Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction

Mahin Vazifehdan, Mohammad Hossein Moattar [*], Mehrdad Jalali

Department of Software Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran

ARTICLE INFO

Article history:
Received 27 July 2017
Revised 20 November 2017
Accepted 10 January 2018
Available online xxxx

Keywords:
Breast cancer recurrence
Missing value imputation
Classification
Tensor factorization
Bayesian network

ABSTRACT

Data mining and machine learning approaches can be used to predict breast cancer recurrence. However, real datasets often include missing values for various reasons. In this paper, a hybrid imputation method is proposed with respect to the dependency between the attributes and the type of incomplete attributes in order to especially improve the prediction of breast cancer recurrence. After splitting the dataset into two discrete and numerical subsets, first missing values of the discrete fields are imputed using Bayesian network. Then, using Tensor factorization, the integrated dataset, which comprises of the filled-subset of the previous stage and numerical missing values subset, is constructed so that both continuous missing values are imputed and the accuracy of imputation is enhanced. We evaluated the proposed method versus six imputation methods i.e. mean, Hot-deck, K-NN, Weighted K-NN, Tensor factorization and Bayesian network on three datasets and used three classifiers, namely decision tree, K-Nearest Neighbor and Support Vector Machine for recurrence prediction. Experimental results show that the proposed method has as average 0.26 prediction improvement. Also, the prediction performance of the proposed approach outperforms all other imputation-classifier pairs in terms of specificity, sensitivity and accuracy.

1. Introduction

Nowadays, breast cancer is the second deadliest cancer in Iran. After years of study and research, there are still many unanswered questions facing researchers in various domains, such as prediction, diagnosis and treatment. According to the latest statistics, in Iran, the mean annual number of new cases of breast cancer is approximately 10,000. Among these cases, approximately 2500 patients lose their lives (Sharfian et al., 2015). Women comprise approximately 98% of breast cancer patients and it is worth mentioning that the average age of breast cancer diagnosis in Iranian women is a decade lower than that of the world average (Sharfian et al., 2015).

Recurrence is one of the major problems in breast cancer that means possibility of regrowth of cancer cells in surgery or related areas. The likelihood of post-surgery recurrence affects breast cancer patients' lives at any time. Therefore, recurrence prediction is the main factor for successful treatment of this disease (Kim, 2012). Even though, a large amount of patient information is collected in medical datasets. To benefit from the collected data of patients and increase the accuracy of prediction, a number of researchers have utilized data mining and machine learning approaches for predicting breast cancer (Choi and Jiang, 2010). Classification algorithms are widely used for discovering valuable information from datasets, which can be applicable in the real world. The aim of classification is to predict a class label for each existing sample in the dataset (Zheng et al., 2014). Based on number of features, number of instances, number of classes and the degree of imbalance, results of classification approaches are different.

However, datasets are not always complete. They often include missing values in some samples. This is a major challenge in utilizing data mining approaches for breast cancer prediction. This may occur due to different reasons, such as lack of response from the patients, human errors or system faults for collecting information. Although some of the learning algorithms can work with incom-

[*] Corresponding author.
E-mail addresses: mahinvazifehdan@mshdiau.ac.ir (M. Vazifehdan), moattar@mshdiau.ac.ir (M.H. Moattar), jalali@mshdiau.ac.ir (M. Jalali).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2018.01.002
1319-1578/© 2018 Production and hosting by Elsevier B.V. on behalf of King Saud University.

plete data, most of them are not able to handle missing values. They discard the samples that contain at least one missing value or assign a valid value to the corresponding attribute (Zheng et al., 2014; García-Laencina, 2015; Tutz and Ramzan, 2015; Little and Rubin, 2002). Removing incomplete data is an acceptable method but only when there is a little proportion of missing values i.e., 5%. With the increase of missing ratio, using this method leads to valuable information loss. Imputation of missing values is thus necessary for making efficient predictions using data mining tools (García-Laencina, 2015).

Since 1980, many techniques for missing data imputation have been suggested (García-Laencina, 2015). Ref. Little and Rubin (2002) mentions three patterns of missing values that can affect the performance of imputation method which are: 1) Missing Completely at Random (MCAR), when missing value belongs to an instance that does not depend on either the observed data or the missing data. 2) Missing at Random (MAR) in which missing value belongs to an instance that only depends on the observed data and not the missing data. And finally 3) Missing Not at Random (MNAR), when missing value belongs to an instance that depends on unobserved data. Most studies assume that the pattern of missing values is MAR (García-Laencina, 2015; Dauwels et al., 2012). Thus, in this research, the same assumption is made.

