Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values

Pedro J. García-Laencina a, Pedro Henriques Abreu b,c,* , Miguel Henriques Abreu d, Noémia Afonso d

a Centro Universitario de la Defensa de San Javier (University Centre of Defence at the Spanish Air Force Academy), MDE-UPCT, Calle Coronel Lopez Peña, s/n, 30720 Santiago de la Ribera, Murcia, Spain
b Department of Informatics Engineering, Faculty of Sciences and Technology, University of Coimbra, Pólo II, Pinhal de Marrocos, 3030-290 Coimbra, Portugal
c Centre for Informatics and Systems, University of Coimbra, Pólo II, Pinhal de Marrocos, 3030-290 Coimbra, Portugal
d Portuguese Institute of Oncology of Porto, Rua Dr. Antonio Bernardino de Almeida, 4200-072 Porto, Portugal

Abstract

Breast cancer is the most frequently diagnosed cancer in women. Using historical patient information stored in clinical datasets, data mining and machine learning approaches can be applied to predict the survival of breast cancer patients. A common drawback is the absence of information, i.e., missing data, in certain clinical trials. However, most standard prediction methods are not able to handle incomplete samples and, then, missing data imputation is a widely applied approach for solving this inconvenience. Therefore, and taking into account the characteristics of each breast cancer dataset, it is required to perform a detailed analysis to determine the most appropriate imputation and prediction methods in each clinical environment. This research work analyzes a real breast cancer dataset from Institute Portuguese of Oncology of Porto with a high percentage of unknown categorical information (most clinical data of the patients are incomplete), which is a challenge in terms of complexity. Four scenarios are evaluated: (I) 5-year survival prediction without imputation and 5-year survival prediction from cleaned dataset with (II) Mode imputation, (III) Expectation-Maximization imputation and (IV) K-Nearest Neighbors imputation. Prediction models for breast cancer survivability are constructed using four different methods: K-Nearest Neighbors, Classification Trees, Logistic Regression and Support Vector Machines. Experiments are performed in a nested ten-fold cross-validation procedure and, according to the obtained results, the best results are provided by the K-Nearest Neighbors algorithm: more than 81% of accuracy and more than 0.78 of area under the Receiver Operator Characteristic curve, which constitutes very good results in this complex scenario.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Today, cancer is the main cause of death worldwide [1]. Based on Siegel [1], in 2014 breast cancer will be the most prevalent cancer in women. Despite the most recent therapeutic developments, it will remain the first cause of death for cancer in female gender worldwide, presenting more than two hundred and thirty thousand new cases and provoking more than 15% of the estimated deaths.

Nowadays, cancer patient treatments must be based on the best evidence gathered from randomized clinical trials. The setting of the disease probably influences the clinical endpoints but quality of life must be present transversally, not only in the palliative treatments but also in the adjuvant setting, where the patients could live many years with secondary toxicities. Oncologists have to not only learn with this evidence but also understand that in the majority of situations, trials fit patients (with good performance status and no organ dysfunctions) are not real world ones, and some adjustments are needed when the new drugs become routinely available. Clinical records are an important issue to assure quality of care for centers self-evaluation, and to analyze the real impact of new treatment options in real patients. Analysis of data from the same center reflects the homogeneity of their approaches, based on local protocols, and predicting survival using machine learning techniques constitutes an important tool to identify patients subgroups for recurrence and/or death and to adjust local approach to improve outcomes. Consequently, over the years, many research studies have emerged trying to predict the survival of patients using data mining and machine
learning techniques [2–4]. In spite of the fact that these studies presented high levels of prediction accuracy, they presented some important issues: some of them are based on simulated values or universal datasets (e.g., SEER1), which cannot be linear translated for the realities of all countries. Due to that fact, a national database or in the absence a database for a specialized cancer center (IPO Porto is the biggest Portuguese cancer center) may constitute an important tool to express a country reality. Also, they based their knowledge on complete datasets (without missing values) which is far from the reality of clinical databases.

Missing Data (MD) can result from a huge variety of events. In summarized form, MD can be produced at random or not at random (to obtain a complete description about MD, please consult Section 3). Throughout the years, many authors have tried to address MD in women breast cancer context [5–8]. However, these works presented some issues concerning a real clinical environment: very low percentage of MD and, also, the nature of the variables that compose the datasets (in most cases, continuous variables).

