Computers and Electrical Engineering 000 (2017) 1–18

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆

UshaRani Yelipe[a,*], Sammulal Porika[b], Madhu Golla[a]

[a] VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India
[b] JNTUH College of Engineering, Karimnagar, India

ARTICLE INFO

Article history:
Received 5 March 2017
Revised 21 November 2017
Accepted 22 November 2017
Available online xxx

Keywords:
Imputation
Medical record
Clustering
Classifiers
Missing values
Prediction

ABSTRACT

Medical data is usually not free from missing values and this is also true when data is collected and sampled through various clinical trials. Existing Imputation techniques do not address the problem of high dimensionality and apply distance functions that also have the curse of high dimensionality. There is a need to turn up with innovative approaches and methods for accurate and efficient analysis of medical records. This research proposes an improved imputation approach called IM-CBC (Imputation based on class-based clustering) and a classifier termed as the Class-Based-Clustering Classifier(CBCC-IM). Experiments are performed on nine benchmark datasets and the recorded results using IM-CBC imputation approach are compared to ten imputation approaches using classifiers KNN, SVM and C4.5 and to the CBCC classifier using Euclidean distance and fuzzy gaussian similarity functions. Results obtained prove that the performance of classifiers is improved or atleast nearer to the existing approaches. CBCC-IM classifier records highest accuracy when compared to all other classifiers on benchmark datasets such as Cleveland, Ecoli, Iris, Pima, Wine and Wisconsin.

1. Introduction

Medical Data Imputation is currently an active area of research. Imputation of medical records requires knowledge of statistical methods and application of data mining principles. Many times, mining of medical data for knowledge extraction requires handling missing values by performing imputation such that the imputed value results in better classification rates. Although, several imputation techniques are available, most of them fail to give better classification rates. One of the simplest (or default) technique to handle missing values is to just remove those records having missing values. This technique of handling incomplete medical records is suitable only when the number of such incomplete records is very less and there is no knowledge about the missing pattern. However, there is also a chance for information loss since the valuable data is also removed. Some common approaches for handling missing values are indicator method, imputation of longitudinal data, regression based imputation, rough estimation of missing data values by using concept of mean, median and mode to specify a few of them [8].

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. S. A. Aljawarneh.

* Corresponding author.
E-mail address: usharani_y@vnrvjiet.in (U. Yelipe).

https://doi.org/10.1016/j.compeleceng.2017.11.030
0045-7906/

Please cite this article as: U. Yelipe et al., An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.030

Statistics defines the imputation process as substitution of incomplete (or missing) values. If a single attribute value or data element value is imputed, it is called as “unit imputation”. Alternately, when the missing data or incomplete data is handled at the component level, then it is called as “item imputation”. Imputation can also affect accuracy and efficiency of classifiers when not handled appropriately and correctly. Missing values are common in medical data [10].

Imputation process implicitly requires several data mining pre-processing techniques to be applied which is one of the default steps in data mining process. Another concern is, which attributes of medical data records should be considered as significantly promising attributes. Feature extraction and selection may be applied for such purpose. It is to be taken care that only those dimensions that do not affect final classification accuracies shall only be discarded [20]. Statistical approaches and techniques for data analysis require value for each data element. Missing values [10] restrict application of the statistical techniques and data analysis is thus not possible. It is here application of imputation can help to perform data analysis and classification of disease levels.

In [1] a decision tree based approach is discussed which debates on the choice of handling missing values. Clustering is a widely known learning technique which can also be adopted to handle incomplete medical data. One such approach for handling incomplete medical data is studied in [2]. Imputation is performed using “support vector regression” and “clustering” in [3]. Studies such as [4,5] address handling mixed attributes with missing values. A new framework [6] for performing imputation is discussed.

Auto regression based approach [7] is proposed to handle incomplete records. In [11], C5.0 is extended by adding two imputation approaches called IITMV (Intelligent imputation technique). Their approach involves obtaining a tree using C5.0 functions and applying hot-deck and EM-Imputation approaches. In [12] MMSD imputation technique is proposed to improve the classifier accuracy and accuracies obtained are compared to those achieved using mean, median, hot-deck, mean method based step digression and kNN based imputations. Density measure is used [13], to impute the incomplete pattern by finding best matching record and results obtained are compared to fuzzy c-means, k-means based imputation and fuzzy c-means with genetic algorithm based imputation.

Various imputation techniques such as mean, mode, kNN, Hot-deck, EM and C5.0 are compared in [14] and a review and discussion on which imputation is to be chosen is outlined. Experiments are conducted on synthetic datasets [14]. Our previous research [15–17] addresses missing value imputation applying clustering technique.

Medical records have many percentages of missing values, which directly influence their usefulness in terms of accuracy for classification algorithms. A class mean imputation based on the *k*-Nearest Neighbour Hot deck imputation approach to impute both nominal and continuous missing data value in datasets is demonstrated [21]. Gira [22] presented ratio type imputation approach for estimation of population data. Nishanth [23] proposed a k-mean and multilayer perceptron based imputation method for financial data that is used for predicting the severity of phishing attacks in financial firms.

Tang and Ishwaran [24] presents a machine learning based imputation method called random forest missing data algorithm. This approach enhanced the performance of all random forest procedure improved with increasing correlation of features. In [25] a column-wise guided data imputation (cGDI) method is demonstrated. The novelty resides in the selection of the most suitable model from a multitude of imputation method for each individual feature based on learning process on the known variable.

Most of the research related to text mining, text classification, intrusion and anomaly detection, web mining, temporal data mining, medical applications considered using only traditional distance measures. Song and Shepperd [21] presents a fuzzy similarity measure for text classification and clustering. Vangipuram et al. [26] presents similarity measure for temporal pattern mining which holds downward closure property. This research has been primarily inspired from [18,30]. In our previous research [15,16], we propose novel imputation approach to fill missing values.

Section 2 introduces the imputation process which is based on the concept of class-based clustering. The proposed imputation algorithm, imputation measure and similarity computation are discussed in Section 3 and resulting classification accuracies from experiments conducted are reported in Section 4 of this paper. Section 5 concludes this paper.

2. Imputation based on class-based clustering approach (CBC-IM)

Our approach is a class-based clustering approach. In this approach, we cluster records that do not have incomplete (or missing) values, i.e records in G^1. The total number of clusters is equal to the number of class labels (or may also be of user choice). We then obtain the distance (or similarity in case fuzzy measure is used) from each of these medical records to all cluster centres. In the first approach, we use Euclidean distance measure and in second approach we use fuzzy measure for similarity computation. When applying fuzzy measure, the standard deviation vector of respective clusters is considered. Our approach considers the dimensionality reduction of medical records to a dimension equal to number of class labels. Then we represent all these records as vectors whose dimensionality is equal to the total count of class labels. This is later followed by finding distance between these transformed records and missing attribute value records (transformed records in group, G^{IM}). Imputation is performed by considering each of these records in group, G^{IM} to which the record distance is minimum (or similarity is maximum).

Please cite this article as: U. Yelipe et al., An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.030

2.1. Research objective

We have two research objectives. The first includes proposing an improved approach (IM-CBC or CBC-IM) for imputation of medical data records through applying feature reduction technique using class-based clustering. The second includes achieving improved classification rates using existing classifiers such as SVM, kNN, C4.5 and validate accuracy of the proposed classifier (CBCC).

2.2. Our approach

The proposed imputation approach is discussed in this section. Initially we propose to group records into two diverse groups, one without missing attribute values (G^1) and another group (G^{IM}) having missing attribute values. The idea is to first consider all records in group G^1 (having no missing values) and obtain clusters equal to number of decision labels and use the knowledge of these clusters to achieve dimensionality reduction and perform imputation. From clusters obtained through considering records in first group (G^1), we obtain cluster mean and deviation for all clusters. Each record in (G^1) is transformed to its equivalent "p-dimensional" vector where each element value of the vector is distance (or similarity) value from cluster centre. Similarly, each record in group (G^{IM}) is also transformed to p-dimensional representation but by discarding missing attribute values. We then find distance (or similarity) between missing attribute value record and each transformed record present in group (G^1). Imputation is performed by considering the record present in group (G^1) that has minimum distance (or maximum similarity). 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, the mean of attribute value is filled and for nominal attribute the corresponding attribute value of similar record is chosen and substituted. After imputation is done, we have final set of medical records, with no incomplete records. This record set can be then used for finding classification accuracies. For prediction, we use the same procedure that is used for imputation. For prediction, class labels are determined instead of imputing missing values and accuracies are evaluated when performing classification. The importance of present approach is that we impute and classify medical records by reducing dimensionality.

3. Proposed approach

In this section, we give the similarity measure and the proposed algorithm for imputation and classification using class-based clustering. Analysis of a dataset having missing values and then filling the missing values through different approaches exist in the literature. However, the classification accuracies achieved using these approaches have been not so promising when analysed. It is this reason; which implicitly motivated us to study and address new methods for imputation. We name our method as Imputation Based on Class-based Clustering (IM-CBC or CBC-IM). Our approach, CBC-IM aims at reduction of attribute dimensionality. Imputation Based on Class-Based Clustering (CBC-IM) is achieved in 5 steps of imputation algorithm outlined in section 3.3.

3.1. Imputation measure (similarity measure)

Let R_i and R_j be two medical records and each medical record is defined over 'm' number of different attributes A_1, A_2, A_3, A_4, ... A_m. Further let, RA_{ik} and RA_{jk} denotes kth attribute element value of records R_i *and* R_j respectively. The membership value between R_i and R_j w.r.t lth attribute is computed by using equation (1)

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

The similarity between R_i and R_j is given by equation (2)

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

Where σ_l is the standard deviation of lth attribute column values.

A missing attribute value is denoted by symbol, $\emptyset$ *or* ?. For all such attributes corresponding to two records, the value of $\mathcal{M}^l(R_i, R_j)$ is treated as 1. The basic Gaussian membership function is extended to be used as the fuzzy similarity measure, for estimating similarity. Since, it is fuzzy based, all values lie between 0 and 1.

