN	83.9	52.4	64.5	72.1	80	75.2	94.7	64.7	83.7	95.7
MV-Most Common	83.3	52.4	60.7	70.7	81.2	67.4	92.7	65.7	84.6	95.9
MV-SVDimpute	76.1	53.4	54.4	70.5	83.2	61.7	83.8	65.6	80	94.9
MV-SVMImpute	88.9	53.1	73.2	81.8	92.2	89.8	94.7	89.8	88.3	95.9
MV-WKNNimpute	84.2	52.4	65.4	71.3	80.6	72.4	94	64.7	84.1	95.7

Classifier accuracies on benchmark datasets that includes Australian (AUS) , Cleveland (CLE), ECOLI, GERMAN (GER), HEPATITIS (HEP), WINE, IRIS, PIMA, NEWTHYROID and WISCON are shown in Table 17. Classifiers chosen are C4.5, SVM and KNN. The distance function used in classifiers is Euclidean and proposed fuzzy measure (exponential, EXP).

Table 17. C4.5, SVM and KNN Classifier Accuracies on Benchmark Datasets Using CBCIM-fuzzy Measure and Various Imputation Strategies

DATASETS / IMPUTATION	C4.5 CBCIM-EUC	SVM CBCIM-EUC	KNN CBCIM-EUC	C4.5 CBCIM-EXP	SVM CBCIM-EXP	KNN CBCIM-EXP
AUSTRALIAN(AUS)	86.16	85.24	85.77	85.94	85.36	84.63
CLEVELAND(CLE)	51.15	54.12	55.77	51.81	53.79	56.1
ECOLI	60.11	70.23	64.88	60.41	70.23	64.58
GERMAN(GER)	70	76.9	74.2	70	77.2	74.6
HEPATITIS(HEP)	78.06	87.74	85.8	78.06	85.16	85.16
WINE	41.01	78.65	80.33	39.88	84.88	82.02
IRIS	92.6	92.6	84.66	93.33	92.66	86.66
PIMA	65.1	69.66	70.96	65.18	69.62	70.92
NEWTHYROID	69.76	85.58	84.18	69.76	84.18	83.72
WISCON	91.54	96.41	95.7	91.54	96.41	95.7

Figure 1 shows accuracies of C4.5 classifier on AUS dataset listed in Table 16 using CBC-IM method with Euclidean distance measure. The accuracy result of CBC-IM approach with C4.5 classifier is better than all other imputation approaches. The accuracy achieved with CBC-IM applying C4.5 classifier is 86.16%. Least accuracy is recorded by SVD-Impute approach equal to 65.94%.

Figure 2 shows Comparison of C4.5 classifier accuracies on Benchmark datasets with CBC-IM approach using Euclidean distance measure and various imputation approaches. Figure 3 Compares C4.5, SVM and KNN classifier accuracies on Benchmark datasets using CBC-IM approach with Euclidean and Fuzzy measure. In this paper, the datasets are used from the weblink, http://keel.es/.

