Journal of King Saud University – Computer and Information Sciences xxx (2018) xxx–xxx

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/

Please cite this article in press as: Vazifehdan, M., et al. A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.01.002

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

Please cite this article in press as: Vazifehdan, M., et al. A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.01.002

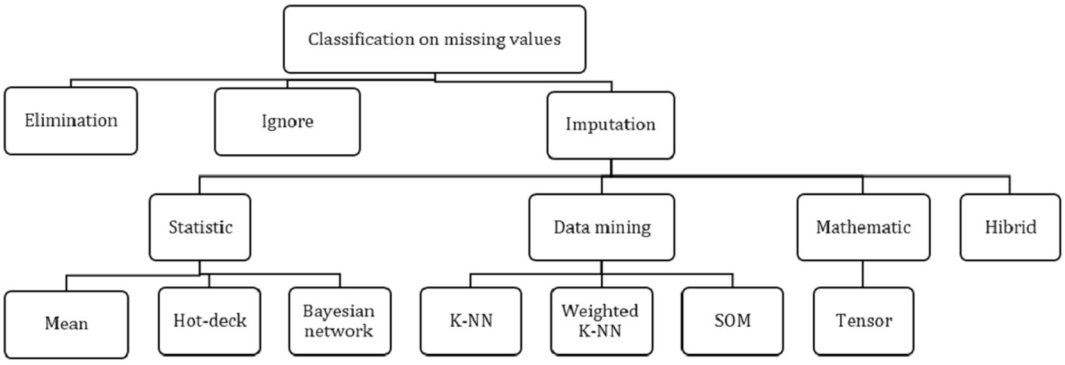

Fig. 1. Classification of missing values imputation methods.

imputation modes including three statistical models, two ML-based models and one mathematical-based model.

3.1. Imputation method

3.1.1. Mean/Mode imputation

This method is one of the earliest and most popular models that is known by different names such as mean imputation (Jerez, 2010), mean substitution (Dauwels et al., 2012), most common method (Purwar and Singh, 2015), etc. In this method, the missing value is imputed by using the mean of observed data (for continuous attribute) and the most frequent value (mode) (for discrete attribute) of the corresponding attribute (Zheng et al., 2014; Little and Rubin, 2002; Cohen, 1995; Vapnik et al., 1996; Purwar and Singh, 2015; Malarvizhi and Thanamani, 2012). Ref. (Malarvizhi and Thanamani, 2012) states that imputation error rates of median and standard deviation are less than that of the mean imputation method. A disadvantage of this method which can be mentioned is lack of attention to dependency between attributes.

3.1.2. Hot-deck imputation

Hot-deck is an old statistical model. The procedure is as follows: each incomplete record is compared with all other examples and missing value of the corresponding record is replaced by the value of the same attribute of the most similar record. Mean and Hot-deck belong to single methods, which means that they impute the missing value with only a single sample but their main drawback is the lack of attention to the correlation between attributes. On the other hand, the computational cost might be increased because of comparison with every example (Clark and Niblett, 1989; Cohen, 1995; Purwar and Singh, 2015; Malarvizhi and Thanamani, 2012).

3.1.3. K-Nearest neighbor imputation

K-NN (Cover and Hart, 1967) is one of the most frequently used machine learning approaches. In this model, the missing value of each example is imputed with k most similar neighbor values in training space. The mean value of neighbors is used for numerical attributes and the mode value is used for discrete attributes (Zheng et al., 2014; Little and Rubin, 2002; Cohen, 1995; Vapnik et al., 1996; Purwar and Singh, 2015). Two parameters play an important role in the performance of imputation which are the number of neighbors and the distance measurement. Euclidean distance is a well-known distance measure for this purpose and we apply it in our experiments as follows:

$$d_{p,q} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{1}$$

where n is the number of attributes of each instance and p_i is the i th attribute value in p instance. One of the disadvantages of KNN is its high dependency on k and it often gives an acceptable and reliable answer for k values between 5 and 10 while higher k impose negative effect on the performance of the imputation (Tutz and Ramzan, 2015). The K-NN imputation works on gene data better than mean and Singular Value Decomposition (SVD) methods as presented in Ref. (Troyanskaya, 2001). 'Hot-deck' is also known as the KNN imputation where k = 1.

3.1.4. Weighted K-NN model

In the K-NN model, in order to impute the missing value, the first neighbor and the k th neighbor are equally important, while the first neighbor, in general, is more important than other neighbors. Therefore, Weighted K-NN is defined as developed K-NN model so that to any neighbor a specific weight is assigned where the first neighbor (closest) has the highest weight value and the k th neighbor (furthest) has the lowest weight value (Tutz and Ramzan, 2015; SolaroEmail et al., 2017).

3.1.5. Tensor model

Tensors or multi-dimensional arrays are known as hypermatrices. They are extensions of vectors (first-order tensor) and matrices (second-order tensor) that introduce arrays of higher order ($N > 2$). For example, a third-order tensor is an array with elements, $X_{i,j,k}$. Tensor factorization was introduced by Hitchcock (Hitchcock, 1927). Tensor factorization is generally a computationally expensive task but with a high precision (Yang, 2017). Tucker and Canonical Polyadic (CP) are two tensor factorization models which are commonly used. In particular, CP initially divides the overall tensor into rank-one tensors and next, using tensor multiplication, most frequently used type of which is Kronecker, it reconstructs the primary tensor. Kronecker product of two matrices, $A \in R^{I\times J}$ and $B \in R^{K\times L}$ can be obtained as follows (Acar, 2009):

$$A \otimes B = \begin{bmatrix} a_{11}Ba_{12}B \dots a_{1J}B \\ a_{21}Ba_{22}B \dots a_{2J}B \\ \cdots\cdots\cdots\cdots \\ a_{I1}Ba_{I2}B \dots a_{IJ}B \end{bmatrix} \tag{2}$$