In this research work, a five-year survival prediction approach for an incomplete breast cancer dataset from IPO (Institute Portuguese of Oncology of Porto) with particular characteristics is proposed (a full description of the dataset will be illustrated in Section 2). These characteristics consist of a high percentage of missing values (18%): from the sixteen variables, only two features are complete and three features have more than 40% of MD. Moreover, the majority of cases (patient information) are incomplete (only 3% of the samples are completely observed) and more than 40% of the samples have more than three missing values. To the best of our knowledge, a study with this kind of characteristics has never been proposed. To achieve that goal, four scenarios were evaluated: Survival prediction without imputation (K-Nearest Neighbors (KNN) and Classification Trees (CT)), Survival prediction from cleaned dataset with Mode imputation, Survival prediction from cleaned dataset with Expectation-Maximization (EM) imputation and Survival prediction from cleaned dataset with KNN imputation. In addition, for all these scenarios, and due to the high rate of MD in three variables, the authors also consider reducing the dataset by discarding those variables in order to measure their impact on the performance results. With respect to survival prediction, four classification algorithms are used: CT, KNN, Logistic Regression (LR) and Support Vector Machines (SVM).

The achieved results prove that, without imputation, it is impossible to create accurate models to predict survival in this type of contexts. Also, KNN algorithm proves to be the best option for this scenario presenting an accuracy of more than 81% and an area under the ROC (Receiver Operator Characteristic) curve of more than 0.78, which constitute very interesting results for this type of datasets (high percentage of unknown categorical data). In terms of application, and as it has been already mentioned, the obtained models will be used to assure quality of care for Centers self-evaluation, and to analyze the real impact of new treatment options in real patients. The analysis of the data can reflect the homogeneity of their approaches, based on local protocols, and predicting survival using machine learning techniques constitutes an important tool to identify patient subgroups for recurrence and/or death and to adjust local approach to improve outcomes.

The remainder of this paper is organized as follows: Section 2 presents a brief characterization of the used breast cancer dataset. Section 3 outlines the methodological steps followed in this project and also the algorithms applied to deal with missing data and Section 4 reports the collected results. Finally, in Section 5 some conclusions are drawn regarding the achieved results as well as some discussion about the performed work and other works presented in the literature.

2. The IPO breast cancer dataset

Four medical doctors constructed the dataset that was used in this research work composed of 399 women breast cancer patients for the same oncological center (IPO). Each patient was characterized by an input vector2 of 16 variables including age, tumor site and topography, contralateral breast involvement, histological type, degree of differentiation, variables included in TNM classification (T: tumor size, N: number of nodes involved, M: number of metastasis), tumor stage (according to [9]), expression of hormonal receptors, expression of HER2 and treatment type (including surgery type, chemotherapy regimen, hormonotherapy type, if applied). The variables names as well as their type and percentage of MD are illustrated in Table 1. It should be noted that these 16 relevant variables have been previously selected by the IPO medical staff according to international guidelines, professional experience, knowledge, previous decisions and observed outcomes. Then, and following the objectives of the IPO study, the goal was to construct survival prediction models based on these significant variables.

As it is easy to analyze, almost all the variables that compose the dataset are categorical and present MD (the percentage varies between 0 and almost 81% with an average of 17.99% and a standard deviation of 21.67%). Only three variables presented completed values (Age, Topography and Contralateral Breast Involvement), HER2 having presented the highest percentage of MD (almost 82%). Besides, it is important to remark that the categories in each discrete variable are not sparsely populated and, due to this, a detailed histogram analysis for each variable was not included.

Another possible analysis is to check how many imputation values are needed to complete the database. The majority of the research works normally presented a higher percentage of complete data (e.g., normally more than 50% of the total cases) and needed to impute less than three values per patient [6]. In this research project and proving once again the complexity and the novelty of this approach, only 3% of the patients present complete data (12 of 399) and more than 40% of the patients have more than three missing values. The relation between the number of missing values and the number (absolute) of patients is expressed in Fig. 1. The survival target variable is encoded as a binary variable with values 0 and 1, which, respectively, means that a patient did not survive or survived. In the dataset under study, 117 and 282 cases belong to class 0 and 1, respectively. This work is focused on the 5-year survivability prediction for breast cancer.

3. Methods

This section describes all the MD imputation methods applied in our breast cancer dataset and the decision models that were used to predict survival. The following subsections are structured as follows. First, we introduce some notions on MD in order to understand how it has been addressed in our survival problem. Then, the three imputation methods applied for incomplete discrete variables are explained: Mode imputation (Mimp), EM imputation (EMimp) and KNN imputation (KNNimp). Also, the four decision approaches applied in this work were described: KNN classification (KNNclas), Classification Trees (CTclas), Logistic

1 For more information, consult http://seer.cancer.gov/data/

2 In this paper, the terms input vector, case, sample, and instance are used as synonyms.
3.1. Notions on handling missing data

The appropriate way to handle incomplete input data depends in most cases upon how input variables became missing [10–12]. Little and Rubin [13] define three unique types of MD mechanisms, which have been widely accepted in the literature. Data can be missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). First, in the MCAR mechanism, the missingness is independent of the incomplete variables and any other input variables of the dataset. Second, the missingness in a MAR situation is independent of the incomplete variables but the MD can depend on other observed variables. Third, when the MD is MNAR, the missingness is non-random and it depends on the missing variables. With respect to our breast cancer dataset, the MD mechanism can be considered as MAR since the probability that a value is missing only depends on the availability of the clinical information during the data collection process in the IPO. Besides, after a clinical staff validation, our procedure has been accepted as a valid one for the breast dataset.

