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/© 2017 Elsevier Ltd. All rights reserved.

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