Tensor reconstructs missing values with the linear combination of other features. If x be considered as a three-rank tensor of size $I \times J \times K$ and R be assumed as the number of broken matrices or rank of tensor, CP factorization is created by factor matrices A, B and C of size $I \times R$, $J \times R$ and $K \times R$ respectively such that the following equation holds for all values of $i = 1 \dots I$, $j = 1 \dots J$ and $k = 1 \dots K$:

$$x_{ijk} = \sum_{r=1}^{R} a_{ir}b_{jr}c_{kr} \tag{3}$$

Please cite this article in press as: Vazifehdan, M., et al. A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.01.002

The principal objective of CP factorization is minimizing the reconstruction error rate of primary tensor so that sum of one-rank tensors has the least difference with the original tensor and the following f function has the lowest value (Wang et al., 2017).

$$f(A,B,C) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(x_{ijk} - \sum_{r=1}^{R} a_{ir}b_{jr}c_{kr}\right)^2 \quad (4)$$

The CP factorization cannot suitably impute the missing data with high percentage of missing-ness; nevertheless, improved CP model, namely, Weighted CP algorithm (CP-WOPT) has been proposed so that a weight tensor with the same size as the original tensor is considered for imputing missing values. Therefore, f function can be formulated as below:

$$f(A,B,C) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left\{w_{ijk}\left(x_{ijk} - \sum_{r=1}^{R} a_{ir}b_{jr}c_{kr}\right)\right\}^2 \quad (5)$$

where w is a non-negative weight tensor which can be initialized as follows for all $i = 1 \ldots I, j = 1 \ldots J$ and $= 1 \ldots K$ (Acar et al., 2009):

$$w_{ijk} = \begin{cases} 1 \text{ if } x_{i,j,k} \text{ is know} \\ 0 \text{ if } x_{i,j,k} \text{ is unknow} \end{cases} \quad (6)$$

3.1.6. Bayesian network model

Bayesian network is known as belief network and it also belongs to the family of probabilistic graphic models (Dauwels et al., 2012; Franzin et al., 2017; Dempster et al., 1977; De Campos, xxxx). This network is composed of a directed acyclic graph (DAG) where nodes are associated to attributes that represent both dependency among attributes and a joint probability distribution (Pr_M) over a collection of discrete variables. Bayesian network can be denoted as a triple $M = (\mathcal{G}, X, P)$ where $\mathcal{G} = (V_\mathcal{G}, E_\mathcal{G})$ is the dependency graph of X variables including $V_\mathcal{G}$ as a set of m nodes (a node per variable) and E_G as a set of edges relationships between variables. P is a set of conditional probabilities, $Pr_M(X_i|PA_i)$, where PA_i refers to the nodes which X_i depends on them (called parents of X_i, it may be empty or a subset of variables $V_\mathcal{G}$). The main capability of Bayesian network is its Markov structure, which means that each attribute X_i can be conditionally independent of non-descendants while having its parent (pa_i). Bayesian network can show the joint probability distribution as the following equation:

$$Pr_M(X_1 \ldots X_m) = \prod_i Pr_M(X_i|PA_i) \quad (7)$$

Despite the fact that learning and inference are quickly performed in Bayesian network, it has a major challenge. This method is mostly used for datasets containing discrete, binary data where every attribute V_i commonly has a finite number of values ($v_{i_1} \ldots v_{i_{n_i}}$). Although it can be trained with continuous values (Rancoita, 2014). There are generally two issues in Bayesian network: network structure and learning its parameters. In network structure, it is tried to detect the best DAG for a given database. On the other hand, the learning of parameters means setting of parameters of conditional probability distribution based on corresponding dataset. Expectation Maximization algorithm (EM) (Franzin et al., 2017; Dempster et al., 1977) is one of the best methods in Bayesian network. It is a repetitive procedure for estimating the highest probability where only subset of the total data is observed. Therefore, its advantage is that it is successfully trained with missing data. This method consists of two steps; first, Expectation step (E-step) which calculates log probability of data and based on this calculated log, current structure and networks parameters are characterized. In maximization step (M-step)), we begin to find the parameters for maximizing the previous step probabilities and updating network structure. This procedure is repeated until it enhances neither the network structure nor parameters values. The main drawback of EM method is the use of a local searching function to build the best DAG in M-step while local search requires much computation time. Hence, Ref. Rancoita (2014) employed available global searching functions such as K2 local search (Little and Rubin, 2002), Branch-and-bound (BB) (De Campos, xxxx), dynamic programing (DP) (Silander and Myllym, 2004) and linear integer programing (IP) (Jaakkola and Meila, 2010) for this purpose. Learning process usually starts by running K2 algorithm for finding an improved solution. If optimum network structure is obtained in a period of time, the process is stopped or returned; otherwise; one of the other methods (BB, IP or DP) are applied depending on the number of variables.

3.2. Predictive models

3.2.1. Decision tree

Decision tree is a supervised model and a useful, understandable and simple method for classification. Its inputs and outputs are labeled training data and an organized sequential tree structure, respectively. One of its basic advantages can be stated as decomposition of complex problems into smaller and simpler problems. Classification process consists of labeling the input instances by a traversing from the root node to the leaf nodes with respect to the fact that the leaf nodes are class labels. There are two issues involved in building decision tree: (1) growing the tree as long as training set instances are accurately classified. (2) Pruning until the unnecessary nodes are eliminated in order to improve the overall accuracy (Zheng et al., 2014; Buyse, 2006; Jaakkola and Meila, 2010). C4.5 is a kind of decision tree which can be used for both discrete and continuous data. Hence, we benefited from this method for our classification task.

3.2.2. K- nearest neighbor classifier

This method is a well-known instance-based method. The label class of a new instance is predicted using the majority of k nearest neighbors labels which belong to training instances based on a certain distance measure.

K-NN involves two important issues: (1) the number of neighbors (k), (2) distance measure (d). If the number of neighbors is low, the outlier examples may affect the results, while large number of neighbors may face the interference of unrelated data (Zheng et al., 2014; Buyse, 2006; Clark and Niblett, 1989; Purwar and Singh, 2015; Cortes and Vapnik, 1995) .In this study, the optimal number of neighbors can be obtained using cross validation procedure.