Intuitively, the easiest way to handle MD is simply discarding them and, after that, applying a standard complete data analysis [11,13]. This treatment can only be applied when the incomplete portion of the dataset is very reduced in comparison with the sample size [11]. To discard MD also entails that the decision model will be trained under the complete data assumption. Due to this fact, this procedure is valid only if there is not any incomplete pattern during the operation mode of the decision model [12], which is very unlikely. Although the elimination of MD is a widely used approach in practice [3], the previously cited situations are not common.

On the one hand, MD may occur on several combinations of the input variables and, in this context, the incomplete cases represent a significant portion of the whole dataset [14]. As a real example, in our breast cancer dataset, there are only 3% of complete cases and the elimination of those with MD will invalidate any kind of analysis in the dataset. On the other hand, if the MD phenomenon appears in the training phase of the survival model, it is very likely that the new patient instance may contain also MD and, in this case, the trained model would not be able to provide survival predictions.

According to the literature [6,12,14], more appropriate treatments for dealing with MD are the following ones:

- **Imputation based approaches**: Using the available complete data, the MD are estimated and filled by plausible values. After that, the classifier is designed using the imputed dataset. Therefore, two different and consecutive stages could be considered: imputation and classification.
- **Avoiding explicit imputations**: In this kind of approaches, the decision methods are able to deal with unknown values, i.e., they can handle MD during the classifier design. Then, classification is performed without a previous MD imputation.

This research work applies and analyzes MD treatments for survival prediction in breast cancer patients (for a full description, see Section 4). To perform such analyses, representative methods for imputation and classification were used. Regarding the imputation procedures (Mimp, EMimp and KNNimp) and, in addition to their acceptance in other related works [6], it should be remarked that they have been chosen due to their well-known capabilities to impute MD in discrete variables (not all imputation methods are suitable to handle discrete data). Meanwhile, the four classification methods applied in this work (KNNclas, CTClas, LRclas and SVMclas) have also shown their usefulness in other breast cancer studies [3,15–19]. Note that two of them, KNNclas and CTClas, can directly classify incomplete input cases without a previous MD imputation, as we explain in Section 3.3.

### 3.2. Imputation methods

Over the years, past works have tried to perform a MD estimation used a mean/mode imputation (Mimp) approach. For any continuous variable, missing values are imputed using the mean of the observed values; meanwhile, for discrete variables, the mode (the most frequently observed) is used. This method presents obvious disadvantages like the under representation of the variability in the dataset and, also, it completely ignores the relationship between the different variables of the dataset.

#### 3.2.1. EM imputation

A more sophisticated approach is using mixture models trained with the Expectation-Maximization (EM) algorithm [20]. A mixture model provides a general semi-parametric model for the joint distribution of all variables by mixing together several base distribution functions in a convex combination manner. In particular, continuous data can be modeled as a mixture of Gaussians; and for discrete data, it can be modeled as a mixture of Bernoulli densities.
(binary variables) or by a mixture of multinomial densities (categorical variables) [21]. When the dataset is composed of mixed continuous, binary and categorical variables (as in our dataset), the mixture model assumes a joint density with mixed components of the three variable types. The EM algorithm has been successfully applied for training mixture models. It is an efficient iterative procedure to compute the Maximum Likelihood (ML) estimate with missing data [21]. Each iteration of the EM algorithm consists of two steps [20]. The Expectation step (E-step) computes the log likelihood of the data, and the Maximization step (M-step) finds the parameters that maximize this likelihood. Missing values in the input features can be naturally handled in the EM algorithm, by computing the expectation of the sufficient statistics for the E-step over the incomplete values as well as the mixture components. Thus, in the E-step, MD are estimated given the observed data and the current estimate of the model parameters. This is achieved by means of the conditional expectation and, then, the estimate of the MD from the E-step is used instead of the actual MD. In the M-step, the likelihood function is maximized under the assumption that the MD are known. The iterative process of the EM algorithm is repeated until convergence. Note that a detailed description of this approach is out-of-the-scope of this work. More specific details can be found in [20–24].