Please cite this article as: U. Yelipe et al., An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.030

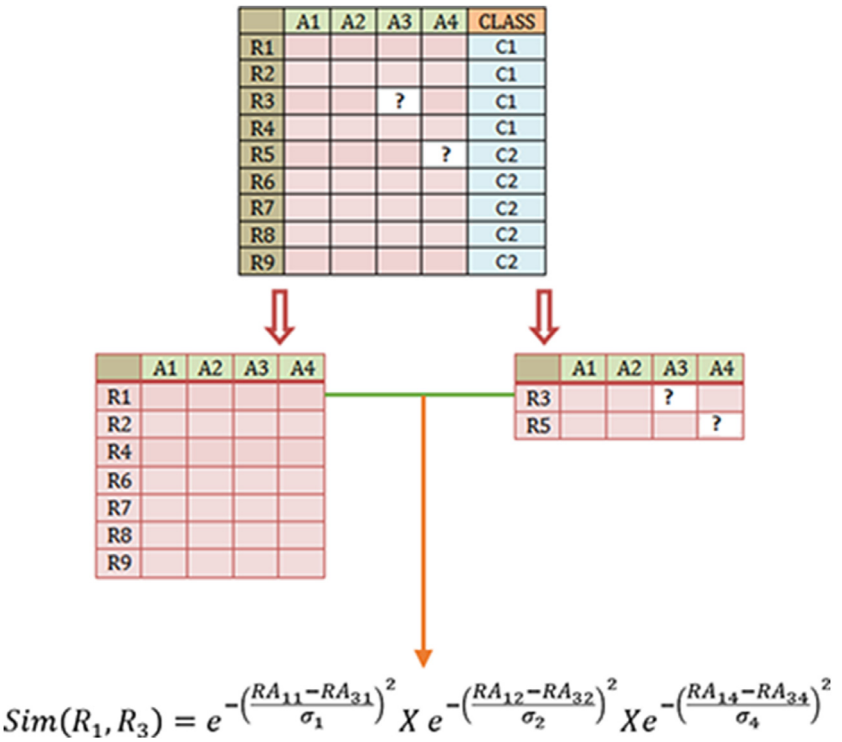


Fig. 1. Computation of membership value using proposed imputation measure.

3.2. Computation using similarity measure

We have expression for similarity computation given by Eq. (3) motivated from [18]

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

The similarity value between two records is computed as shown in Fig. 1

3.3. Imputation algorithm

Imputation algorithm is explained step by step as outlined below

Step-1: The first step in class-based clustering imputation involves classifying medical records in to two groups (G^1 and G^{IM}). G^1 has complete records and G^{IM} has incomplete records (those need to be imputed).

Step-2: All records in G^1 are clustered using k-means clustering algorithm. The mean and deviation of each cluster are recorded. The deviation and mean computed shall act as representatives for the clusters formed.

Step-3: The third step involves expressing records as fuzzy vectors through computation of fuzzy similarity (or distance vectors using Euclidean distance). For this, records from G^1 and G^{IM} are considered and their similarities or dissimilarities are computed w.r.t the mean of these clusters by applying either fuzzy measure or the Euclidean distance measure. When considering records from G^{IM} the mean vector element value corresponding to incomplete attributes value in missing record of G^{IM} is neglected for both cases when similarity or dissimilarity is computed.

Step-3 results in all records transformed to a new representation. The dimensionality of all these records shall be the count representing the total number of class labels. It is to be noted that the standard deviation is computed only for transformed dimensions of G^1.

Step 4: After records are transformed to new dimensions, then each record from the group, G^{IM}, is chosen and the similarity (or dissimilarity) value of these records is computed for each record in the group, G^1.

Step 5: In this step, missing attribute value is filled. For instance, A^y (missing attribute value) in R^{IM} (medical record) present in the category, G^{IM}, is filled by choosing, R^x in G^1, for which similarity of R^{IM} is maximum.

i) Nominal attribute – Attribute value imputation is carried by filling corresponding attribute value of record in G^1 for which R^{IM} records the maximum similarity.
ii) Numerical attribute – Mean of corresponding attribute value considering records in that class is filled.

Please cite this article as: U. Yelipe et al., An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.030

Table 1
C4.5 classifier accuracies with various imputation approaches.

Methods	AUS	CLE	ECOLI	GER	HEP	WINE	IRS	PIM	WISCON
CBC-IM-EUC	84.63	52.14	55.65	69.8	83.87	41.01	92.6	65.1	93.99
CBC-IM-FUZZY	84.49	51.48	58.03	70.1	84.51	39.88	95.33	65.1	93.41
IM-EUC-CBC FUZZY	82.02	51.48	52.97	70.5	81.29	39.88	90	65.1	93.27
MV-BPCA	81.59	54.12	42.55	70	80	39.88	85	65.1	88.12
MV-EM	75.79	51.48	41.96	70	80	38.76	85.3	70.18	93.56
MV-FKMeans	82.31	51.15	77.08	70	83.22	39.88	91.3	75.39	92.56
MV-Ignore	83.33	53.53	43.67	69.5	81.25	33.96	93.1	67.81	93.26
MV-Kmeans	83.47	51.48	58.92	70	83.22	39.88	83.3	65.1	94.42
MV-KNN	84.92	51.81	52.08	70	84.51	39.88	92	65.1	92.99
MV-Most Common	82.02	51.48	55.05	69.9	83.22	39.88	91.3	65.1	92.7
MV-SVDimpute	65.94	51.48	41.66	68.5	83.22	39.32	82	68.88	94.42
MV-SVMImpute	82.17	54.45	62.5	70	77.41	44.38	95.3	44.38	93.27
MV-WKNNimpute	84.78	51.81	51.19	70	84.51	39.88	91.3	65.1	92.99

Table 2
SVM classifier accuracies with various imputation approaches.

Methods	AUS	CLE	ECOLI	GER	HEP	WINE	IRS	PIM	WISCON
CBC-IM-EUC	83.62	52.14	60.41	71.8	80	85.39	90.66	63.05	96.56
CBC-IM-FUZZY	82.6087	51.87	61.01	73.3	83.22	82.58	90.66	62.1	96.13
IM-EUC-CBC FUZZY	79.13	52.14	58.92	74.2	80	70.78	84	63.93	95.99
MV-BPCA	76.08	50.16	53.86	67.9	79.35	59.55	81.3	75.91	59.55
MV-EM	77.82	52.8	52.97	71.4	79.35	65.16	82	69.01	95.13
MV-FKMeans	81.01	52.47	78.57	70.5	80	71.91	94.57	76.17	96.56
MV-Ignore	81.88	53.87	53.44	70.5	83.75	54.71	93.1	67.01	96.04
MV-Kmeans	82.46	54.45	62.2	72	82.58	78.08	78.1	63.54	96.13
MV-KNN	83.91	52.47	64.58	72.1	80	75.28	94.7	64.71	95.7
MV-Most Common	83.33	52.47	60.71	70.7	81.29	67.41	92.7	65.75	95.99
MV-SVDimpute	76.08	53.46	54.46	70.5	83.22	61.79	83.8	65.62	94.99
MV-SVMImpute	88.98	53.13	73.21	81.8	92.25	89.88	94.7	89.88	95.99
MV-WKNNimpute	84.2	52.47	65.47	71.3	80.64	72.47	94	64.71	95.7

Table 3
kNN classifier accuracies with various imputation approaches.

Methods	AUS	CLE	ECOLI	GER	HEP	WINE	IRS	PIM	WISCON
CBC-IM-EUC	79.56	54.12	54.16	68.1	79.35	79.77	84.7	62.1	95.7
CBC-IM-FUZZY	79.85	54.12	55.35	69.1	78.06	76.96	86	62.1	93.41
IM-EUC-CBC -FUZZY	78.69	54.12	51.78	68.5	80.0	57.3	81.33	64.19	95.27
MV-BPCA	74.05	49.17	47.67	67.6	74.83	56.74	78.0	73.82	56.54
MV-EM	73.62	55.77	52.97	69.6	74.83	54.49	90.0	65.62	95.27
MV-FKMeans	76.95	56.1	73.21	69.5	79.35	63.48	87.33	65.88	95.99
MV-Ignore	79.71	55.89	52.29	73.0	80.0	60.37	85.0	67.54	95.6
MV-Kmeans	80.86	56.67	57.44	68.3	79.35	68.53	68.53	64.45	96.13
MV-KNN	81.88	55.77	55.35	69.5	76.77	69.1	90.0	64.84	95.99
MV-Most Common	80.72	55.77	51.48	70.0	78.7	55.61	84.0	60.8	95.85
MV-SVDimpute	70.72	55.77	46.72	68.0	76.77	55.05	82.0	62.1	95.13
MV-SVMImpute	83.91	55.77	64.28	71.1	86.45	83.7	91.33	83.7	96.13
MV-WKNNimpute	82.17	55.77	55.05	69.4	78.06	63.48	88.66	64.58	95.99

Step-6: For disease prediction and classification, the procedure discussed in steps 1 to 5 is repeated by considering training and testing records having no missing values. We name the classifier as CBCC.

Please cite this article as: U. Yelipe et al., An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.030

Table 4
Classifier accuracies on benchmark datasets.

Dataset (Clusters)	SVM	kNN	C4.5	CBCC-IM-EUC
Clev (7)	52.1452	54.1254	52.1452	56.4301
Ecoli(13)	60.4167	54.1667	55.6548	83.7896
Iris(3)	90.6667	84.6667	92.6667	94.666
Pima(3)	63.0508	62.1094	65.1042	66.6765
Wine(11)	85.3933	79.7753	41.0112	97.1568
Wisconsin(13)	96.5665	95.7082	93.9914	96.708

Table 5
Classifier accuracies on benchmark datasets.

Dataset (Clusters)	SVM	kNN	C4.5	CBCC-IM-FUZZY
Clev (13)	51.87	54.1254	51.4851	58.7419
Ecoli(11)	61.0119	55.3571	58.0357	79.8
Wine(7)	82.5843	76.9663	39.8876	92.6797
Wisconsin(5)	96.1373	93.41	93.4192	96.2795

Table 6
Classifier accuracies – with EUCLIDEAN and Fuzzy measures.