3.2.3. Support vector Machine

Support vector machine (Cortes and Vapnik, 1995) is a kernel-based method that is also widely used for classification. If the training set is linear separable, SVM makes hyper planes with maximum margin; otherwise, it is mapped to other space with greater dimension to be linearly separable (Vidyasagar, 2017). Though SVM particularly works for dataset containing two classes and its basic idea is finding the optimal discrimination between two classes, there are also ways to be extend it for multi-class datasets, i.e. one-against-all (OAA) and one-against-one (OAO). OAA approach requires k separators for k-class classification such that every separator is used to separate one class from all other classes. While OAO approach creates binary vector machines for each possible combination of classes (a binary vector machine per possible combination of classes). OAA algorithm is generally considered a k $(k-1)/2$ binary vector machine (Choi and Jiang, 2010).

Please cite this article in press as: Vazifehdan, M., et al. A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.01.002

4. Proposed approach

As stated before, Bayesian network is a strong model for representing conditional dependencies between variable, but this model lacks generalization to continuous variables due to the necessity of prior knowledge for the conditional densities. Due to this reason, Bayesian network is proper for imputing missing values when data is of finite domain variables. On the other hand, tensor factorization estimates missing features with the linear combination of others (not necessarily continuous or discrete) and therefore the estimated value is usually precise. However, constructing tensors in the presence of large amount of missing values is erroneous. Therefore, for better value estimation, in the proposed approach the categorical values are firstly estimated using Bayesian network and then the partially completed dataset is fed to the tensor factorization approach for imputing continuous missing values.

Tensor factorization learns the latent structure and dependencies among different dimensions of the tensor which represents variables. This property is definitely the advantage that some other missing value imputation approaches lack. When Bayesian network can model dependencies between relatively fewer variables, tensor factorization is an effective approach for capturing dependencies in higher dimensional data. On the other hand, Bayesian network represent dependencies in the form of conditional probability functions which are hard to estimate and highly depend on prior assumptions.

Fig. 2 shows the proposed framework for missing values imputation and especially for predicting breast cancer recurrence. The framework consists of four subsections: splitting original dataset into discrete attributes set and continuous attributes set, Bayesian network imputation, reconstruction and imputation using tensor and class prediction with classifiers. Since it is important whether the data is pre-processed or not before missing values imputation, it should be stated that we have not applied any pre-processing to the data. In the first step, we split vertically the entire dataset into two subsets: a subset with continuous missing attributes and a subset with categorical missing attributes. Secondly, Bayesian network, computes the probability of possible values. Then, the most likely value replaces the discrete missing value. This procedure is repeated for each missing value until the discrete dataset is completed. Bayesian network model for categorical attributes of Omid dataset is shown in Fig. 3.

After estimating non-numerical missing values, integrated dataset from discrete imputed subset and the subset of numerical missing data are reconstructed using tensor; and continuous missing values are also imputed using this reconstruction. Reconstruction process is executed as long as the convergence condition is satisfied. That is, the difference between successive estimated values is minimized and no change can be occurred in the next iterations. We have applied mean squared deviation (MSD) as the convergence measurement, which can be formulated as follows:

$$MSD = 1/n\sum_{i=1}^{n}(output(t)_i - output(t-1)_i)^2 \quad (8)$$

In which $output_i(t)$ denotes the estimated value in tth iteration and n denoted the total number of missing values.

After imputing missing values, SVM, DT and K-NN classifiers are applied on complete dataset for prediction.

The proposed method is summarized as follows:

Input:

1. An $n \times d$ original dataset with missing values

Output:

1. An $n \times d$ imputed dataset
2. Breast cancer recurrence prediction with imputed datasets

Step 1: Split the dataset into two subsets with continuous and discrete values.

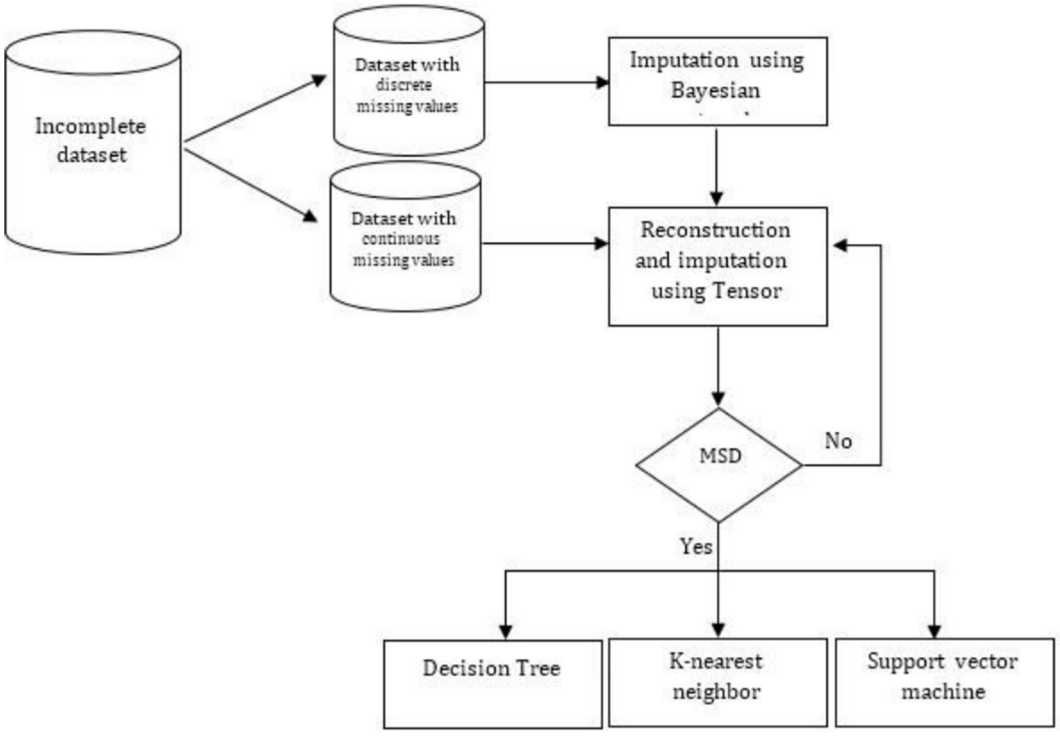

Fig. 2. The proposed framework.