3.2.2. KNN for imputation

The KNN algorithm is an instance-based learning approach [22,25,26]: it stores all the training instances and, when a new input vector is encountered, the prediction is performed by considering its K closest training instances according to a given distance metric. Then, there are two important issues to be defined: the parameter K (number of nearest neighbors) and the distance metric. Considering several possible values for K, its optimal value is usually chosen by cross-validation. With respect to the similarity measurement (distance metric), it has to deal with mixed (continuous and discrete) and incomplete feature values. A recommended distance function is the Heterogeneous Euclidean-Overlap Metric (HEOM) [27]. Considering two input vectors, \( x_A \) and \( x_B \), the HEOM distance can be calculated by

\[
d(O, \text{HEOM}) = \sqrt{\sum_{j=1}^{n} d_j(x_A, x_B)^2},
\]

with \( d_j(x_A, x_B) \) being the distance between the two cases on its \( j \)-th attribute, where

\[
d_j(x_A, x_B) = \begin{cases} 
1 & \text{if } x_j \text{ is missing in } (x_A) \text{ or } (x_B), \\
\frac{d_O(x_A, x_B)}{d_O(x_A, x_B) + d_O(x_B, x_B)} & \text{if } x_j \text{ is a discrete variable}, \\
\frac{d_N(x_A, x_B)}{d_N(x_A, x_B) + d_N(x_B, x_B)} & \text{if } x_j \text{ is a continuous variable}.
\end{cases}
\]

In Eq. (2), it is considered that the distance varies from 0 to 1 (the maximal distance value). If either of the input values is missing in the \( j \)-th variable, its distance is 1. If both input values are available, HEOM uses the overlap metric, \( d_0 \), for categorical attributes and the normalized Euclidean distance, \( d_N \), for continuous attributes:

\[
d_O(x_A, x_B) = \begin{cases} 
0 & \text{if } x_A = x_B, \\
1 & \text{otherwise;}
\end{cases}
\]

\[
d_N(x_A, x_B) = \frac{|x_A - x_B|}{\max(x_A) - \min(x_A)}.
\]

Given an incomplete input vector \( x \), and for each missing value in \( x \), the KNN imputation method (KNNimp) [27–29] finds the corresponding set of \( K \) closest training cases with observed values in the incomplete feature to be imputed. Then, a different set of \( K \) Nearest Neighbors is obtained for each missing value. Suppose that the \( j \)-th feature of \( x \) is missing and \( V = \{v_k\}_{k=1}^n \) denotes its set of \( K \) Nearest Neighbors, the unknown value \( x_j \) is estimated by a local approximation using the \( j \)-th feature values of \( V \). Depending on the nature of the \( j \)-th variable, the imputed value is given by the mean (continuous data) or the mode (categorical data) of \( |V|_j = 1 \). Particularly, we use the weighted scheme for KNNimp [29], where the contribution of each \( v_j \) in the mean/mode computation is properly weighted according to its distance to \( x \), in order to give greater weights to closer neighbors. It is based on the well-known Dudani’s rule [30], where the nearest neighbor receives a weight of 1, the furthest neighbor a weight of 0, and the remaining neighbors are scaled linearly between 0 and 1. Finally, and according to the literature, it has been shown that KNNimp provides a robust approach for imputing MD in several real-world problems [27–29].

3.3. Decision methods

In the next subsection, the decision methods applied in this research work (KNNclas, CTClas, LRClas and SVMclas) will be briefly described.

3.3.1. KNN classifier

When the KNN method is used to classify an unlabeled pattern \( x \), the distances from \( x \) to the labeled instances are computed using the HEOM, its \( K \) Nearest Neighbors are found, and their class labels are then used to assign a class to \( x \). In this work, the class is determined using the weighted voting scheme proposed by Dudani [30], where closer instances have greater weights. Due to the fact that the distance metric can deal with MD, it should be remarked that KNNclass can directly perform decisions with incomplete patterns.

3.3.2. Classification Trees

Classification Trees (CT) are defined by recursively partitioning the input space from a root node to multiple branch nodes [26,31]. With the exception of the root node, all other nodes in a CT have exactly one incoming edge. A node with outgoing edges is known as internal node. All other nodes are called leaves (or decision nodes) and each leaf is assigned to a class. In a CT, each internal node divides the instance space into two or more sub-spaces according to a certain condition test of the input features. Typically, a single attribute is considered in each internal node, so that the input space is partitioned according to the attribute value. For continuous variables, the condition refers to a range. An input vector is classified in the CT by sorting it from the root to a leaf, according to the results of the conditions tested along the path. CT are computationally efficient, can easily handle mixed variables (continuous and discrete) and the rules generated by them are relatively easy to interpret and understand, particularly in healthcare contexts.