Dataset (Cluster)	CBCC-IM-EUC	Dataset (clusters)	CBCC-IM-FUZZY	Dataset (clusters)	IM-EUC-CBCC- FUZZY
Aus (3)	79.2753	Aus(11)	75.3623	Aus(13)	71.3043
Clev (7)	56.4301	Clev(13)	58.7419	Clev(9)	53.8064
Ecoli(13)	83.7896	Ecoli(11)	79.8	Ecoli(5)	76.1942
German(13)	68	German(15)	64.2	German(7)	63.7
Hep(7)	77.4166	Hep(3)	72.8333	Hep(3)	72.2499
Iris(3)	94.666	Iris(5)	82	Iris(3)	92.1241
Pima(3)	66.6765	Pima(11)	64.8552	Pima(1)	86.6666
Wine(11)	97.1568	Wine(7)	92.6797	Wine(11)	64.9766
Wisconsin(13)	96.708	Wisconsin(5)	96.2795	Wisconsin(3)	96.4223

Table 7
Benchmark datasets description.

S.NO	Data set name	Abbreviation	#Attributes (R/I/N)	#Examples	#Classes	%MVs (Examples)
1	Iris + MV	IRI	4 (4/0/0)	150	3	32.67%
2	Pima + MV	PIM	8 (8/0/0)	768	2	50.65%
3	Wine + MV	WIN	13 (13/0/0)	178	3	70.22%
4	Australian + MV	AUS	14 (3/5/6)	690	2	70.58%
5	Ecoli + MV	ECO	7 (7/0/0)	336	8	48.21%
6	Cleveland	CLE	13 (13/0/0)	303	5	1.98%
7	German + MV	GER	20 (0/7/13)	1000	2	80.00%
8	Hepatitis	HEP	19 (2/17/0)	155	2	48.39%
9	Wisconsin	WIS	9(0/9/0)	699	2	2.29%

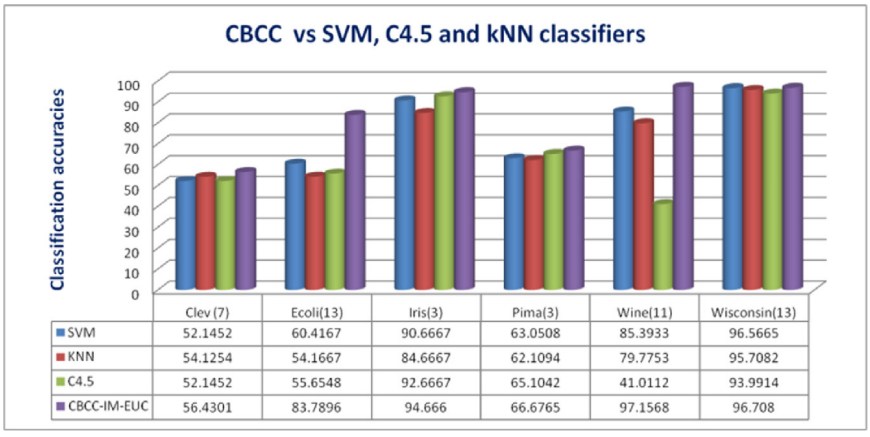
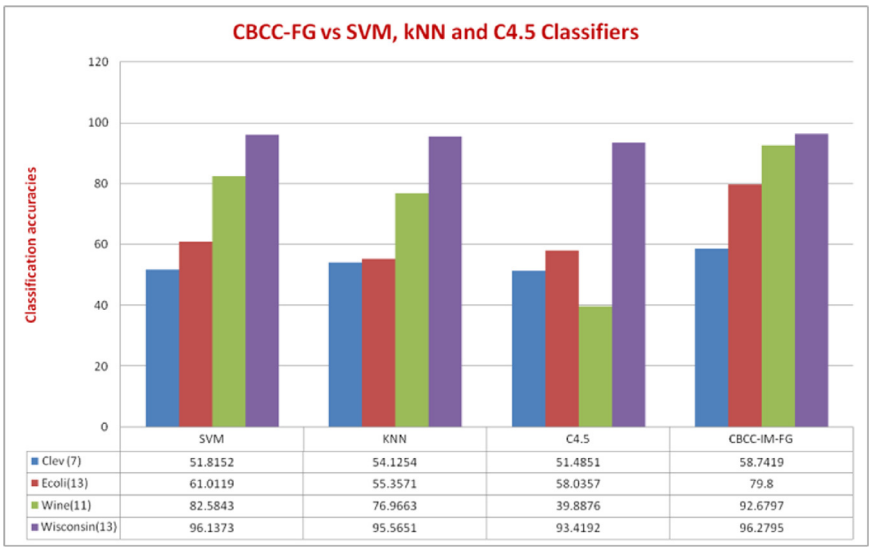


Fig. 2. CBCC-Euc vs SVM, kNN, C4.5 classifier accuracies.

Please cite this article as: U. Yelipe et al., An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.030



	SVM	KNN	C4.5	CBCC-IM-FG
Clev (7)	51.8152	54.1254	51.4851	58.7419
Ecoli(13)	61.0119	55.3571	58.0357	79.8
Wine(11)	82.5843	76.9663	39.8876	92.6797
Wisconsin(13)	96.1373	95.5651	93.4192	96.2795

Fig. 3. CBCC-FG vs SVM, kNN, C4.5 classifier accuracies.

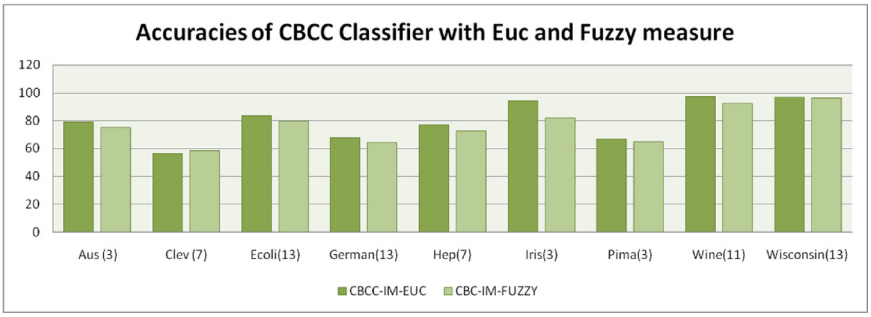
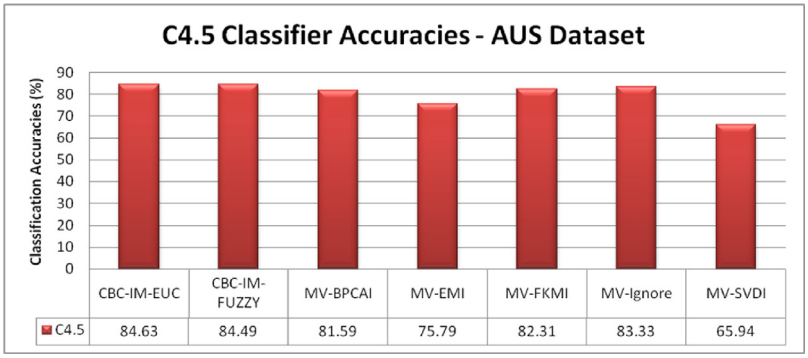


Fig. 4. CBCC-FG vs CBCC-EUC classifier accuracies.



	CBC-IM-EUC	CBC-IM-FUZZY	MV-BPCAI	MV-EMI	MV-FKMI	MV-Ignore	MV-SVDI
C4.5	84.63	84.49	81.59	75.79	82.31	83.33	65.94

Fig. 5. C4.5 classifier accuracies – various imputation approaches.

Please cite this article as: U. Yelipe et al., An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.030

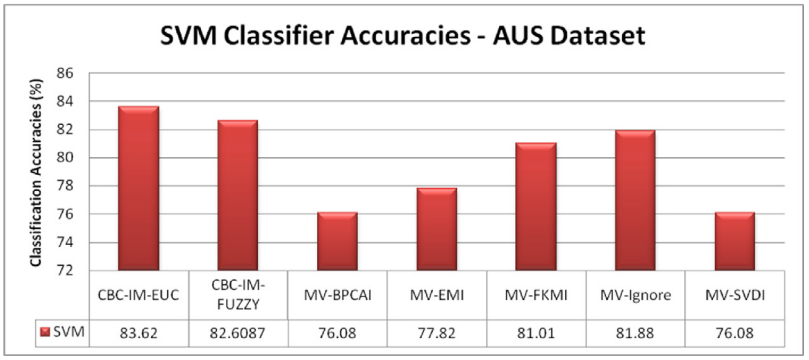


Fig. 6. SVM classifier accuracies– various imputation approaches.

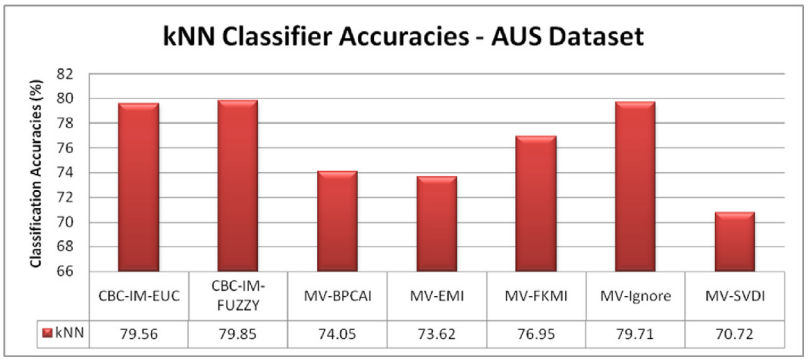


Fig. 7. kNN classifier accuracies– various imputation approaches.

