Expert Systems with Applications 42 (2015) 5621–5631

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Hybrid prediction model with missing value imputation for medical data

Archana Purwar *, Sandeep Kumar Singh

Department of Computer Science and Information Technology, JIIT Noida, India

ARTICLE INFO

Article history:
Available online 5 March 2015

Keywords:
Missing value imputation
Multilayer Perceptron (MLP)
K-means clustering
Data mining

ABSTRACT

Accurate prediction in the presence of large number of missing values in the data set has always been a challenging problem. Most of hybrid models to address this challenge have either deleted the missing instances from the data set (popularly known as case deletion) or have used some default way to fill the missing values. This paper, presents a novel hybrid prediction model with missing value imputation (HPM-MI) that analyze various imputation techniques using simple K-means clustering and apply the best one to a data set. The proposed hybrid model is the first one to use combination of K-means clustering with Multilayer Perceptron. K-means clustering is also used to validate class labels of given data (incorrectly classified instances are deleted i.e. pattern extracted from original data) before applying classifier. The proposed system has significantly improved data quality by use of best imputation technique after quantitative analysis of eleven imputation approaches. The efficiency of proposed model as predictive classification system is investigated on three benchmark medical data sets namely Pima Indians Diabetes, Wisconsin Breast Cancer, and Hepatitis from the UCI Repository of Machine Learning. In addition to accuracy, sensitivity, specificity; kappa statistics and the area under ROC are also computed. The experimental results show HPM-MI has produced accuracy, sensitivity, specificity, kappa and ROC as 99.82%, 100%, 99.74%, 0.996 and 1.0 respectively for Pima Indian Diabetes data set, 99.39%, 99.31%, 99.54%, 0.986, and 1.0 respectively for breast cancer data set and 99.08%, 100%, 96.55%, 0.978 and 0.99 respectively for Hepatitis data set. Results are best in comparison with existing methods. Further, the performance of our model is measured and analyzed as function of missing rate and train-test ratio using 2D synthetic data set and Wisconsin Diagnostics Breast Cancer Data Sets. Results are promising and therefore the proposed model will be very useful in prediction for medical domain especially when numbers of missing value are large in the data set.

1. Introduction

Research in the field of predictive data mining for medical applications is a significant and moving area. Generally, a medical practitioner collects his/her knowledge from patient's symptoms and confirmed diagnosis. Diagnosis is usually made either by evaluating the current test results of the patients or by referring to the previous decisions made on other patients with same test results. The accuracy of diagnosis of patient's disease like diabetes, breast cancer and others is greatly relenting on a experts' experience (Meesad & Yen, 2003). Due to the pace at which numbers of patients are increasing, it has become cumbersome to make diagnostics decisions. On the other side, development of new computational methods and tools makes relatively easy to make decisions from dedicated databases of electronic patient records. As an example, numerous classifiers have been developed for diagnosis and screening of diabetes, cancer and liver disorders (Seera & Lim, 2014). A number of classification systems have been developed in the literature like RBQ, LVQ C4.5, CART, Bayesian Tree, ANN + FNN, HPM, Sim + F2, real coded GA, and FMM-CART-RF to support diagnosis of diabetes as well as breast cancer disease (Seera & Lim, 2014). Various hybrid models are also proposed in the literature (Kahramanli & Allahverdi, 2008; llango & Ramaraj, 2010; Patil, Joshi, & Toshniwal, 2010) namely hybrid system consisting of artificial neural network and fuzzy neural network, HPM consisting of K-means clustering with J48 and HPM with F-score consisting of feature selection using F-score, K-means clustering and SVM.

Researchers have developed a large number of classification systems to improve accuracy of predictive classification. The proposed model uses K-means clustering as a means to analyze 11 missing data imputation (MVI) techniques under study and selects the best imputation method. This best imputation method is applied on the data set before pattern extraction and subsequently

* Corresponding author.
E-mail addresses: archana.purwar@jiit.ac.in (A. Purwar), sandeepk.singh@jiit.ac.in (S.K. Singh).

http://dx.doi.org/10.1016/j.eswa.2015.02.050
0957-4174/

applying prediction. Moreover, to the best of our knowledge, none of the hybrid prediction models have combined use of K-means clustering and MLP for prediction of diseases.

This paper makes a novel contribution by first analyzing 11 MVI techniques experimentally and finds the best method for handling the missing values in the data set using K-means clustering. Consequently, it improves the quality of data. Moreover, it also aims at predictive classification using novel model that can classify records in the test data set using training data set. Multi layer Perceptron (MLP) has a capability to learn from examples and can generalize beyond the training data (Carpenter & Markuzon, 1998; Downs, Harrison, Kennedy, & Cross, 1996; Mukhopadhyay, Changhong, Huang, Mulong, & Palakal, 2002). Due to these characteristics of neural networks, MLP with backpropagation is investigated for developing a valid and useful prediction model. K-means clustering is also used to develop proposed hybrid prediction model with missing value imputation (HPM-MI) to extract correct instances from data before applying MLP for classification.