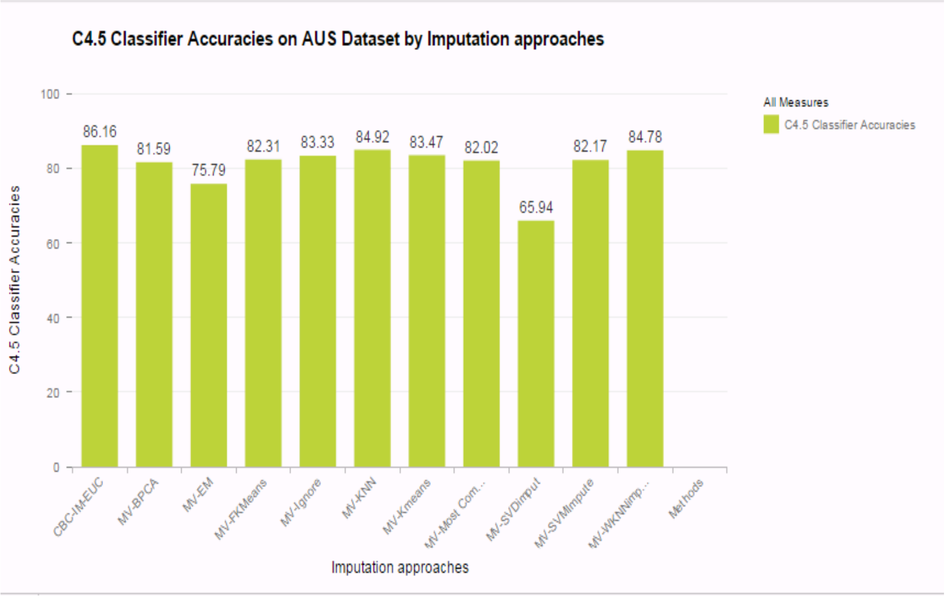

Figure 1. Comparison of C4.5 Classifier Accuracies on Benchmark Datasets with CBC-IM Approach Using Euclidean Distance Measure and Various Imputation Approaches

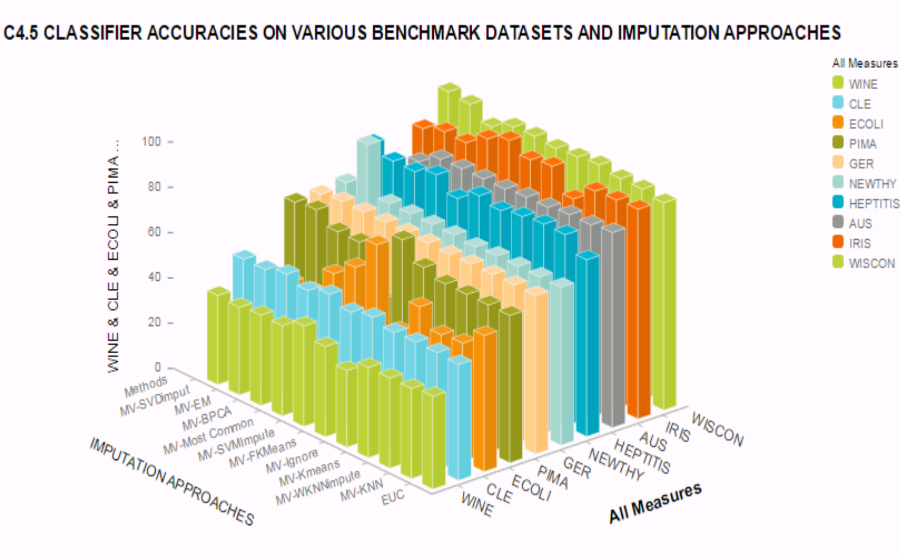

Figure 2. Comparison of C4.5 Classifier Accuracies on Benchmark Datasets with CBC-IM Approach Using Euclidean Distance Measure and Various Imputation Approaches