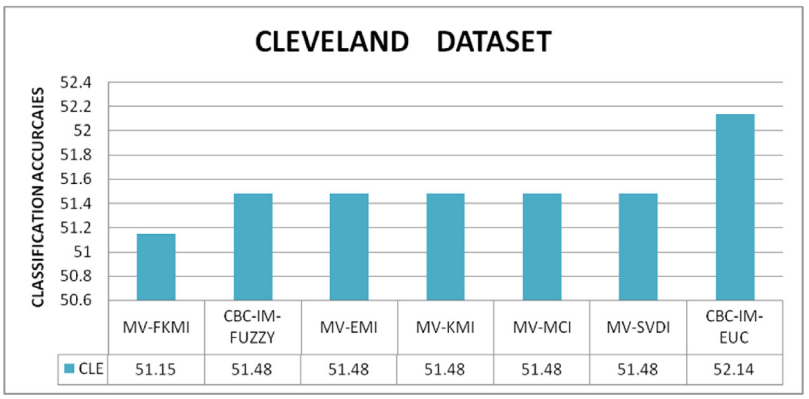


Fig. 8. C4.5 accuracies on CLEVELAND Dataset for various imputation approaches.

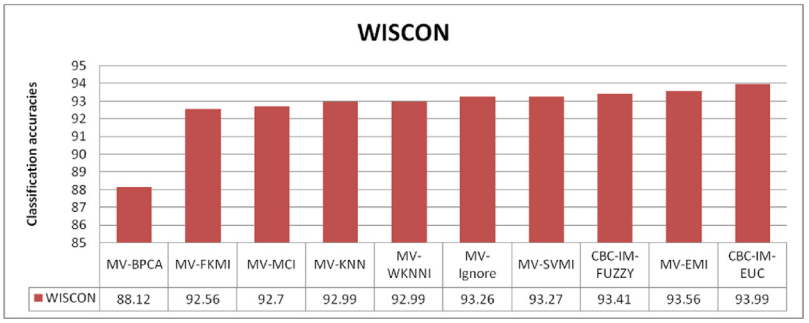


Fig. 9. C4.5 accuracies on WISCON Dataset for various imputation approaches.

Please cite this article as: U. Yelipe et al., An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.030

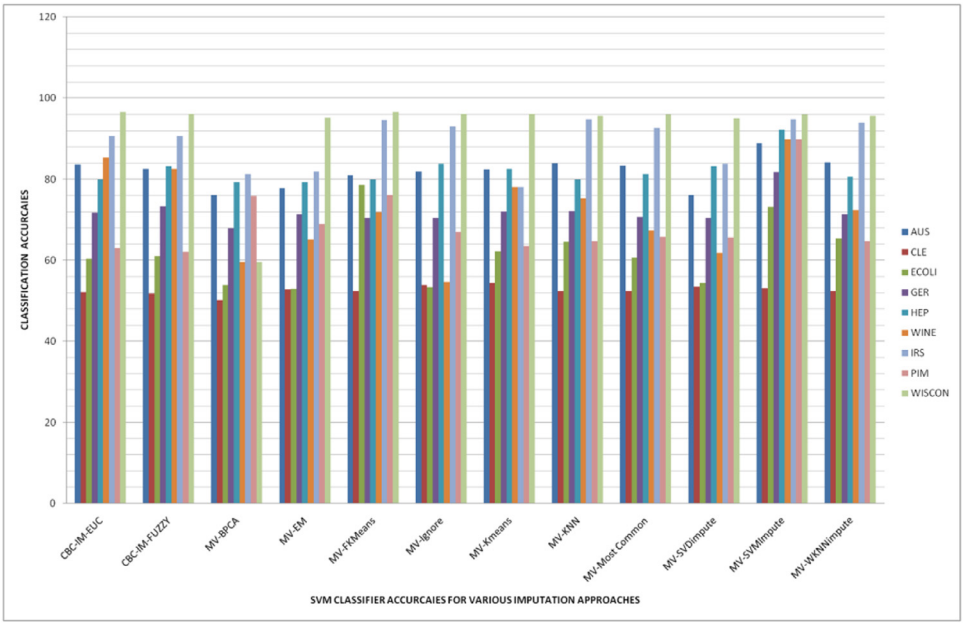


Fig. 10. SVM classifier accuracies for various imputation approaches on benchmark datasets.

3.3.1. Algorithm: imputation of medical data values

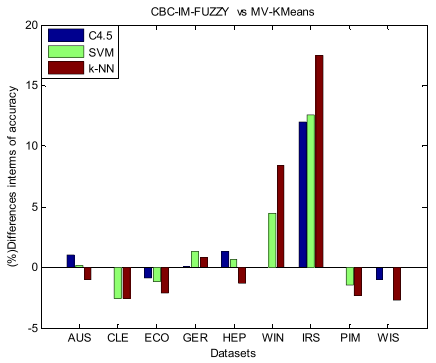
Input : Medical Record Dataset (U) consisting of missing attribute values.
Output : Imputed medical records
Variable notations:
$n \rightarrow$ total numer of medical records
i, j $\rightarrow$ index of medical records
k $\rightarrow$ index of medical record attribute
$P_k \rightarrow k$th attribute variable
$M_i \rightarrow i$th Medical record
$M_i(P_k) \rightarrow k$th attribute value in the ith medical record
$\emptyset \rightarrow$ Missing attribute value
$|L| \rightarrow$ Number of unique class label
$C[r] \rightarrow r$th cluster
$U \rightarrow$ Medical record dataset with class lables
$\mu_r \rightarrow$ Mean of rth cluster with k-dimensionality

Begin of algorithm

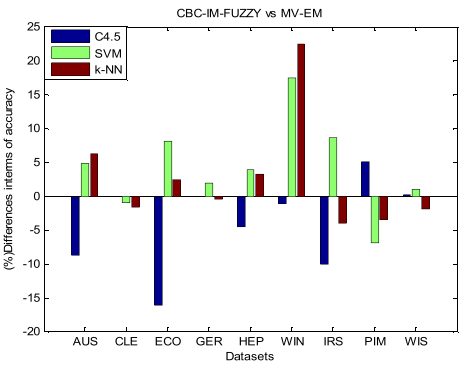
1. Split the given medical record dataset into two sets, one is without missing attribute value which is represented by Group1= $\{M_i | \forall$ i,k and $M_i(P_k) \neq \emptyset\}$ and other set is with missing attribute values that is denoted as Group2= $\{M_i | \forall$ i,k and $M_i(P_k) = \emptyset\}$.
2. Generate number of clusters equal to number of class label by applying k-means clustering algorithm and considering records in Group1 i.e C[U, r] = KMeans (Group1, |L|)
3. Compute mean of r-clusters where mean of the rth cluster is denoted using $\mu_r = \langle \mu_r^1, \mu_r^2, \mu_r^3, \ldots\ldots \mu_r^k \rangle$ and each μ_r^k is obtained applying, $\mu_r^k = \frac{\sum_{j=1}^{j=y} M_j(P_k)}{|j|}$, $\forall$ j, r with record, $M_j \in r$th cluster
4. Find the similarity value of each medical record to all clusters applying the proposed fuzzy measure. i.e $Sim(M_i , C[r]) = \prod_{k=1}^{k=m} \exp^{-\left(\frac{(M_i(P_k)-\mu_r^k)}{\sigma}\right)^2}$ where μ_r^k represents mean value of rth cluster considering kth attribute.
5. Represent all medical records from both groups in terms of similarity values to cluster centre. i.e dimensionality of each medical record is now equal to number of clusters.
6. Find the similarity value between each record in group-2 to every record in group-1 and the record the medical record in group-1 that has highest similarity value to medical record in group-2.
7. Apply imputation process steps outlined from Step-1 to Step-5 and impute the suitable missing attribute value in medical record dataset accordingly.
8. Stop the imputation process when all missing attribute values in dataset are imputed successfully.
9. Output the final imputed record dataset.

End of algorithm

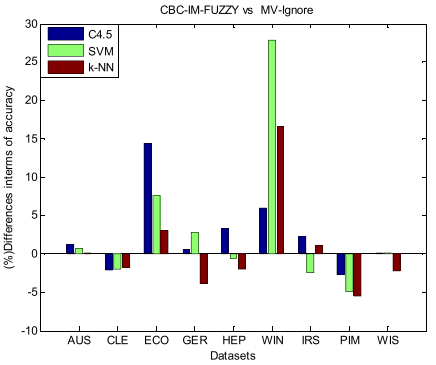
Please cite this article as: U. Yelipe et al., An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.030



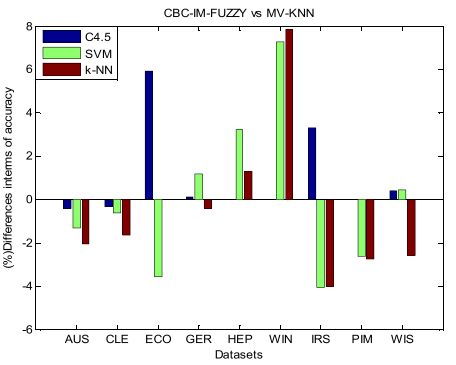
(a) CBC-IM-FUZZY vs MV-Kmeans



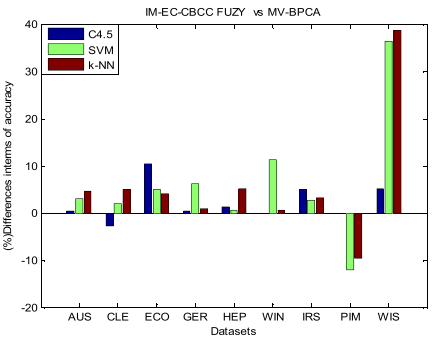
(b) CBC-IM-FUZZY vs MV-EM



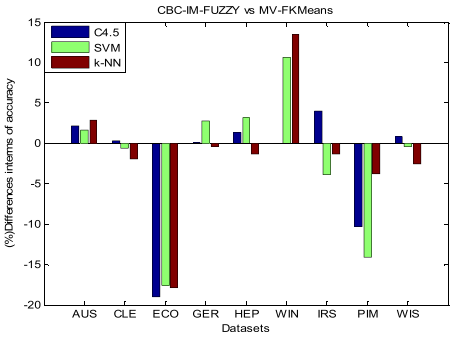
(c) CBC-IM-FUZZY vs MV-Ignore



(d) CBC-IM-FUZZY vs MV-KNN



(e) CBC-IM-FUZZY vs MV-BPCA



(f)CBC-IM-FUZZY vs MV-FKMeans

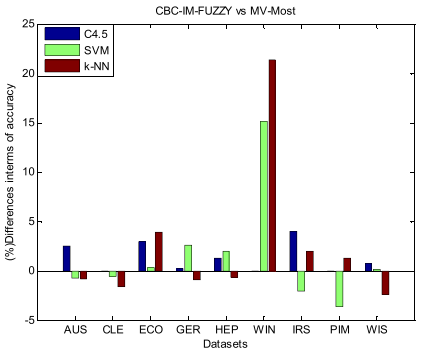
Fig. 11. Comparison of CBC-IM-FUZZY to all other approaches.