Please cite this article in press as: Vazifehdan, M., et al. A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.01.002

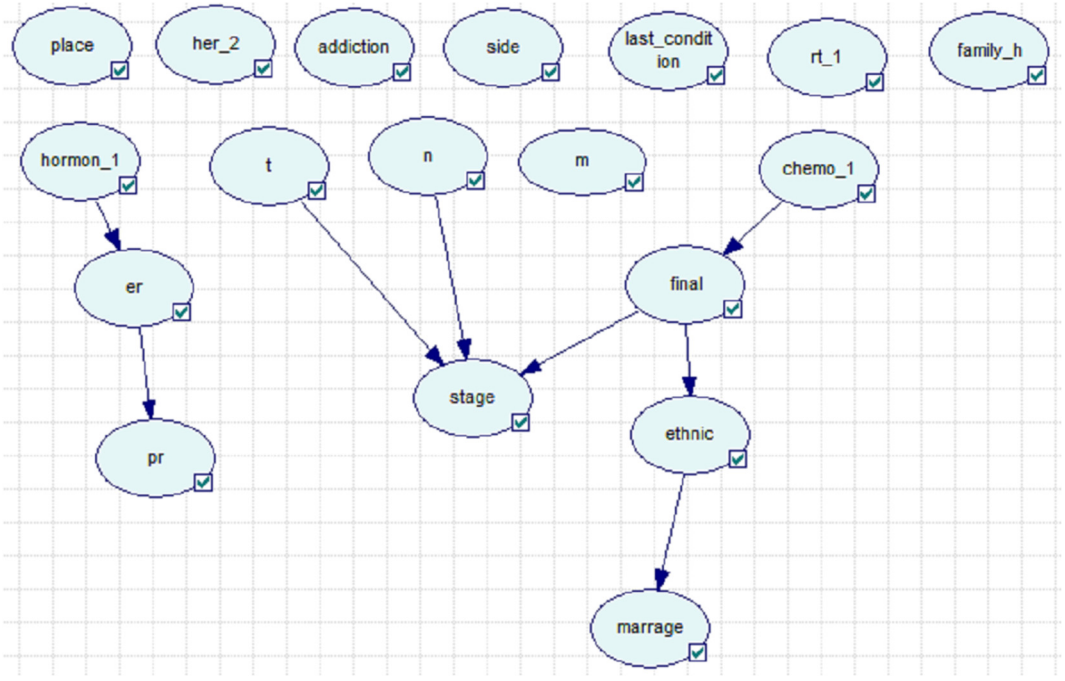

Fig. 3. Dependency graph of discrete attributes in Omid dataset.

Step 2: Impute discrete missing values subset using Bayesian network.

Step 3: Integrate of discrete imputed subset and continuous missing values subset.

Step 4:

- Convert the dataset to tensor.
- Divide the entire tensor into rank-one tensors.
- Create a non-negative weight tensor that is the same size as the original tensor.
- Impute continuous missing values and reconstruct rank-one tensors by CP-Wopt function and repeat it until the convergence is reached.

Step 5: Return the existing data of original dataset and preserve imputed values.

Step 6: Apply the classification models (DT, K-NN and SVM) via complete imputed dataset.

5. Evaluation

One of the main goals of the experiments is to examine the effect of several imputation methods on the accuracy of subsequent prediction models. We start by detailed description of our real dataset and experimental settings and then proceed to performance evaluation and analysis.

5.1. Datasets

We used women breast cancer dataset of Omid Hospital of Mashhad, Iran that is collected during 2000–2010. This dataset includes clinical and laboratory data for 217 instances with 22 variables, namely, Age, Height, Weight, Overall survival (period of admission until complete treatment), Ethnicity, Place, Marital Status, Tumor Size (T), The number of involved cells (N), Metastasis (M), Side of cancer, Last condition of patient, Family History, Addiction, Radiotherapy, Chemotherapy, Hormone therapy, Stage, Final Status, PR, ER and HER2. These variables have been approved by Omid treatment center.

Disease Free Survival (DFS) is assumed as target class which medically means the period from admission of the patient until the recurrence of the patient's disease (it may be before leaving hospital or after that). The values of DFS range from 0 to 149 months; therefore, we grouped them into 4 categories: in the first class, relapsed instances belong to the first 11 months (the number of instances in this group is 60), in the second class, relapsed instances range between 11 and 34 months (51 instances), between 34 and 56 in the third class (51 instances), and between 56 and 149 months in the fourth class (55 instances). The highest missing rate belongs to HER with 29.41%. The brief description of the statistical information and missing percent for each variable is shown in Table 1. Only 96 instances of 217 instances are complete.

Also, to evaluate the performance of proposed method, we used two known datasets: Wisconsin and Cleveland, which are publically available to researchers in UCI machine learning repository (Blake and Merz, 1998). Hence, there is no missing value in Wisconsin dataset; we only used this dataset to evaluate the precision of the proposed missing value imputation in term of estimation error. The basic information of datasets is summarized in Table 2.

5.2. Evaluation settings

In general, the experiments are implemented on a system with hardware characteristics including Intel core 7 3.40 GHz, 16 G RAM, 2 TG Hard Disk and Windows 7. We employed Matlab (R2014a) as well as Tensor Toolbox (Bader and Kolda, 2015) and Poblano Toolbox (Dunlavy et al., 2010) for executing methods used in this paper.

5.3. Evaluation setup

To evaluate the performance of the proposed method for missing value estimation, we inserted different amounts (i,e, 5, 10 and

Please cite this article in press as: Vazifehdan, M., et al. A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.01.002

Table 1
Attributes of Omid dataset for breast cancer recurrence prediction.