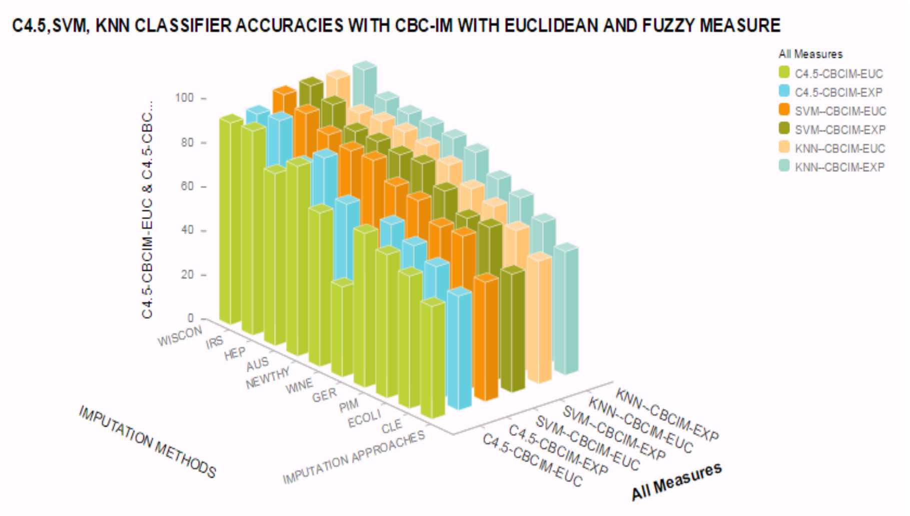

Figure 3. Comparison of C4.5, SVM and KNN Classifier Accuracies on Benchmark Datasets Using CBC-IM Approach with Euclidean and Fuzzy Measure

7. CONCLUSIONS

This research is aimed at devising a procedure to handle attribute missing values and impute them. A novel dimensionality reduction approach is proposed. Our approach first reduces dimensionality of all medical records by transforming them into p-dimensional records where the resultant dimensionality is the number of classes. The lower dimension medical records are later used to carry imputation. In present case, we use Euclidean measure and class based clustering concept to perform dimensionality reduction and apply proposed measure to find similarity between records to choose best record for imputation. Results show CBC-IM approach with Euclidean and proposed measure has been better compared to rest of the imputation approaches discussed in results section. Accuracies obtained using CBC-IM approach are comparatively better to MV-BPCA, MV-EM, MV-FKMeans, MV-Kmeans, MV-Ignore, MV-KNN, MV-Most Common, MV-SVDimpute, and MV-WKNNimpute approaches.

REFERENCES

Aljawarneh, S. (2011). Cloud security engineering: avoiding security threats the right way. *International Journal of Cloud Applications and Computing, 1*(2), 64-70.

Aljawarneh, S. A., Moftah, R. A., & Maatuk, A. M. (2016). Investigations of automatic methods for detecting the polymorphic worm signatures. *Future Generation Computer Systems, 60*, 67-77.

Aljawarneh, S., Yassein, M.B., & Talafha, W.A. (2017). A resource-efficient encryption algorithm for multimedia big data. *Multimedia Tools and Applications*, 1–22, doi:10.1007/s11042-016-4333-y

Atif Khan, John, A., Doucette, & Robin Cohen. (2013). Validation of an ontological medical decision support system for patient treatment using a repository of patient data: Insights into the value of machine learning. *ACM Transactions on Intelligent Systems and Technology*, *4*(4), 1–31.

Bilal Hawashin, Ayman Mansour, & Shadi Aljawarneh. (2013). An efficient feature selection method for Arabic text classification. *International journal of computer applications, 83*(17).

Farhangfar, A., Kurgan L., & Pedrycz, A. (2007). A Novel Framework for Imputation of Missing Values in Databases. In Part A: Systems and Humans, *Proceedings of the IEEE Transactions on Systems, Man and Cybernetics, 37*(5), 692-709.

Jau-HueiLin, Peter J., & Haug. (2008). Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. *Journal of Biomedical Informatics*, *41*(1), 1–14.

Karla, L., Caballero Barajas, & Ram Akella. (2015). Dynamically modeling patient's health state from electronic medical records: a time series approach. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15),* 69-78, doi: 10.1145/2783258.2783289

Kirkpatrick, B., & Stevens, K. (2014). Perfect Phylogeny Problems with Missing Values. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 11*(5), 928-941.

Luengo, J., García, S., & Herrera, F. (2010). 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.

Miew Keen Choong, Charbit, M., & Hong Yan. (2009). Autoregressive-model-based missing value estimation for dna microarray time series data. *IEEE Transactions on Information Technology in Biomedicine, 13*(1), 131-137.

Qiang Yang, Ling, C., Xiaoyong Chai, & Rong Pan. (2006). Test-cost sensitive classification on data with missing values. *IEEE Transactions on Knowledge and Data Engineering,18*(5), 626-638.