4. Results and discussions

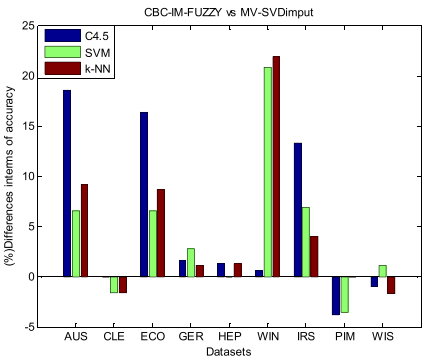
In this section, we are going to discuss the classifier accuracy results achieved on various benchmark datasets [9,19,20] by applying proposed imputation method. Results are discussed in two perspectives i.e using existing classifiers and proposed classification technique.

I. Our imputation approach is applied to benchmark datasets and classification results achieved are compared to ten imputation strategies that include MV-BPCA, MV-EM, MV-FKMEANS, MV-Ignore, MV-Kmeans, MV-kNN, MV-MostCommon, MV-SVDIMPUTE, MV-SVMIMPUTE, MV-WKNN-IMPUTE. The classifiers chosen are kNN, SVM and C4.5. In Table 1, Table 2 and Table 3, CBC-IM-EUC denotes imputation performed using class-based clustering with euclidean distance. Similarly, CBC-IM-Fuzzy denotes imputation performed using class-based clustering with fuzzy measure and IM-EUC-CBC-FUZZY indicates imputation performed using Euclidean distance but class-based clustering performed using fuzzy measure. It is on this imputed benchmark datasets, classification accuracies are recorded applying SVM, kNN and C4.5 classifiers.

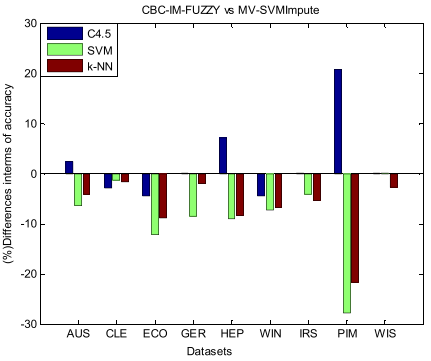
Please cite this article as: U. Yelipe et al., An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.030



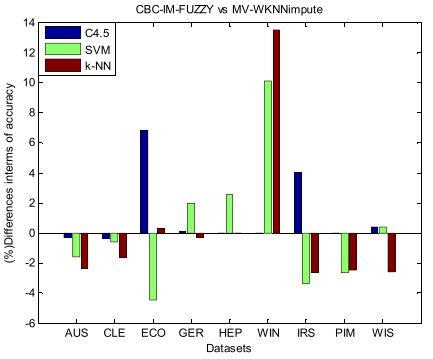
(g)CBC-IM-FUZZY vs MV-Most



(h)CBC-IM-FUZZY vs MV-SVDimpute



(i) CBC-IM-FUZZY vs MV-SVMI



(j) CBC-IM-FUZZY vs MV-WKNNI

Fig. 11. Continued

II. Our imputation method is applied to benchmark datasets and the classification results obtained using CBCC classifier using Euclidean and Fuzzy measure are compared to kNN, SVM and Naïve Bayes, C4.5 classifiers.

4.1. Results achieved using existing classifiers

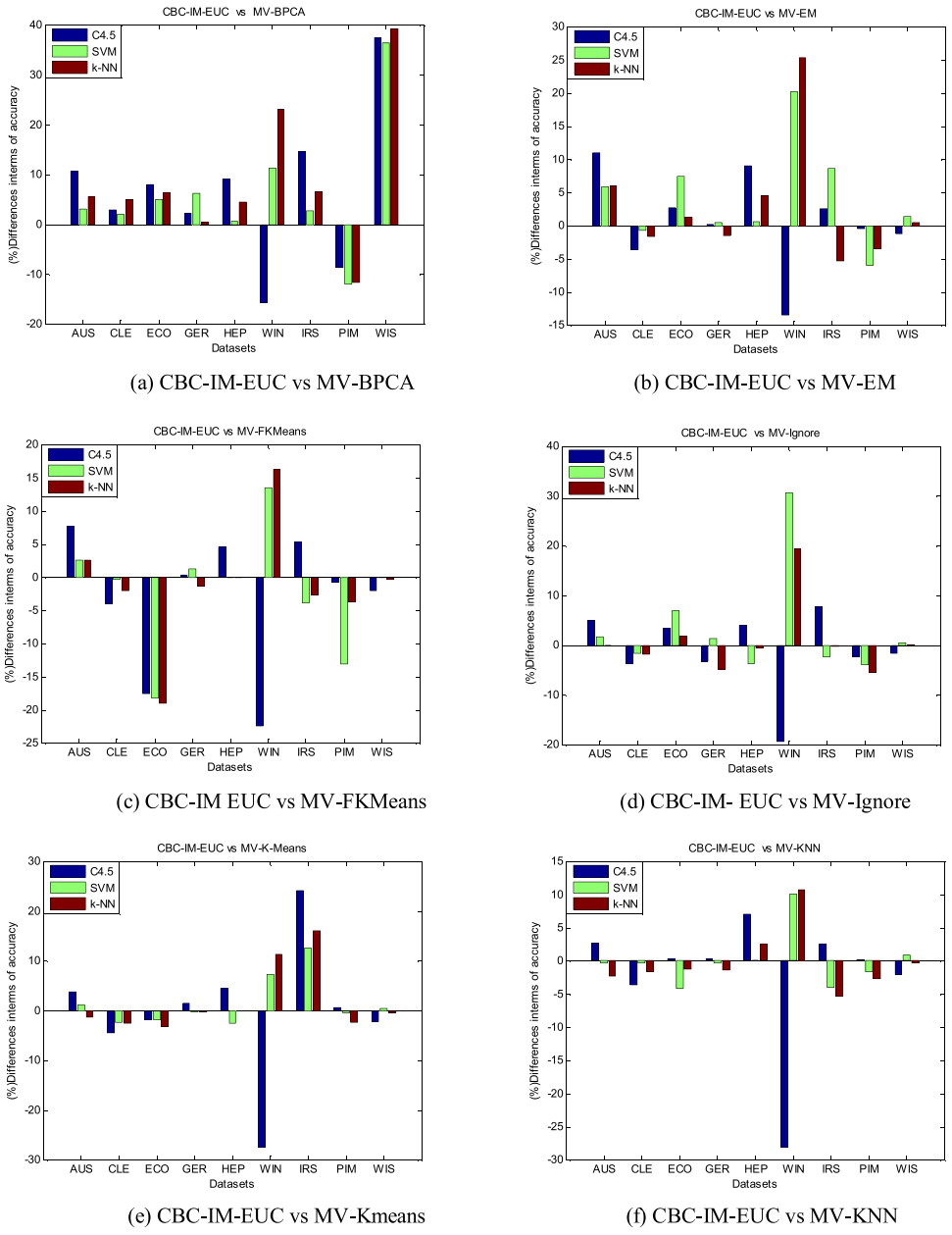
Table 1 shows results achieved on benchmark datasets Australian, Cleveland, Ecoli, German, Hepatitis, Wine, Iris, Pima and Wiscon datasets [19,20] using proposed imputation approach with Euclidean and fuzzy measure. The results achieved using C4.5 decision tree classifier with ten imputation techniques and nine benchmark datasets are compared to accuracies obtained using CBC-IM-EUC imputation approach. Some of these findings are discussed below

i) Percentage accuracy of C4.5 classifier recorded on **WISCOSIN** benchmark dataset using proposed approach is 93.99 (CBC-IM-EUC) which is better to other imputation approaches MV-BPCA (88.12%), MV-EM (93.56%), MV-FKMeans (92.56%), MV-Ignore (93.26%), MV-KNN (92.9942%), MV-Most Common (92.7%), MV-WKNN impute (92.99%), MV-SVM Impute (93.27%).
ii) Percentage accuracy of C4.5 classifier recorded on **AUS** dataset using proposed approach (CBC-IM-EUC) is 84.63 which is better to other imputation approaches MV-BPCA (81.59%), MV-EM (75.79%), MV-FKMeans (82.31%), MV-Ignore (83.33%), MV-Kmeans (83.47%), MV-Most Common (82.02%), MV-SVD impute (65.94%), MV-SVM Impute (82.17%).
iii) Percentage accuracy of C4.5 classifier recorded on **CLE** dataset using proposed approach (CBC-IM-EUC) is 52.14 which is better to other imputation approaches MV-EM (51.48%), MV-FKMeans (51.15%), MV-Kmeans (51.48%), MV-Most Common (51.48%), MV-SVD impute (51.48%), MV-WKNN Impute (51.81%).