Missing (%)	Variance	Mean/Mode	Type (Range)	Attributes	No.
0	163.07	61.34	Numerical (32–85)	Age	1
0	38.21	154.19	Numerical (135–195)	Height	2
0	188.01	62.04	Numerical (39–112)	Weight	3
0	979.54	46.46	Numerical (0–159)	Overall Survival	4
0.69	1.57	1	Categorical (1–8)	Ethnic	5
0	0.15	1	Binary	Place	6
3.46	0.11	1	Binary	Marriage	7
2.76	0.65	2	Categorical (1–4)	T	8
4.15	0.89	1	Categorical (1–4)	N	9
0	0.08	0	Binary	M	10
7.26	0.26	1	Binary	Side	11
0	0.17	1	Binary	Last Condition	12
0.69	0.19	2	Binary	Family History	13
5.19	0.09	1	Binary	Addiction	14
0	0.24	1	Binary	Radiotherapy	15
0	0.06	1	Binary	Chemotherapy	16
0	0.25	0	Binary	Hormone therapy	17
0.34	0.42	2	Categorical (1–4)	Stage	18
4.84	0.58	1	Categorical (1–7)	Final	19
20.41	4.04	0	Binary	PR	20
20.76	3.97	1	Binary	ER	21
29.41	0.66	1	Categorical (1–3)	HER2	22

Table 2
Basic information of the datasets used in this paper.

Datasets	Records	Attributes	Missing values	Pure records
Omid	217	22	Yes	96
Wisconsin	569	32	No	569
Cleveland	303	13	Yes	297

15 percent) of random missing values in the evaluation datasets, even Wisconsin dataset that does not contain missing values. Since our real dataset is of various ranges, we compared the accuracy of the imputation methods using the normalized root mean square error (NRMSE) that can be defined as follows (Dauwels et al., 2012):

$$NRMSE = \frac{1}{\max - min}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x_i')^2} \tag{9}$$

where x_i is the real value and $x\prime_i$ is the imputed value. x_{max} and x_{min} are maximum and minimum values, respectively.

One of the most popular and well-known measures for examining the classification performance is accuracy which is applied to recurrence prediction in this study. It refers to the ability of the model for correct prediction of class label of unobserved cases (García-Laencina, 2015). Furthermore, sensitivity and specificity measures have been used to analyze correct and incorrect decisions by the corresponding classifier. These three measures can be calculated as below:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{11}$$

$$Specificity = \frac{TN}{FP + TN} \tag{12}$$

Where TP, TN, FP and FN denote true positive, true negative, false positive and false negative, respectively. For example if a dataset has two classes, true positive indicates the number of correct classifications that belong to the first class and true negative is the number of correct classifications that belong to the second class. On the other hand, false positive and false negative give us respectively the number of instances that are incorrectly predicted in the first class while they belong to the other class and the number of instances that are incorrectly predicted in the second class while they belong to the first class.

When the number of classes is more than two, to evaluate the performance of classifier, we must obtain the above equations for each class separately such that each class is considered as first class and all other classes as second class. After computing the above equations (Eqs. 10–12) for every class, we apply the average of these values for the final result.

The proposed approach of missing data imputation is compared with six imputation methods, i.e., mean, Hot-deck, K-NN, Weighed K-NN, Tensor-based imputation (Dauwels et al., 2012) and Bayesian network imputation (Rancoita, 2014). In literature, the mentioned approaches are applied for both numerical and categorical attributes in the same way. These approaches do not use different paradigms for different variable. Therefore, we have applied these approaches for both discrete and numeric values the same. We firstly empty 5%, 10% and 15% of a whole dataset. Then, we estimated the missing values via the imputation methods and compared the resulted values with the actual ones using NRMSE measurement (Eq. (9)).

Table 3
Parameters of some of imputation and prediction models.

Parameter	Method	Task
Number of nearest neighbors = 5;Distance = Standardized Euclidean	K-NN W-KNN	Imputation
Kernel function = RBF;Order of the RBF kernel = 4	SVM	Prediction
Number of nearest neighbors = 5;Distance = Pearson Correlation	K-NN	

Please cite this article in press as: Vazifehdan, M., et al. A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.01.002

Table 4
NRMSE of imputation methods on three datasets with 5–15% missing rates.

Datasets	Imputation Methods	NRMSE (lower value is better)		
		Missing rate: 5%	Missing rate: 10%	Missing rate: 15%
Omid	Mean	0.31	0.34	0.36
	Hot-deck	0.66	0.74	0.82
	K-nn	0.54	0.70	0.73
	W-knn	0.37	0.40	0.42
	Tensor	0.18	0.21	0.25
	Bayesian network	0.15	0.17	0.20
	Proposed Method	**0.09**	**0.12**	**0.16**
Wisconsin	Mean	0.32	0.33	0.34
	Hot-deck	0.35	0.37	0.38
	K-nn	0.27	0.29	0.32
	W-knn	0.17	0.19	0.21
	Tensor	0.11	0.17	0.20
	Bayesian network	0.09	0.15	0.17
	Proposed Method	**0.06**	**0.11**	**0.13**
Cleveland	Mean	0.33	0.35	0.35
	Hot-deck	0.55	0.59	0.62
	K-nn	0.58	0.61	0.63
	W-knn	0.35	0.37	0.38
	Tensor	0.16	0.21	0.24
	Bayesian network	0.13	0.19	0.22
	Proposed Method	**0.08**	**0.11**	**0.14**

In this work, 5-fold cross validation procedure is used in order to evaluate predictive models. The dataset is split randomly into 5 folds. One fold is considered for test and all the others for training. To ensure the stability of the results, the number of experiments is five. We reported only the average results for each experiment. Table 3 introduces the parameter settings for imputation and classification methods.