One of the most popular CT approaches is the C4.5 algorithm [31,32], which uses entropy-based measures (typically, the gain ratio) as splitting criterion during the tree construction process. Another advantage of this algorithm is that a CT can internally work with MD. This algorithm uses a probabilistic approach to handle missing values [33], which consists of two main steps. The first step is the penalizing of gain ratios for incomplete features: only cases with known values are taken into account during the calculation of gain ratio, which is also corrected by a factor that represents the probability with which a given attribute is known. The second step is the partitioning of cases, which assigns a weight according to the probability of each possible feature value. For incomplete features, these probabilities are estimated based on the observed values at a given node. For example, suppose that \( x_j \) is a binary feature with MD, if a node contains 70% of known
values equal to 0 and the remaining ones equal to 1, it means that 0.7 and 0.3 are, respectively, the probability of a missing value in \( x_i \) that would be equal to 0 and 1. In this example, when a missing value appears in \( x_i \), a weight equal to 0.7 is now distributed down the branch with \( x_i = 0 \) and a weight of 0.3 down the other tree branch. Note that when the feature value is known, its weight is established equal to one. During the testing phase, if a test case is incomplete, it explores all available branches and decides the class label by the most probabilistic value.

### 3.3.3. Logistic Regression

Logistic Regression (LR) is a generalization of a linear regression for binary classification [22]. Given a set of features (continuous, discrete or a mixture of both types) and a binary target, LR computes a linear combination of the inputs and it is passed through the sigmoid (or logistic) function. Due to this, the LR output continuously ranges from 0 to 1, and it can be interpreted as a class probability. The parameters of a LR model are obtained using an optimization method based on Maximum Likelihood Estimation (MLE). In this work, the Iteratively Reweighted Least Squares (IRLS) method has been applied, which is the most common approach due to its easier implementation and competitive results.

### 3.3.4. Support Vector Machines

Support Vector Machines (SVM) is a kernel-based algorithm based on statistical learning theory [34–36], which is originally formulated for binary classification problems. A kernel is a function that transforms the input data into a high-dimensional space [36]. In general, kernels can be linear (the dot product) and nonlinear (such as the Gaussian or the polynomial) functions. Once the kernel-based transformation is done, SVM determines the optimal hyperplane for separating the two classes in the high-dimensional space. It is done by maximizing the margin between both classes while minimizing misclassification errors. The SVM classifier follows from the solution to a convex quadratic programming problem involving inequality constraints. In order to reduce its complexity and computational requirements, the Least Squares formulation of SVM (LS-SVM) has been proposed [37], which only involves equality constraints. LS-SVM reduces the optimization problem to the resolution of a system of linear equations, which can be done in an efficient manner [37]. In this work, SVM classifiers are trained using the LS-SVM method and also, based on the obtained favorable results from the preliminary runs and its easier implementation, a linear kernel is chosen.

### 3.4. Performance Evaluation

In this work, four well-known measures for performance evaluation in survival prediction problems were employed: accuracy, sensitivity, specificity and area under ROC Curve. The first three measures are given by

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \tag{6}
\]

\[
\text{Specificity} = \frac{TN}{TN + FP} \tag{7}
\]

where \( TP, TN, FP \) and \( FN \) denote true positives, true negatives, false positives and false negatives, respectively. These four values are summarized in a 2 × 2 confusion matrix (see Table 2).

In addition to Eqs. (5)–(7), the area under the ROC curve (AUC) is an effective way to measure the overall performance of a survival prediction model. This measure represents a trade-off between sensitivity and specificity. AUC takes values from 0 to 1, where 0 indicates a perfectly inaccurate model and 1 reflects a perfectly accurate model. In general, a value of 0.5 for AUC is considered as the lower bound. The higher the AUC is, the greater the overall performance of the model to correctly predict survival.

### 3.4.1. Cross validation

In order to make an accurate performance evaluation of the different approaches, this work uses the nested stratified 10-fold Cross-Validation (CV) procedure [22]. Following this, our dataset is randomly divided into ten disjoint equal-sized folds by stratified sampling, which ensures that each class is present in the same proportion in every fold. In the standard 10-fold CV procedure, nine folds are used for training the model and the excluded fold is used to test the trained model. This process entails ten iterations in order to ensure that each fold has been used once for testing purposes. The obtained results from all iterations are accumulated as the final result from the CV process in the method under evaluation. In a previous related work [3], Delen et al. used the standard 10-fold CV procedure to evaluate the performance of three decision methods (neural network, decision tree and LR) for breast cancer survival prediction. However, at this point, it is important to state that most methods require to choose some parameter values for designing the model (such as the hidden layer size of a neural network or the number of neighbors in the KNN algorithm) and this selection of parameters has to be done independent of the training and test sets. Particularly, Delen et al. [3] did not consider this fact during the CV process and reported the obtained results for the best model according to the test set, which leads to overestimating the performance. Model selection (i.e., to choose the set of model parameters) cannot be performed using the test set. For this reason, an independent validation set has to be used for model selection. In this work, the nested version of CV has been used, which comprises an outer CV for the test performance evaluation of an inner CV that trains and validates the model. Therefore, in each of the ten iterations of the outer CV process, one fold is chosen for testing and the remaining nine folds are considered in the inner CV process, which at the same time comprises nine iterations with one fold as validation set and eight folds as training set. Fig. 2 depicts a scheme of the nested 10-fold CV procedure. Note that, for each parameter setting of a model, ninety combinations of folds are evaluated by considering all folds during training, validation and test phases. At the end, we choose that set of model parameters which gives better average performance results in validation.