Table 2 shows results achieved on benchmark datasets Australian, Cleveland, Ecoli, German, Hepatitis, Wine, Iris, Pima and Wiscon datasets using proposed imputation approach with Euclidean distance measure and fuzzy measure. Percentage accuracy recorded on **WISCOSIN** benchmark dataset using proposed approach is 96.56 (CBC-IM-EUC) and 96.13 (CBC-IM-FUZZY) which is better to other accuracies obtained using SVM classifier on imputation approaches MV-BPCA (59.55%), MV-EM (95.13%), MV-Ignore (96.04%), MV-Kmeans (96.13%), MV-Most Common (95.99%), MV-SVD impute (94.99%), MV-SVM Impute (95.99%). Accuracies of SVM on all other benchmark datasets is depicted in Table 2

Table 3 shows results achieved on benchmark datasets Australian, Cleveland, Ecoli, German, Hepatitis, Wine, Iris, Pima and Wiscon datasets using proposed imputation approach with Euclidean distance measure. The results achieved using kNN classifier for CLEVELAND dataset is 54.12% using proposed approach with Euclidean and fuzzy measures. This is better compared to MV-BPCA. It is noted that the kNN Classifier is showing slightly lower classification rates when compared to other

Please cite this article as: U. Yelipe et al., An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.030



(a) CBC-IM-EUC vs MV-BPCA

(b) CBC-IM-EUC vs MV-EM

(c) CBC-IM EUC vs MV-FKMeans

(d) CBC-IM- EUC vs MV-Ignore

(e) CBC-IM-EUC vs MV-Kmeans

(f) CBC-IM-EUC vs MV-KNN

Fig. 12. Comparison of CBC-IM-EUC to all other approaches.

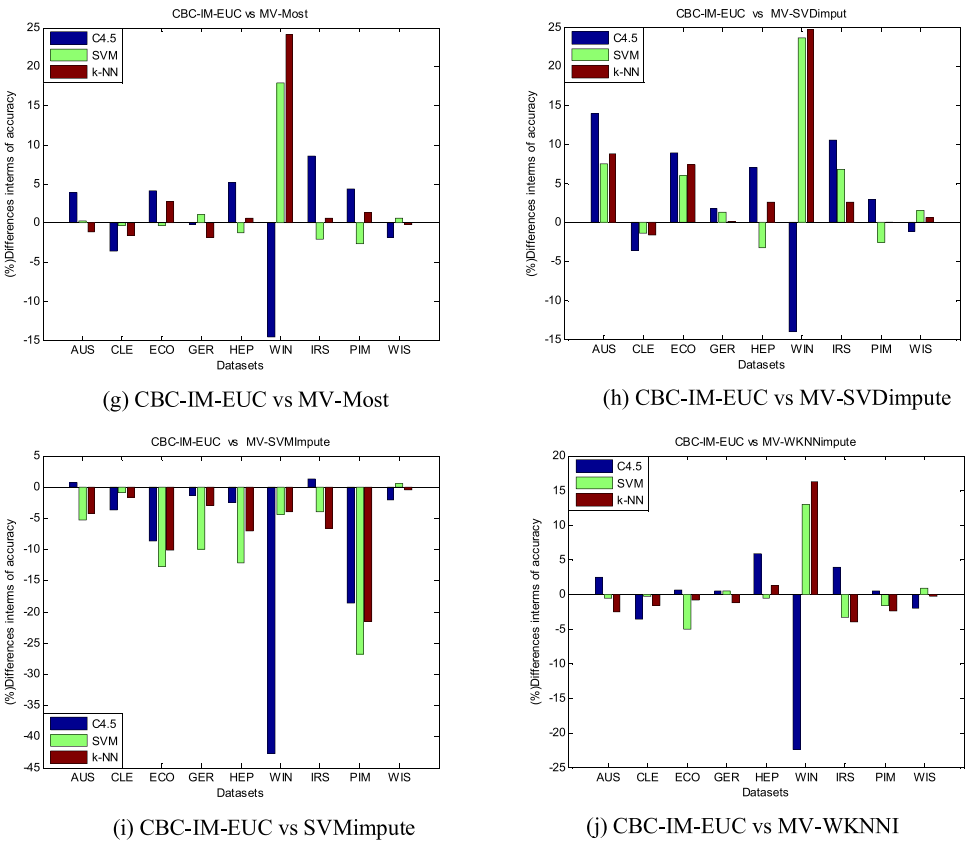
imputation approaches with highest recorded accuracy being 56.67% using MV-Kmeans imputation. This disadvantage is overcome with the proposed classifier, CBCC discussed in later sub-section.

4.2. Comparison of results achieved with proposed classifier using Euclidean and Fuzzy measures to SVM, kNN and C4.5

Table 4 shows classifier accuracies achieved using proposed imputation approach with CBCC classifier (our approach) for Cleveland, Ecoli, Iris, Pima, Wine and Wisconsin benchmark dataset.

i) **Wine** – The accuracy achieved using CBCC-IM-EUC is 97.15% where as it is observed that the accuracy of SVM, kNN and C4.5 classifiers are 85.39%, 79.77% and 41.01% respectively.
ii) **Wisconsin** – The accuracy achieved using CBCC-IM-EUC is 96.70% where as it is observed that the accuracy of SVM, kNN and C4.5 classifiers are 96.56%, 95.71% and 93.99% respectively.

Please cite this article as: U. Yelipe et al., An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.030



(g) CBC-IM-EUC vs MV-Most

(h) CBC-IM-EUC vs MV-SVDimpute

(i) CBC-IM-EUC vs SVMimpute

(j) CBC-IM-EUC vs MV-WKNNI

Fig. 12. Continued

iii) **Cleveland** – The accuracy achieved using CBCC-IM-EUC is 56.43% where as it is observed that the accuracy of SVM, kNN and C4.5 classifiers are 52.14%, 54.12% and 52.14% respectively.
iv) **Ecoli** – The accuracy achieved using CBCC-IM-EUC is 83.78% where as it is observed that the accuracy of SVM, kNN and C4.5 classifiers are 60.42%, 54.16% and 55.65% respectively.
v) **Iris** – The accuracy achieved using CBCC-IM-EUC is 94.66% where as it is observed that the accuracy of SVM, kNN and C4.5 classifiers are 90.66%, 84.66% and 92.66% respectively.

Table 5 shows classifier accuracies achieved using proposed imputation approach with CBCC classifier (our approach) with fuzzy measure for Cleveland, Ecoli, Iris, Pima, Wine and Wisconsin benchmark dataset.

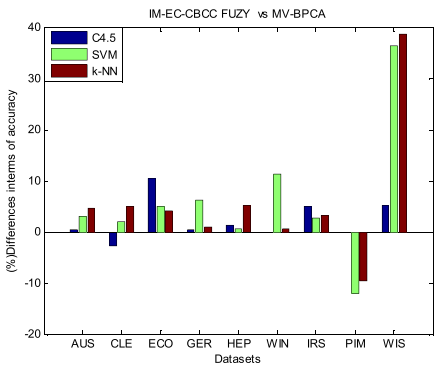
i) Wine – The accuracy achieved using CBCC-IM-FUZZY is 92.67% where as it is observed that the accuracy of SVM, kNN and C4.5 classifiers are 82.58%, 76.96% and 39.88% respectively.
ii) Wisconsin – The accuracy achieved using CBCC-IM-FUZZY is 96.27% where as it is observed that the accuracy of SVM, kNN and C4.5 classifiers are 96.14%, 95.56% and 93.41% respectively.
iii) Cleveland – The accuracy achieved using CBCC-IM-FUZZY is 58.74% where as it is observed that the accuracy of SVM, kNN and C4.5 classifiers are 51.81%, 54.12% and 51.48% respectively.
iv) Ecoli – The accuracy achieved using CBCC-IM-FUZZY is 79.8% where as it is observed that the accuracy of SVM, kNN and C4.5 classifiers are 61.01%, 55.36% and 58.04% respectively.

Using proposed classifier, CBCC-IM-FUZZY the accuracy for Cleveland dataset is better compared to other classifiers and imputation strategies. This accuracy is the highest recorded accuracy on Cleveland dataset when compared to all other approaches.

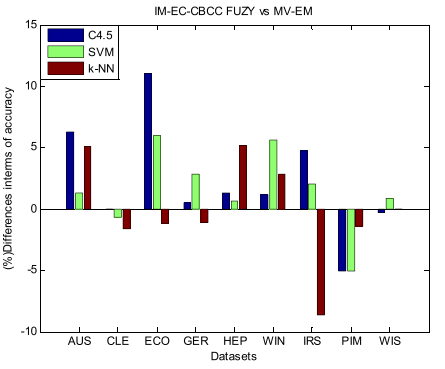
4.3. Comparison of results obtained for proposed classifier using Euclidean (CBCC-IM-EUC) and Fuzzy (CBCC-IM-FUZZY) measures

Table 6 gives the comparison of classification accuracies obtained using proposed imputation with CBCC Classifier using Euclidean and fuzzy measures for various benchmark datasets. The clusters formed for AUS dataset is 3, Cleveland and Hepatitis is 7, German, Wisconsin, Ecoli is 13. In Table 6, CBCC-IM-EUC, CBCC-IM-FUZZY, IM-EUC-CBCC-FUZZY denotes accuracies of class-based clustering classification approach with imputation and classification carried using Euclidean distance

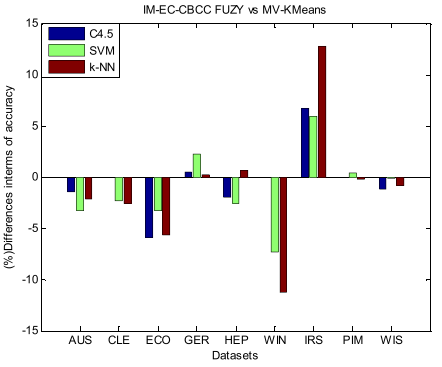
Please cite this article as: U. Yelipe et al., An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.030