5.4. Experimental results

Imputation of the missing data according to the proposed method is conducted to improve breast cancer recurrence prediction. Table 4 shows the results of NRMSE for estimation approaches on the three datasets (Omid, Wisconsin and Cleveland dataset). Also Fig. 4 illustrates the results using curves. In these results,

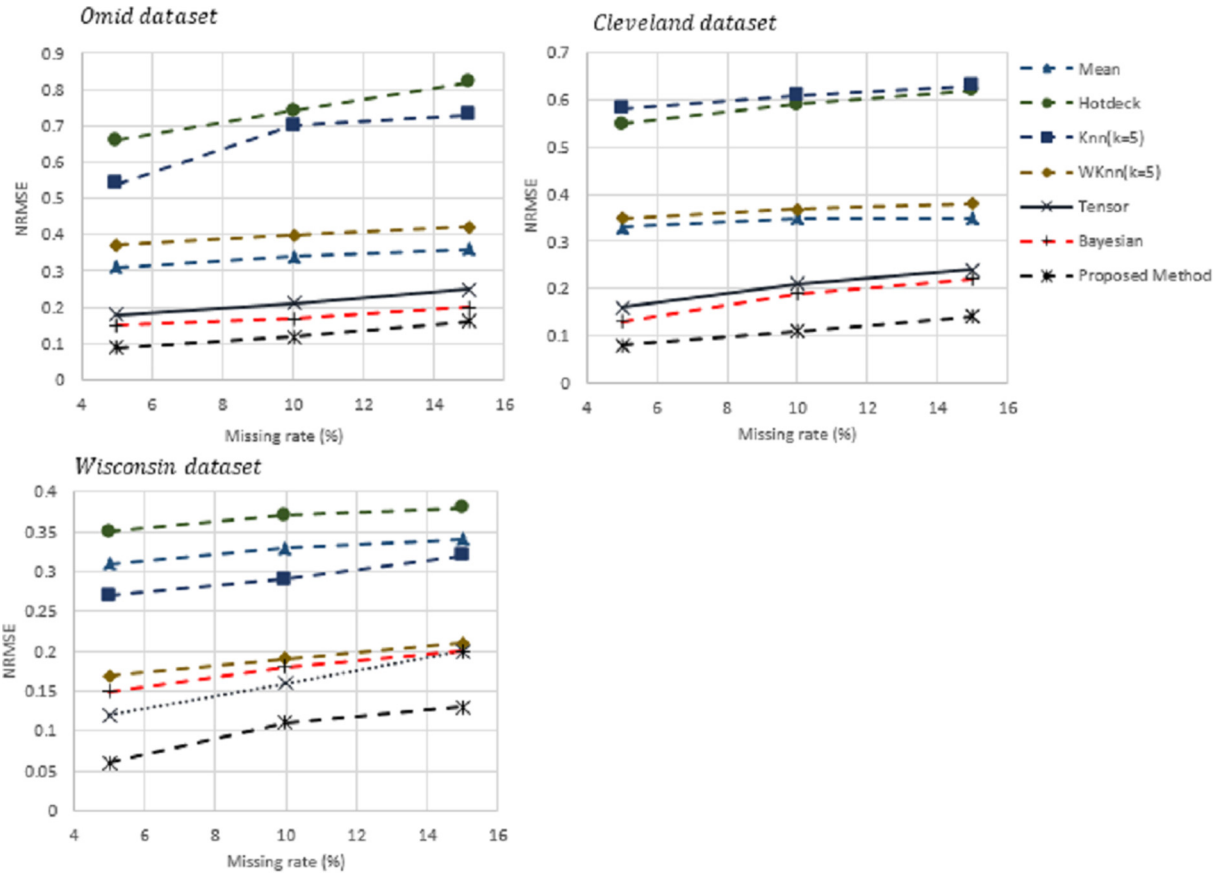

Fig. 4. NRMSE of imputation methods for Omid, Wisconsin and Cleveland datasets.

Please cite this article in press as: Vazifehdan, M., et al. A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.01.002

the proposed method obtained the lowest error rates and was more efficient than the other methods (the NRMSE for Omid dataset is 0.12, the NRMSE for Wisconsin dataset is 0.10 and the NRMSE for Cleveland dataset is 0.11).

In these results, the proposed method obtained the lowest error rate and is better than other methods. As expected, since the number of discrete categorical attributes are more than continuous attributes, Bayesian network based imputation performance is superior compared to Tensor imputation. On these datasets, W-KNN, KNN and Hot-deck do not work very well in dealing with continuous missing values. Table 5 shows the results of classifiers (Eqs. 10–12) on Omid dataset. The proposed method has achieved the best result with an average accuracy of 89.29%, sensitivity of 78.55% and specificity of 92.83% with C4.5 classifier which has an increased accuracy as compared with Tensor-based imputation and Bayesian network imputation. The same results are also depicted in Fig. 5.

6. Discussion and conclusion

The recurrence of breast cancer affects the lives of patients even many years after Decision TreeKNNSVMsurgery. In recent years, machine learning and data mining methods have increasingly improved predictions and helped medical professionals. Extracting verified information from collected medical data is considered as a major challenge. Due to the increase of cancer patients, especially breast cancer patients, the research about this field is very important. The existence of missing values in medical data is a main challenge in this field. A precise estimation of missing values, which leads to better decision, or at least assist experts for this purpose, is valuable in cancer diagnosis and recurrence prediction. In this paper, a new approach for missing values imputation is proposed with respect to dependencies among variables and the type of incomplete variable which significantly affects imputation using Tensor and Bayesian networks for both categorical and numerical

Table 5
Breast cancer recurrence prediction accuracy on Omid dataset with 10% imputed missing values.