It should be remarked that, in order to compensate for the imbalanced class distributions during the training of the survival models, random under-sampling is performed inside the inner CV process. Other related works in breast cancer survival have also applied random under-sampling [38]. In this work, for each fold of the CV process, a subset of majority class examples is randomly chosen to provide the same proportion of balance during the training stage of the classifiers. Particularly, after this sampling procedure, each fold is composed of twelve samples for each class. As the under-sampling may be inaccurate due to the randomness, it has been repeated ten times and, then, for each iteration of the inner CV process, the decision model is trained ten times. Therefore, in the whole nested 10-fold CV process, a decision model
with a set of parameters is trained nine hundred times. Finally, it is important to say that all methods are evaluated with the same samples in each fold, which ensures that fair performance comparisons are carried out.

The nested 10-fold CV procedure is used to select both the non-trainable parameters of the imputation and the decision methods. The explored values are in the following intervals:

- **KNNimp and KNNclas**: In each approach, for the number of nearest neighbors: \[1 - 40\].
- **EMimp**: For the number of mixture components: \[2 - 20\].
- **SVMclas**: For the regularization parameter: \([10^{-3} - 10^{+1}]\).

The parameters have been selected by considering the obtained performance for survival prediction in the validation sets, particularly, according to the AUC. Note that, in some imputation-based approaches for survival prediction, a 2-D grid search is required to both tune the parameters of the imputation method (e.g., the number of components in EMimp) and the decision method (e.g., the value of \(K\) in KNNclas). The paired-sample t-test at the 5% significance level has been used to compare the performance results (e.g., AUC values) obtained with the applied approaches.

4. Experiments

One of the main goals of the experiments conducted in this work is to evaluate how different MD treatments affect the survival prediction performance in our breast cancer dataset. Particularly, and as we have already mentioned in Section 3.1, two main MD treatments are considered: first, modeling our survival prediction problem by decision methods which are able to handle MD without imputation; second, to apply MD imputation methods in our incomplete dataset and, then, to design survival prediction models using the resulting dataset, with complete and imputed instances. Next, the experimental results are shown and discussed.

4.1. Survival prediction without MD imputation

In a first set of experiments, we have evaluated the survival prediction performance of two well-known methods, KNNclas and CTclas, without a previous MD imputation. As it has been described in Section 3.3, both methods can directly handle unknown input features for predicting survival. Nevertheless, according to our experiments without imputation, both classifiers completely fail because they provide the same class as output for all patient data. In particular, for small values of the \(K\) parameter (less than 10), KNNclas predicts that the patients would not survive (its output is always 0); and for higher values of \(K\) (more than 10), its output is equal to 1 (patients would be alive). In the case of CTclas, the predicted output class is 0 for all patients. The incorrect behavior of both decision methods in this survival problem is produced by the high percentage of unknown feature values, most samples being incomplete. Anyway, and due to the fact that there are three variables with more than 40% of MD (HER2, Chemotherapy regime and Hormonotherapy type), we have repeated our analysis discarding them. However, the same
results are obtained: both decision methods cannot model the breast cancer survival task. For all these reasons, MD imputation must be considered as a solution in our survival problem.