Rest of the paper is grouped in five sections. Section 2describes the study of data mining methods namely imputation methods, K-means clustering, MLP and review of prediction models. Then, Section 3depicts the proposed model. Evaluation of proposed model is done in Section 4. Section 5 shows the results and its discussion. Finally, paper is concluded by Section 6.

2. Background study

This section reviews a few data mining methods and predictive classification models.

2.1. Data mining methods

As the amount of data stored in medical databases is increasing, there is growing need for efficient and effective techniques to extract the information. Previous researches have given evidence that medical diagnosis and prognosis is amended by employing data mining techniques on clinical data (Hammer & Bonates, 2006; Saastamoinen & Ketola, 2006; Tsirogiannis et al., 2004). This has been possible due to extensive availability of data mining techniques and tools for data analysis. Predictive modeling requires that the medical informatics researchers and practitioners need to select the most appropriate strategy to cope with clinical prediction problem (Bellazzi & Zupan, 2008). This section discusses mining techniques used to develop the proposed model.

2.1.1. Missing value imputation

2.1.1.1. Introduction. In real-life databases, incomplete data or information as shown in Table 1 is frequent owing to the presence of missing values in the attributes. First row in Table 1 shows name of the variables in the data set while other rows show the values of these variables. The values denoted by '?' in Table 1 represent the missing values. Missing values can occur due to large number of reasons such as errors in the manual data entry procedures, equipment errors or incorrect measurements. The presence of missing values (MVs) in data mining produces several problems in the knowledge extraction process such as loss of efficiency, complications in managing and analyzing data. It may also result in bias decisions due to differences between missing and complete data.

Table 1
Sample data showing missing values by '?'.

A	B	C	D	E	F	G	H	Class
6	148	72	35	?	33.6	0.62	50	Yes
1	85	66	29	?	26.6	0.35	31	No
8	183	64	?	?	23.3	0.67	32	Yes
1	89	66	23	94	28.1	0.16	21	No
?	137	40	35	168	43.1	2.28	33	Yes

In order to solve these problems, two approaches are found in the literature. First approach consists of missing data toleration techniques which integrate the techniques of missing values handling in specific data mining algorithms such as in classification (David, 2007; Saar-Tsechansky, 2007), clustering (Hathaway & Bezdek, 2002) and feature selection (Aussem & de Morais, 2008). Second type of approach consists of missing data imputation techniques which fill in missing values before using complete-data methods on data sets. One advantage of imputation is that the treatment of missing data is independent of the succeeding mining algorithm, and people can select a suitable learning algorithm after imputation (Qin, Zhang, Zhu, Zhang, & Zhang, 2007).

In our proposed HPM-MI, we have used best proven MVI approach to fill the missing value before pattern extraction and classification on the data set. We have validated our model on three benchmark data sets i.e. Pima Indian Diabetes, Wisconsin Breast Cancer and Hepatitis data set from UCI repository (Newman, Hettich, Blake, & Merz, 2007) having 763, 16 and 167 missing values respectively and two complete data sets namely 2D synthetic data set as well as Wisconsin Diagnostics Breast Cancer (WDBC) in which missing values were artificially induced.

2.1.1.2. Imputation techniques. In order to analyze the impact of various MVI techniques (Koren, Bell, & Volinsky, 2009; Luengo, García, & Herrera, 2011; Takács, Pilászy, & Németh, 2008), experimentation is done to choose the best possible one to handle missing values present in the data sets under study. The following approaches have been empirically assessed to find the missing values:

- *Case deletion:* The examples that have any missing value in their attributes are removed from the data set.
- *Most Common Method* (*MC*): Missing values present in the data set is substituted by mean value for numerical and mode for nominal attributes.
- *Concept Most Common* (*CMC*): This method calculates the missing values similar to MC method but it considers only the same class in which MV is missing.
- *K-Nearest Neighbor* (*KNNI*): Firstly, this method finds the *k* nearest neighbors and then, the most common value among all neighbors is taken for nominal attributes, and the mean value is used for numerical attributes.
- *Weighted Imputation with K-Nearest Neighbor* (*WKNN*): This method calculates the distance of each missing value instances from its neighbors. This distance is used to calculate the weight. MV is computed by weighted mean for numerical attributes. For nominal attributes, imputed value is the category with highest weight.
- *K-means Clustering Imputation* (*KMI*): All the instances are clustered using K-means clustering. The instances in each cluster are considered nearest neighbors of each other. The missing value is computed in similar manner as in KNNI method.
- *Imputation with Fuzzy K-means Clustering* (*FKMI*): After the fuzzy clustering of the data set, missing values are computed as weighed sum of all centroids, using the membership function of each cluster as the weight.
- *Support Vector Machines Imputation* (*SVMI*): SVM model is trained to predict missing attributes from the complete instances which do not have missing values. During testing, missing values are predicted using other attributes by setting missing attribute as a class attribute whose value is intended to be predicted.