	C4.5			K-NN			SVM		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Mean	75.37	91.92	87.78	70.47	90.21	85.16	53.70	72.36	69.79
Hot-deck	76.65	92.35	88.30	70.71	90.52	85.61	48.70	70.57	67.70
K-nn	77.42	92.48	88.72	71.09	90.62	85.71	**58.85**	72.49	70.83
W-knn	76.64	92.31	88.47	71.49	90.62	85.75	63.70	72.36	70.83
Tensor	77.63	92.60	88.94	71.79	**90.81**	86.08	55.62	**73.70**	**71.87**
Bayesian network	77.21	92.40	88.50	71.36	90.70	85.80	55.62	72.95	70.40
Proposed Method	**78.55**	**92.83**	**89.29**	**71.99**	90.80	**86.16**	58.75	73.07	71.35

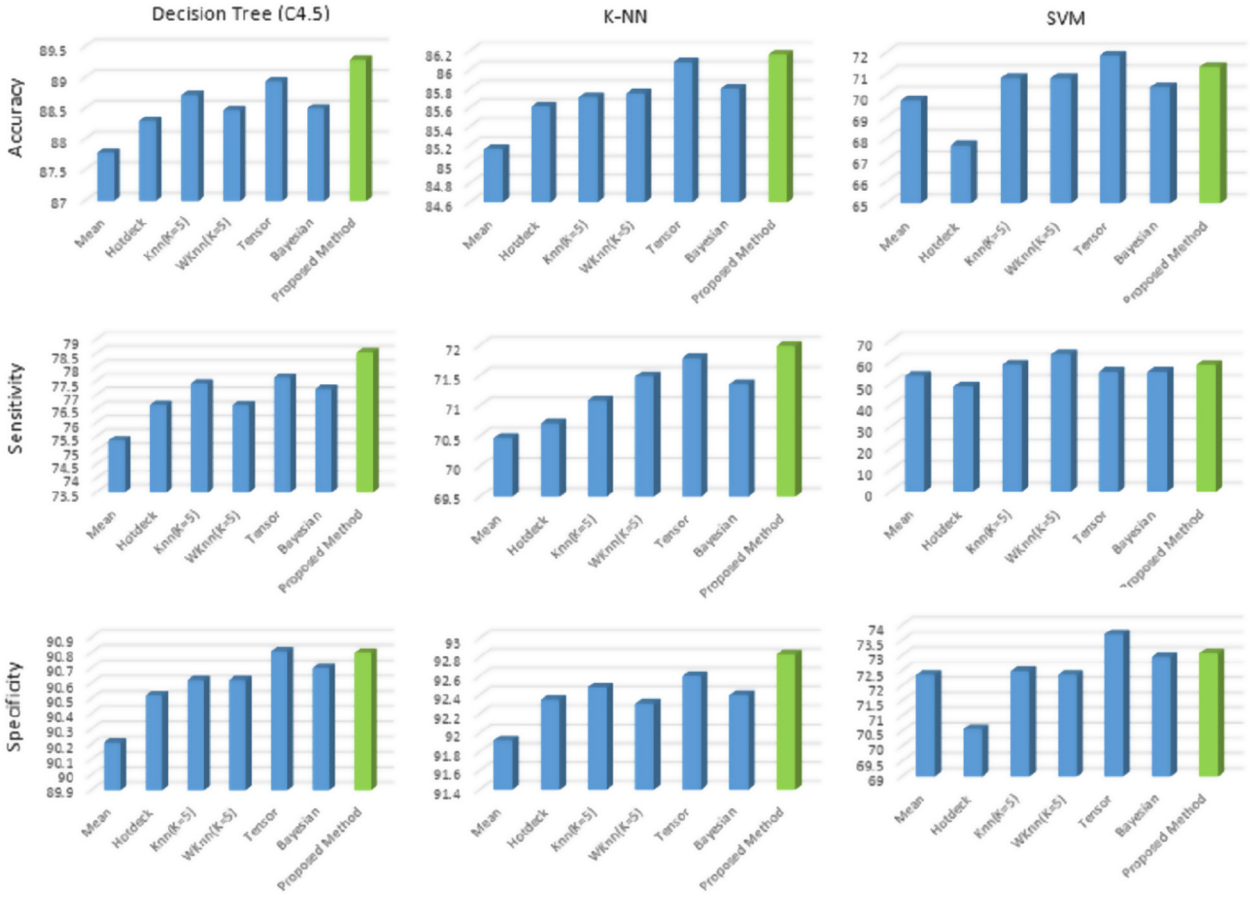

Fig. 5. Classification accuracy, sensitivity and specificity of DT, K-NN, and SVM models on Omid dataset for Breast cancer recurrence prediction.

Please cite this article in press as: Vazifehdan, M., et al. A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.01.002

variables. The proposed method for replacing missing data has also been evaluated using several different imputation methods including mean, Hot-deck, K-NN, Weighted K-NN, Tensor and Bayesian network using NRMSE criterion. Although both tensor and Bayesian networks are capable of imputing missing values, they are more capable in continuous and discrete missing data imputation, respectively. We also used three predictive models, namely, SVM, K-NN and DT and three popular measures of accuracy, sensitivity and specificity in order to predict breast cancer recurrence. These measures are applied using 5-fold cross validation on all datasets created by imputation methods. Finally, experimental results are reported for each imputation-classifier pair.

In general, the results show that the proposed method can practically increase both quality of the data and prediction quality, and it is more useful and efficient than all other compared techniques for subsequent classification. Unfortunately, our method suffers from some limitations, such as the lack of suitable prediction by SVM with RBF kernel and more computational overhead than all other methods due to checking convergence. However, we sought to obtain the following objectives: 1) the improvement of breast cancer recurrence prediction 2) proposing an approach for imputing missing data 3) paying attention to the dependency between attributes and type of incomplete attribute. Future works can focus on noise elimination of the data, feature selection and the use of other classification models for enhancing prediction accuracy. In addition, although Bayesian network and Tensor factorization are selected due to their effectiveness in previous reports, one can evaluate other missing value estimation approaches (i.e. deep neural networks or clustering-based approaches) in the proposed hybrid iterative framework.

Acknowledgements

The authors would like to express their gratefulness to Omid Oncology and Treatment Center of Mashhad for providing women breast cancer data.