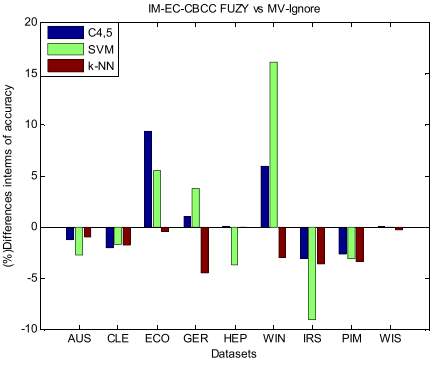
(a) IM-EUC-CBC-FUZZY vs MV-BPCA



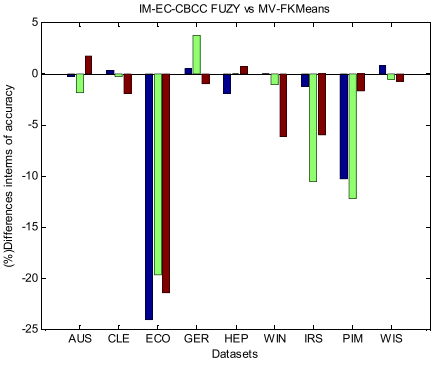
(b) IM-EUC-CBC-FUZZY vs MV-EM



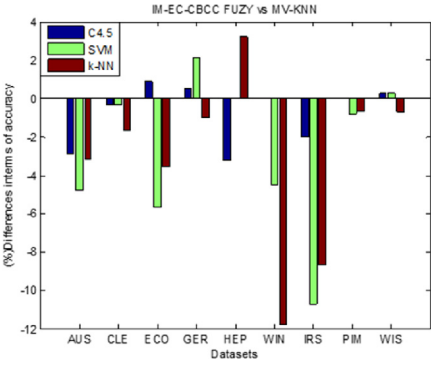
(c) IM-EUC-CBC-FUZZY vs MV-KMeans



(d) IM-EUC-CBC-FUZZY vs MV-Ignore



(e) IM-EUC-CBC-FUZZY vs MV-FKMeans



(f) IM-EUC-CBC-FUZZY vs MV-KNN

Fig. 13. Comparison of IM-EUC-CBC-FUZZY to other approaches.

function, class-based clustering classification approach with imputation and classification carried using fuzzy measure, class-based clustering classification approach with imputation carried using Euclidean distance function and classification using proposed fuzzy membership function.

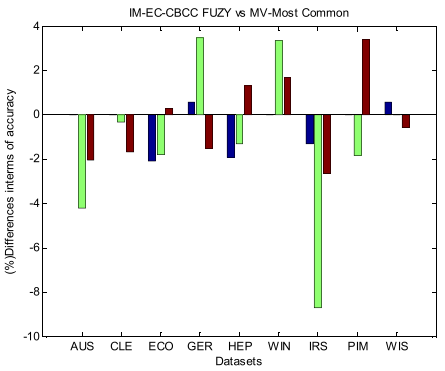
Graphs depicted in Figs. 2 and 3 show comparison of classification accuracies of CBCC Classifier using CBC-IM with Euclidean and fuzzy gaussian (FG) measures to accuracies achieved using well known classifiers SVM, kNN and c4.5 for benchmark datasets depicted in the graph.

Fig. 4 depicts the classification accuracies achieved using CBCC classifier with Euclidean and fuzzy measure for well-known benchmark datasets.

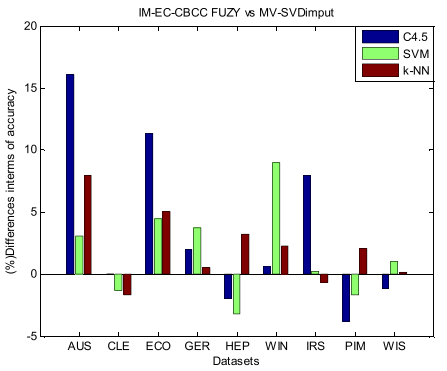
Graphs in Figs. 5–7 depict the classification accuracies achieved using C4.5, SVM and kNN classifier with imputation using CBC-IM adopting Euclidean and fuzzy measure over AUS dataset.

Graphs in Figs. 8 and 9, depict the classification accuracy achieved using C4.5 classifier with imputation using CBC-IM adopting Euclidean and fuzzy measure over CLEVELAND and Wiscon datasets respectively. Graph in Fig. 10, depicts the classification accuracy achieved using SVM classifier for various datasets comparing various imputation techniques.

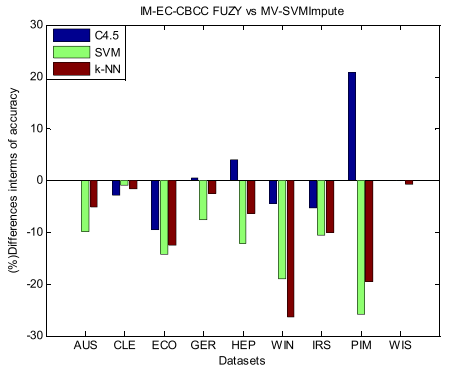
Please cite this article as: U. Yelipe et al., An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.030



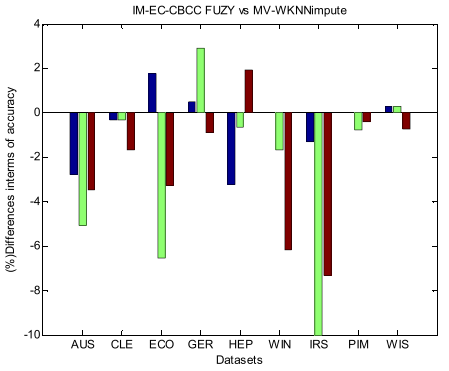
(g) IM-EUC-CBC-FUZZY vs MV-Most Common



(h) IM-EUC-CBC-FUZZY vs MV-SVDI



(i) IM-EUC-CBC-FUZZY vs MV-SVMI



(j) IM-EUC-CBC-Fuzzy vs MV-WKNNI

Fig. 13. Continued

4.4. Non-parametric statistical test

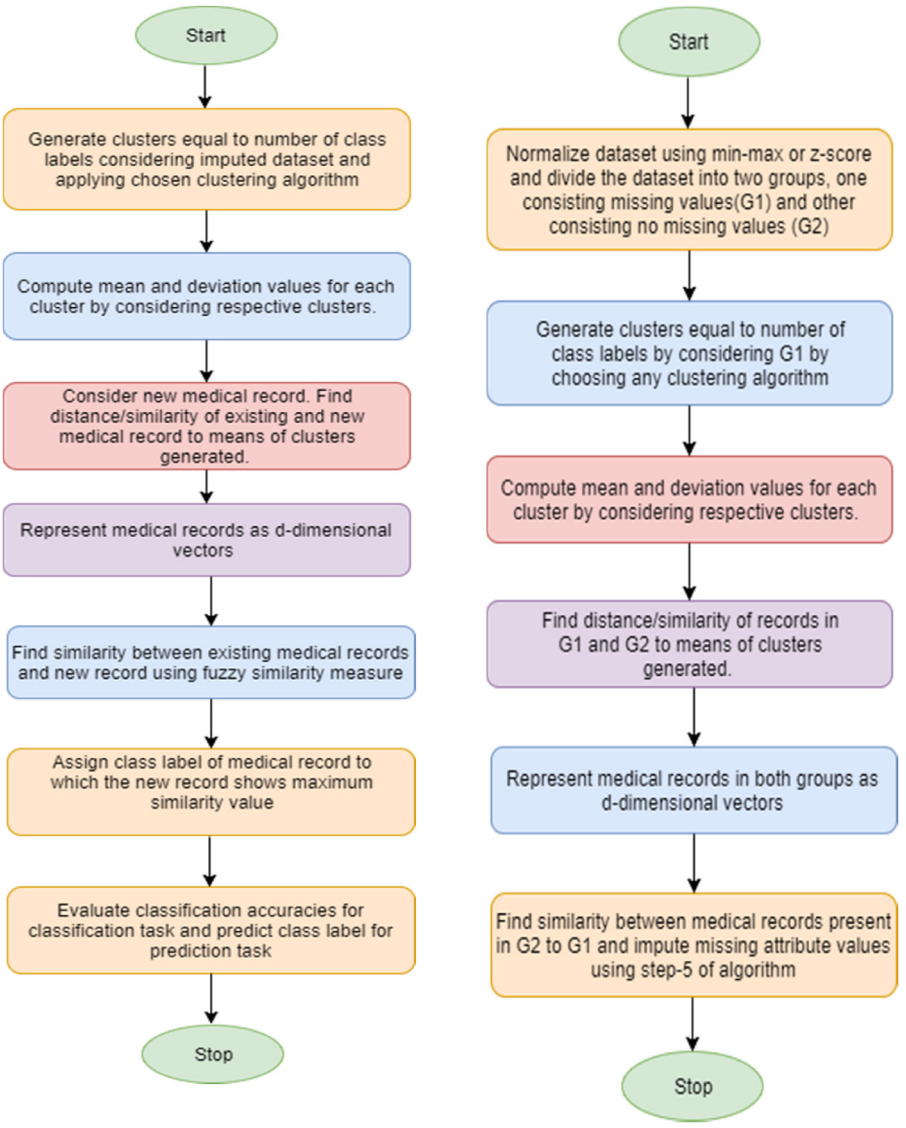
In this study, we use the non-parametric test that is also known as Wilcoxon signed-rank test which is equivalent to the paired *t*-test. This pairwise test aims to detect significant differences between the behaviour of two algorithms [27,28]. It does not assume normality of the sample mean of the accuracy. In addition, it is most robust with respect to outliers [29].

In Figs. 11–13 all bar graphs above the baseline denote the percentage of accuracy in the form of improvement, which are obtained based on the proposed methods known as CBC-IM-EUC, CBC-IM-Fuzzy and IM-EC-CBCC Fuzzy Imputation algorithms. The positive bar chart above the baseline represents the improvement in terms of accuracy. The bars above base line represents positive test classifiers and below the base line indicates negative test classifier accuracy of CBC-IM-EUC, CBC-IM-Fuzzy and IM-EUC-CBC-Fuzzy imputation approaches with other popular imputation algorithms.

In this study, C4.5 classifier using CBC-IM-EUC approach has scored around 37.45% percentage improvement in terms of accuracy when compared with MV-BPCA Imputation approach for Wiscon dataset. Application of C4.5 classifier with CBC-IM-Fuzzy imputation approach has scored approximately 20.72% percentage improvement in terms of classifier accuracy when compared with MV-SVM imputation approach for Pima dataset. In addition, C4.5 classifier with IM-EUC-CBC-FUZZY imputation has scored around 20.72% percentage improvement in terms of classifier accuracy when compared with MV-SVM imputation approach for Pima dataset.