Rajesh Kumar Gunupudi, Mangathayaru Nimmala, Narsimha Gugulothu, &Suresh Reddy, Gali. (2017). CLAPP: A self constructing feature clustering approach for anomaly detection, *Future Generation Computer Systems*. doi:10.1016/j.future.2016.12.040

Shadi A. Aljawarneh, Vangipuram Radhakrishna, P.V.Kumar, V. Janaki. (2017). G-SPAMINE: An approach to discover temporal association patterns and trends in internet of things. *Future Generation Computer Systems,* Retrieved from http://dx.doi.org/10.1016/j.future.2017.01.013

Shobeir Fakhraei, Hamid Soltanian-Zadeh, Farshad Fotouhi, & Kost Elisevich. (2010). Effect of classifiers in consensus feature ranking for biomedical datasets. *Proceedings of the ACM fourth International Workshop on Data and Text Mining in Biomedical Informatics*, 67-68. doi: 10.1145/1871871.1871886

UshaRani, Y., & Sammulal, P. (2015). A novel approach for imputation of missing values for mining medical datasets. *Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research,* Madurai, 1-8, doi:10.1109/ICCIC.2015.7435816

UshaRani, Y., & Sammulal, P. (2016). An efficient disease prediction and classification using feature reduction based imputation technique. *Proceedings of the 2016 International Conference on Engineering & MIS* Agadir, 1-5, doi:10.1109/ICEMIS.2016.7745363

Vangipuram Radhakrishna, P. V. Kumar, V. Janaki & S. Aljawarneh. (2016). A similarity measure for outlier detection in time stamped temporal databases. *Proceedings of the 2016 International Conference on Engineering & MIS (ICEMIS),* 1-5.doi: 10.1109/ICEMIS.2016.7745347

Vangipuram Radhakrishna, P. V. Kumar, V. Janaki & S. Aljawarneh. (2016). A computationally efficient approach for temporal pattern mining in IoT. *Proceedings of the 2016 International Conference on Engineering & MIS (ICEMIS),* 1-4. doi:10.1109/ICEMIS.2016.7745354

Vangipuram Radhakrishna, Shadi A. Aljawarneh, Puligadda Veereswara Kumar, Kim-Kwang Raymond Choo. (2016). A novel fuzzy Gaussian-based dissimilarity measure for discovering similarity temporal association patterns. *Soft Computing*. First Online: 18 November 2016. doi:10.1007/s00500-016-2445-y

Wang Ling, Fu Dongmei, Li Qing, & Mu Zhichun. (2010). Modelling method with missing values based on clustering and support vector regression. *Journal of Systems Engineering and Electronics, 21*(1), 142-147.

Wei-Chao Lin, Shih-Wen Ke, & Chih-Fong Tsai. (2015). CANN: An intrusion detection system based on combining cluster centers and nearest neighbors. *Knowledge-Based Systems*, *78*, 13-21.

Xiaofeng Zhu, Zhang S, Zhi Jin, Zili Zhang, & Zhuoming Xu. (2011). Missing value estimation for mixed-attribute data sets. *IEEE Transactions on Knowledge and Data Engineering, 23(*1), 110-121.

Yelipe UshaRani, & Sammulal, P. (2016). An innovative approach for imputation and classification of medical records for efficient disease prediction. *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies,*1-6. doi: 10.1145/2905055.2905273

Zhang, C., Yongsong Qin, Xiaofeng Zhu, Jilian Zhang, & Zhang, S. (2006). Clustering-based missing value imputation for data preprocessing. *Proceedings of the IEEE International Conference on Industrial Informatics,* 1081-1086. doi: 10.1109/INDIN.2006.275767

Zhang, S, Zhenxing Qin, Ling C.X, & Sheng S. (2005). Missing is useful: missing values in cost-sensitive decision trees. *Proceedings of the IEEE Transactions on Knowledge and Data Engineering, 17*(12), 1689-1693.

Zhenxing Qin, Shichao Zhang, & Chengqi Zhang. (2006*).* Missing or absent? A question in cost-sensitive decision tree. *Proceedings of the 2006 conference on Advances in Intelligent IT: Active Media Technology,* IOS Press, *138,* 118-125.