References

Acar, E., 2009. Unsupervised multiway data analysis: a literature survey. IEEE Trans. Knowl. Data Eng. 21 (1), 6–20.
Acar E. et. al, "Scalable Tensor Factorizations with Missing Data," October, no. October, pp. 701–712, 2009.
Aydilek, I.B., Arslan, A., 2013. A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Inf. Sci. (Ny) 233, 25–35.
Bader B.W., Kolda T.G., and others, "MATLAB Tensor Toolbox Version 2.6." Feb-2015.
Batista, G.E., Monard, M.C., 2003. An analysis of four missing data treatment methods for supervised learning. Appl. Artif. Intell. 17 (5–6), 519–533.
Bishop C.M., Pattern Recognition and Machine Learning, vol. 53, no. 9. 2013.
Blake C., Merz C.J., "Repository of machine learning databases," 1998.
Buyse, M. et al., 2006. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J. Natl. Cancer Inst. 98 (17), 1183–1192.
Cai Z., Heydari M., Lin G., "Missing value imputation," vol. 4, no. 5, pp. 935–957, 2006.
Choi, S., Jiang, Z., 2010. Cardiac sound murmurs classification with autoregressive spectral analysis and multi-support vector machine technique. Comput. Biol. Med. 40 (1), 8–20.
Clark, P., Niblett, T., 1989. The {CN}2 rule induction algorithm. Mach. Learn. 3 (4), 261–284.
Cohen, W.W., 1995. Fast effective rule induction. Proceedings of the twelfth international conference on machine learning, pp. 115–123.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.
Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13 (1), 21–27.
Dauwels, J. et al., 2012. Tensor Factorizat. Missing Data Imput., 2109–2112
De Campos C.P. "Properties of Bayesian Dirichlet Scores to Learn Bayesian Network Structures," pp. 431–436.
Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc Ser B 39 (1), 1–38.
Dunlavy D.M., Kolda T.G., Acar E., "Poblano v1.0: A Matlab Toolbox for Gradient-Based Optimization," Mar. 2010.
Farhangfar, A., Kurgan, L., Dy, J., 2008. Impact of imputation of missing values on classification error for discrete data. Pattern Recognit. 41 (12), 3692–3705.
Franzin, A., Sambo, F., Di Camillo, B., 2017. bnstruct: an R package for Bayesian Network structure learning in the presence of missing data. Bioinformatics 33 (8), 1250–1252.
García-Laencina, P.J. et al., 2015. Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. Comput. Biol. Med. 59, 125–133.
Hitchcock, F.L., 1927. The expression of a tensor or a polyadic as a sum of products. J. Math. Phys. 6 (1), 164–189.
Jaakkola T., Meila M. "Learning Bayesian Network Structure using LP Relaxations," vol. 9, pp. 358–365, 2010.
Jerez, J.M. et al., 2010. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif. Intell. Med. 50 (2), 105–115.
Jerez-Aragonés, J.M., Gómez-Ruiz, J.A., Ramos-Jiménez, G., Muñoz-Pérez, J., Alba-Conejo, E., 2003. A combined neural network and decision trees model for prognosis of breast cancer relapse. Artif. Intell. Med. 27 (1), 45–63.
Kim, W. et al., 2012. Development of novel breast cancer recurrence prediction model using support vector machine. J. Breast Cancer 15 (2), 230–238.
Kohonen, T., 1995. Self organizing maps. Springer Ser. Inf. Sci. 30, 521.
Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data.
Malarvizhi M. R., Thanamani A.S. "K-Nearest Neighbor in Missing Data Imputation," vol. 5, no. 1, pp. 5–7, 2012.
Ng A.Y., "Feature selection, L1 vs. L2 regularization, and rotational invariance," Twenty-first Int. Conf. Mach. Learn. - ICML '04, p. 78, 2004.
Purwar, A., Singh, S.K., 2015. Hybrid prediction model with missing value imputation for medical data. Expert Syst. Appl. 42 (13), 5621–5631.
Rancoita, P.M.V. et al., 2014. Bayesian network data imputation with application to survival tree analysis. Comput. Stat. Data Anal. 93, 373–387.
Sharfian, A. et al., 2015. Burden of breast cancer in iranian women is increasing. Asian Pacific J. Cancer Prevent. 16, 5049–5052.
Silander T., Myllym P. "A Simple Approach for Finding the Globally Optimal Bayesian Network Structure," 2004.
SolaroEmail, N., Barbiero, A., ManziPier, G., Ferrari, A., 2017. A sequential distance-based approach for imputing missing data: forward Imputation. Adv. Data Analysis Classif. 11 (2), 395–414.
Sun, Y., Urquidi, V., Goodison, S., 2010. Derivation of molecular signatures for breast cancer recurrence prediction using a two-way validation approach. Breast Cancer Res. Treat. 119 (3), 593–599.
Troyanskaya, O. et al., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17 (6), 520–525.
Tutz, G., Ramzan, S., 2015. Improved methods for the imputation of missing data by nearest neighbor methods. Comput. Stat. Data Anal. 90, 84–99.
Van't Veer L.J. et al., "Gene expression profiling predicts clinical outcome of breast cancer," vol. 415, no. 345, 2002.
Vapnik, V., Golowich, S.E., Smola, A., 1996. Support vector method for function approximation, regression estimation, and signal processing. Annu. Conf. Neural Inf. Process. Syst., 281–287
Vidyasagar, M., 2017. Machine learning methods in computational cancer biology. Annu. Rev. Control, 1–21.
Wang, K.J. et al., 2014. A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients. Appl. Soft Comput. J. 20, 15–24.
Wang, H., Zhang, Q., Yuan, J., 2017. semantically enhanced medical information retrieval system: a tensor factorization based approach. IEEE Access 5, 7584–7593.
Wilin, A., 2009. Gene Select. Cancer Classif., 389–422
Yang, Fan et al., 2017. LFTF: a framework for efficient tensor analytics at scale. Proc. VLDB Endowment 10 (7), 745–756.
Zheng, B., Yoon, S.W., Lam, S.S., 2014. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst. Appl. 41 (4 PART 1), 1476–1482.

Please cite this article in press as: Vazifehdan, M., et al. A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.01.002