SVM classifier using the CBC-IM-EUC approach has scored around 37.01% and 30.68% percentage improvement in terms of accuracy when compared with MV-BPCA and MV-Ignore Imputation approaches for Wiscon and Wine datasets. Application of SVM classifier with CBC-IM-Fuzzy imputation approach has scored approximately 36.58% and 27.87% percentage improvement in terms of classifier accuracy when compared with MV-BPCA and MV-Ignore for Wiscon and Wine dataset. In addition, SVM classifier with IM-EUC-CBC FUZZY Imputation approach has scored around 36.44% percentage improvement in terms of classifier accuracy when compared with MV-BPCA Imputation algorithm for Wiscon dataset. k-NN classifier using the CBC-IM-EUC algorithm has scored around 39.16%, 25.28% and 24.72% percentage improvement in terms of accuracy when compared with MV-BPCA, MV-EM and MV-SVD imputation approaches for Wiscon and Wine datasets. Application of k-NN classifier with CBC-IM-Fuzzy Imputation algorithm has scored approximately 36.87% and 22.47% percentage improvement in terms of classifier accuracy when compared with MV-BPCA and MV-EM Imputation approach for Wiscon and Wine

Please cite this article as: U. Yelipe et al., An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.030



(a) Classification process (b) Imputation process

Fig. 14. Generalized flow of proposed imputation and classification.

datasets. In addition, k-NN classifier with IM-EUC-CBC-FUZZY imputation approach has scored around 38.73% percentage improvement in terms of classifier accuracy when compared with MV-BPCA imputation approach for Wiscon dataset.

Finally, we conclude that CBC-IM-EUC, CBC-IM-Fuzzy and IM-EUC-CBC-Fuzzy imputation algorithms outperforms the other popular Imputation algorithms with respect to classification accuracy and minimum number of decision rules.

4.5. Benchmark datasets description

The description of nine benchmark datasets used for experiments in this paper is shown in the Table 7. It contains information about name of the benchmark dataset, number of attributes, medical records, decision classes for each dataset, missing values information in datasets.

4.6. Flowchart

A pictorial representation of proposed imputation and classification approaches is depicted in Fig. 14.

Please cite this article as: U. Yelipe et al., An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.030

5. Conclusions

In this paper, we addressed a novel technique for imputation called as class-based clustering imputation (IM-CBC). To estimate the similarity between any two medical records, we have used Euclidean distance and proposed fuzzy-similarity functions. Classification is performed using well known classifiers such as kNN, SVM and C4.5. Experiment results obtained using three classifiers SVM, C4.5 and kNN are evaluated with respect to various imputation approaches. Further, we also extend the proposed imputation approach to build a classifier termed as the Class-based Clustering Classifier (CBCC) using Euclidean and fuzzy measures. Classification results obtained using CBCC classifier are compared to results obtained using other three classifiers. Results and graphs discussed show and prove that the proposed approach has comparatively better classification rates to other classifiers. A non-parametric analysis is also performed by considering imputation approaches and classifiers. In future, we wish to extend our approach with new fuzzy measures for medical data values imputation, classification and prediction.

Acknowledgements

We thank all anonymous reviewers for their valuable suggestions and remarks that helped us to improved quality of the paper. Authors also thank the editor in chief, Manu Malek for valuable suggestions in improving the manuscript quality. I am thankful to my guide, P. Sammulal for guiding me through my research work at JNTU Hyderabad towards successful completion of my Ph.D Programme. Special thanks to Anurag Yepuri, student of Information Technology, department, VNR VJIET for his involvement and participation in shaping part of this research. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Zhang S, Qin Z, Ling C, Sheng S. "Missing is useful": missing values in cost-sensitive decision trees. IEEE Trans Knowl Data Eng 2005;17(12):1689–93.
[2] Zhang C, Qin Y, Zhu X, Zhang J, Zhang S. Clustering-based missing value imputation for data preprocessing. In: 2006 IEEE international conference on industrial informatics; 2006. p. 1081–6.
[3] Wang L, Fu D, Li Q, Mu Z. Modelling method with missing values based on clustering and support vector regression. J Syst Eng Electron 2010;21(1):142–7.
[4] Kirkpatrick B, Stevens K. Perfect phylogeny problems with missing values. IEEE/ACM Trans Comput Biol Bioinf 2014;11(5):928–41.
[5] Zhu X, Zhang S, Zhi J, Zhang Z, Xu Z. Missing value estimation for mixed-attribute data sets. IEEE Trans Knowl Data Eng 2011;23(1):110–21.
[6] Farhangfar A, Kurgan L, Pedrycz. "A novel framework for imputation of missing values in databases,. Part A: Syst Hum, IEEE Trans Syst Man Cybern 2007;37(5):692–709.
[7] Choong MK, Charbit M, Yan H. Autoregressive-model-based missing value estimation for DNA microarray time series data. IEEE Trans Inf Technol Biomed 2009;13(1):131–7.
[8] Zhang Z. Missing data imputation: focusing on single imputation. Ann Trans Med 2016;4(1).
[9] Lichman M. UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science; 2013 http://archive.ics.uci.edu/ml.
[10] Lewis HD. Missing data in clinical trials. N Engl J Med 2012;367:2557–8.
[11] Aljuaid T, Sasi S. Intelligent imputation technique for missing values. In: 2016 international conference on advances in computing, communications and informatics (ICACCI), Jaipur; 2016. p. 2441–5. doi:10.1109/ICACCI.2016.7732423.
[12] Thirukumaran S, Sumathi A. Improving accuracy rate of imputation of missing data using classifier methods. In: 2016 10th international conference on intelligent systems and control (ISCO), Coimbatore; 2016. p. 1–7. doi:10.1109/ISCO.2016.7726908.
[13] Razavi-Far R, Saif M. Imputation of missing data using fuzzy neighborhood density-based clustering. In: 2016 IEEE international conference on fuzzy systems (FUZZ-IEEE), Vancouver, BC; 2016. p. 1834–41. doi:10.1109/FUZZ-IEEE.2016.7737913.
[14] Aljuaid T, Sasi S. Proper imputation techniques for missing values in data sets. In: 2016 international conference on data science and engineering (ICDSE), Cochin; 2016. p. 1–5. doi:10.1109/ICDSE.2016.7823957.
[15] UshaRani Y, Sammulal P. An efficient disease prediction and classification using feature reduction based imputation technique. In: 2016 international conference on engineering & MIS (ICEMIS), Agadir; 2016. p. 1–5. doi:10.1109/ICEMIS.2016.7745363.
[16] UshaRani Y, Sammulal P. An approach for imputation of medical records using novel similarity measure. Recent advances in soft computing ICSC–MENDEL 2016 Advances in intelligent systems and computing, 576. Cham: Springer; 2017.
[17] UshaRani Y, Sammulal P. A class-based clustering approach for imputation and mining of medical records (CBC-IM). IADIS Int J Comput Sci Inf Syst 2017;12(1):61–74.
[18] Jiang JY, Liou RJ, Lee SJ. A fuzzy self-constructing feature clustering algorithm for text classification. IEEE Trans Knowl Data Eng 2011;23(March (3)):335–49.
[19] http://sci2s.ugr.es/keel/datasets.php#sub1.
[20] 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 Netw 23 406–418, doi:10.1016/j.neunet.2009.11.014.
[21] Song Q, Shepperd M. A new imputation method for small software project data sets. J Syst Softw 2007;80(1):51–62.
[22] Gira A. Estimation of population mean with a new imputation method. App Math Sci 2015;9(34):1663–72.
[23] Nishanth KJ, et al. Soft computing based imputation and hybrid data and text mining: the case of predicting the severity of phishing alerts. Expert Syst Appl 2012;39(12):10583–9.
[24] Tang F, Ishwaran H. Random forest missing data algorithms. Statist Anal Data Mining 2017.
[25] Petrozziello A, Jordanov I. Column-wise guided data imputation. Proc Comput Sci 2017;108:2282–6.
[26] Vangipuram R, Shadi AA, Kumar P, Janaki V. A novel fuzzy similarity measure and prevalence estimation approach for similarity profiled temporal association pattern mining. Future generation computer systems; 2017. ISSN 0167-739X, https://doi.org/10.1016/j.future.2017.03.016.
[27] Wilcoxon F. Individual comparison by ranking methods. Biometrics 1945;1:80–3.
[28] Perez-Godoy MD, et al. "Training algorithms for radial basis function networks to tackle learning processes with imbalanced data-sets. Appl Soft Comput 2014;25:26–39.
[29] Benavoli A, et al. A Bayesian Wilcoxon signed-rank test based on the Dirichlet process International conference on machine learning; 2014.
[30] Lin W-C, Ke S-W, Tsai C-F. CANN. Knowl-Based Syst 2015;78, C(April (2015)):13–21. http://dx.doi.org/10.1016/j.knosys.2015.01.009.

Please cite this article as: U. Yelipe et al., An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.030

UshaRani Yelipe is an Assistant Professor at VNR VJIET, India and has nearly ten years of academic experience. She pursued B. Tech in Information Science from Nagarjuna university and M. Tech in computer science and engineering from JNTU Hyderabad. She is pursuing PhD in CSE from JNTU Hyderabad under the guidance of Professor Sammulal. She has to her credit several publications in journals and conferences. Her areas of research interest are data mining and software engineering.

Sammulal Porika is an Associate Professor at JNTUH college of engineering, Karimnagar, India. He has several research scholars working towards their doctoral degree under various universities. He has more than 15 years of teaching experience. His areas of interest include data mining, software engineering, network security.

Madhu Golla is associated as a faculty member in the department of Information Technology, VNR VJIET, Hyderabad, India. He has more than 15 years of teaching experience and has several papers published at national and international conferences and refereed journals. He is a member of IEEE, ISTE, CSI. He obtained his doctoral degree from JNTUH in 2015. He has been at and presented papers at International conferences within and abroad and has also received several best paper awards. His areas of research interest are data mining, machine learning, statistical analysis.

Please cite this article as: U. Yelipe et al., An efficient approach for imputation and classification of medical data values using class-based clustering of medical records☆, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.030