The main goal of this study is to propose an imputation method using a hybrid of two methods to improve the prediction of breast cancer recurrence. Due to existence of discrete and continuous values, especially in medical datasets, at first, we benefited from Bayesian network for imputation of discrete missing values. Then, we have used reconstruction of the dataset by Tensor factorization for improving the performance of imputation. In addition, we compared the proposed method with the imputation based on Tensor (Dauwels et al., 2012) and Bayesian model (Rancoita, 2014) and some other well-known methods such as mean, Hot-deck, k-nearest neighbor and Weighted K-NN on three datasets. Finally, three classifiers namely, Support Vector Machine (SVM), Decision Tree (DT) and K Nearest Neighbored (KNN) are applied on the imputed datasets to predict breast cancer recurrence.

The remainder of this paper is organized as follows: Section 2 reviews previous studies that include both breast cancer recurrence prediction and imputation of the missing values. Section 3 describes the materials and methods that are used in this paper. Section 4 proposes the new approach for imputation. Section 5 explains the details of the experimental studies and discusses the results. Finally, Section 6 summarizes and concludes the paper.

2. Related works

In this section, previous works which are related to either breast cancer recurrence prediction or missing values imputation are reviewed. Jerez-Aragonés et al. (2003) proposed a combination of decision tree and neural network to predict breast cancer recurrence during different periods of time based on clinical and laboratory data. A novel decision tree called control of induction by sample division method (CIDIM), which is a beneficial tool for representing the relationship among features, has been proposed to select the most relevant diagnosis factors. Next, selected factors have been used as inputs to neural network system. Sun et al. (2010) examined the performance of a two-way approach to evaluate the prediction of breast cancer recurrence using three classifiers (linear SVM, SVM-RFE (Wilin, 2009), and L1 Regularized logistical regression (Ng, 2004). Two datasets, namely, Nature (Van't Veer et al., 2002) and JNCI (Buyse, 2006) were experimented, from which one is used as training set and the other as test set. They also developed a feature selection method for their approach.

Kim (2012) investigated a diagnostic model based on SVM to predict breast cancer recurrence (namely BCRSVM) and it has been compared with two other methods, i.e., Neural network and Regression model. Wang (2014) proposed the combinations of SMOTE, PSO, and three popular classifiers including C5, Logistic Regression, and 1-NN for predicting 5-year survivability of breast cancer patients. SMOTE is an over-sampling based method that creates new synthetic instances in the minority class for balancing the dataset. Feature selection is conducted using PSO algorithm as well. Their results indicate that the hybrid of SMOTE, PSO and C5 is the best framework among all possible combinations.

Batista and Monard (2003) proposed three imputation methods, namely, Hot-deck, mean and k-nearest neighbor and compared them on four datasets. These approaches where evaluated using two methods namely, C4.5 decision tree and CN2 (Clark and Niblett, 1989). Farhangfar et al. (2008) examined the effect of six classifiers i.e., C4.5, k-nearest neighbor, RIPPER (Cohen, 1995), Naïve Bayes and SVM with RBF and polynomial kernel on 15 datasets with missing ratios of 5%, 10–50% with 10% increments and five imputation methods.

Jerez (2010) examined three statistical imputation methods, i.e., mean, Hot-deck, and a hybrid of them and three Machine Learning methods, i.e., k-nearest neighbor, self-organization maps (SOM) (Kohonen, 1995) and multi-layer perceptron (MLP) (Bishop et al., 2013) on breast cancer data. They also introduce breast cancer recurrence prediction with neural network as their final objective. The results of their work indicate that ML methods are better than statistical algorithms. Dauwels et al. (2012) utilized Tensor (especially, CP and normalized CP factorization) for imputation of missing data on medical questionnaires. They compared the approach with mean, k-nearest neighbor, and iterative local least square (Cai et al., 2006) with missing ratios of 10%, 20% and 30%. The experimental results suggest that Tensor imputation outperforms the other methods.

Aydilek and Arslan (2013) proposed a combination approach of optimized fuzzy c-means with support vector regression (Vapnik et al., 1996) and genetic algorithm for imputation of missing values. They considered genetic algorithm for optimizing fuzzy c-means parameters including number of clusters and weighting factor. The method is compared with three imputation methods, namely fuzzy c-means, SVR genetic (SvrGa) and Zero imputation with missing ratios of 1% and 5–25% with increment of 5%.

Although acceptable results are obtained in studies related to prediction of breast cancer recurrence, they are not considered as an improvement of recurrence prediction from the perspective of missing values imputation, and their limitation is the use of old statistical methods. Regarding previous missing data estimations, it should be noted that most of them, fill the missing data irrespective of the dependencies between attributes and the type of incomplete attribute.

When facing missing values, classifiers often either remove the instance containing missing value or impute it using various imputation methods. Based on our researches and studies, we categorize imputation models into four groups which are summarized in Fig. 1. Although many other imputation methods fit in these categories, we mention only some instances. Also this paper chooses representative methods from each group both to evaluate the accuracy of the proposed method and create a set of new and well-known methods.

3. Materials and methods

The following is a brief description of each imputation method and each predictive model that is used in this paper. The imputation methods are representative methods from the three