4.2. Survival prediction with MD imputation

In this work, imputation-based approaches are evaluated by considering three methods for MD imputation (Mimp, EMimp and KNNimp) and four methods for survival prediction (KNNclas, CTclas, LRclas and SVMclas). For each combination of imputation and classification methods, the corresponding set of model parameters is chosen according to the validation results in the nested CV process (as it has been explained in Section 3.4.1). In particular, Table 3 shows the obtained test results in terms of four measures of survival prediction: AUC, Accuracy, Sensitivity and Specificity. As a first comment, and in contrast to the previous experiments, it is possible to see from this table that a previous MD imputation makes feasible the modeling of our survival problem, even with a simple MD estimation based on the mode statistic (Mimp). Nevertheless, it is needed to measure the influence of different MD imputation techniques in the survival prediction performance. As it is expected, EMimp and KNNimp enhance the obtained performance with Mimp, independent of the decision method. For a given decision method, and in terms of AUC, the obtained results with KNNimp are statistically significantly better than from EMimp for most classification methods, except with KNNclas where the differences in the obtained AUC values (0.7810 using EMimp and 0.7845 using KNNimp) are not significant for a significance value of 5%. With respect to the decision methods, we have found that the differences between the AUCs provided by SVMclas and LRclas are not statistically significant and, also, that CTclas gives the worst prediction results. Besides that, and according to its low specificity results, CTclas tends to be biased toward the majority class (less than 55% in the best case for Specificity). Although with less effect, this undesirable behavior also appears in LRclas (62.23%) and SVMclas (63.55%). The most robust and accurate approach is the KNNclas method: it provides better AUCs than the other three classifiers for the same MD imputation technique (up to 0.785 with KNNimp) and, also, it is able to reduce the number of FP and, then, increase the number of TN (it achieves more than 70% of Specificity with EMimp and KNNimp).

Due to the high percentage of MD in HER2, Chemotherapy regime and Hormonotherapy type, we model our survival problem using the imputation-based approaches by discarding those variables [14]. By doing this, the total percentage of MD is reduced from 17.99% (with all variables) to 7.32%. The objective is to know how this elimination of features may influence the obtained performance for our survival problem. Table 4 shows the AUC values obtained by the different approaches in this scenario. As it can be seen in this table, and trying to perform a comparison with the AUC values from Table 3, those values slightly increase for all approaches but this performance enhancement is not statistically significant ($p$–value $>$ 0.05) in none of them. Besides, in this scenario, the same discussions are maintained about the capabilities of each imputation and decision method for our survival prediction problem. Therefore, discarding these variables with high percentage of MD does not clearly impact in the obtained performance. Finally, Table 5 shows the non-trainable parameters chosen by the nested CV process for the results of Table 3 (with all features) and Table 4 (discarding features HER2, Chemotherapy regime and Hormonotherapy type). For each combination of classification and imputation approaches, we show the chosen parameter values between brackets (first value is associated with the parameter for

---

Table 3

<table>
<thead>
<tr>
<th>Methods</th>
<th>Evaluation of survival prediction</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision</td>
<td>Imputation</td>
<td>AUC</td>
<td>Accuracy (%)</td>
<td>Sensitivity (%)</td>
</tr>
<tr>
<td>KNNclas</td>
<td>Mimp</td>
<td>0.7321 (4.56e – 2)</td>
<td>78.08 (3.63)</td>
<td>85.08 (3.66)</td>
</tr>
<tr>
<td></td>
<td>EMimp</td>
<td>0.7810 (2.82e – 2)</td>
<td>81.46 (2.34)</td>
<td>85.77 (2.48)</td>
</tr>
<tr>
<td></td>
<td>KNNimp</td>
<td>0.7845 (2.88e – 2)</td>
<td>81.73 (2.32)</td>
<td>86.44 (2.66)</td>
</tr>
<tr>
<td>CTclas</td>
<td>Mimp</td>
<td>0.6563 (3.22e – 2)</td>
<td>68.83 (2.70)</td>
<td>83.86 (3.14)</td>
</tr>
<tr>
<td></td>
<td>EMimp</td>
<td>0.6772 (2.14e – 2)</td>
<td>71.06 (1.77)</td>
<td>85.12 (2.33)</td>
</tr>
<tr>
<td></td>
<td>KNNimp</td>
<td>0.7012 (3.32e – 2)</td>
<td>74.27 (2.77)</td>
<td>85.79 (3.24)</td>
</tr>
<tr>
<td>LRclas</td>
<td>Mimp</td>
<td>0.7164 (2.13e – 2)</td>
<td>75.25 (2.84)</td>
<td>87.20 (3.14)</td>
</tr>
<tr>
<td></td>
<td>EMimp</td>
<td>0.7373 (2.89e – 2)</td>
<td>77.21 (2.66)</td>
<td>88.37 (3.01)</td>
</tr>
<tr>
<td></td>
<td>KNNimp</td>
<td>0.7514 (2.99e – 2)</td>
<td>79.08 (2.65)</td>
<td>88.04 (3.03)</td>
</tr>
<tr>
<td>SVMclas</td>
<td>Mimp</td>
<td>0.7168 (2.80e – 2)</td>
<td>75.39 (2.53)</td>
<td>86.84 (3.00)</td>
</tr>
<tr>
<td></td>
<td>EMimp</td>
<td>0.7348 (3.17e – 2)</td>
<td>77.30 (2.73)</td>
<td>88.38 (3.22)</td>
</tr>
<tr>
<td></td>
<td>KNNimp</td>
<td>0.7540 (3.34e – 2)</td>
<td>79.61 (2.88)</td>
<td>87.24 (3.08)</td>
</tr>
</tbody>
</table>

---

Table 4

<table>
<thead>
<tr>
<th>Methods</th>
<th>Evaluation of survival prediction</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision</td>
<td>Imputation</td>
<td>AUC</td>
<td>Accuracy (%)</td>
<td>Sensitivity (%)</td>
</tr>
<tr>
<td>KNNclas</td>
<td>Mimp</td>
<td>0.7419 (4.22e – 2)</td>
<td>78.08 (3.63)</td>
<td>85.08 (3.66)</td>
</tr>
<tr>
<td></td>
<td>EMimp</td>
<td>0.7808 (3.64e – 2)</td>
<td>81.46 (2.34)</td>
<td>85.77 (2.48)</td>
</tr>
<tr>
<td></td>
<td>KNNimp</td>
<td>0.7860 (2.92e – 2)</td>
<td>81.73 (2.32)</td>
<td>86.44 (2.66)</td>
</tr>
<tr>
<td>CTclas</td>
<td>Mimp</td>
<td>0.6642 (2.98e – 2)</td>
<td>68.83 (2.70)</td>
<td>83.86 (3.14)</td>
</tr>
<tr>
<td></td>
<td>EMimp</td>
<td>0.6909 (1.32e – 2)</td>
<td>71.06 (1.77)</td>
<td>85.12 (2.33)</td>
</tr>
<tr>
<td></td>
<td>KNNimp</td>
<td>0.7093 (2.78e – 2)</td>
<td>74.27 (2.77)</td>
<td>85.79 (3.24)</td>
</tr>
<tr>
<td>LRclas</td>
<td>Mimp</td>
<td>0.7196 (3.16e – 2)</td>
<td>75.25 (2.84)</td>
<td>87.20 (3.14)</td>
</tr>
<tr>
<td></td>
<td>EMimp</td>
<td>0.7445 (2.92e – 2)</td>
<td>77.21 (2.66)</td>
<td>88.37 (3.01)</td>
</tr>
<tr>
<td></td>
<td>KNNimp</td>
<td>0.7639 (3.00e – 2)</td>
<td>79.08 (2.65)</td>
<td>88.04 (3.03)</td>
</tr>
<tr>
<td>SVMclas</td>
<td>Mimp</td>
<td>0.7192 (3.53e – 2)</td>
<td>75.39 (2.53)</td>
<td>86.84 (3.00)</td>
</tr>
<tr>
<td></td>
<td>EMimp</td>
<td>0.7410 (2.89e – 2)</td>
<td>77.30 (2.73)</td>
<td>88.38 (3.22)</td>
</tr>
<tr>
<td></td>
<td>KNNimp</td>
<td>0.7643 (2.64e – 2)</td>
<td>79.61 (2.88)</td>
<td>87.24 (3.08)</td>
</tr>
</tbody>
</table>
classification and the second one to the parameter for imputation). In this table, we use the term n.a. (not applicable) when there is not any parameter to be chosen in the corresponding approach (for example, Mimp does not require any parameter selection). Note that the explored values for the non-trainable parameters have been previously described in Section 3.4.1.

5. Discussions

Breast cancer survival has been widely studied in order to provide robust and accurate prediction models that help us to improve the disease treatment. These models are constructed using historical patient information stored in clinical datasets and, then, they can be used for predicting the breast cancer outcome for new patient data. However, it is important to note that in order to develop this kind of studies some precautions need to be conducted. The used datasets must be representative for a specific healthcare context (encompassing an oncological center or a country region composed of two or more centers) and the patients data cannot be outdated. In healthcare contexts, a dataset can be considered as an outdated/obsolete one if, for instance, the diagnostic methods or the treatments contained in the dataset are no longer available. In this situation, research works that for example, develop prediction models cannot have any applicability in real environments.

One of the first related works was developed by Prentice and Gloecker in 1978 [39], where they applied the well-recognized Cox regression model for survival analysis with breast cancer data. This model, together with the Kaplan–Meier approach, constitutes the traditional methodologies for survival analysis. In addition to this, in the last decades, advanced methods from machine learning have been successfully applied. In 1997, Burke et al. [40] proposed a survival prediction model based on a standard neural network with a fixed architecture and they evaluated its performance on the SEER breast carcinoma dataset. Training and test sets were randomly defined using the hold-out approach and incomplete instances were directly discarded of their survival analysis. After that, in 2005, Delen et al. [3] performed a comparative analysis of three methods (neural network, decision tree and logistic regression) for 5-year breast cancer survival prediction using the SEER dataset. In this study, the performances of these methods were analyzed in terms of accuracy, sensitivity and specificity. According to its experimental results in a 10-fold CV process, the decision tree model outperformed both neural network and logistic regression. It should be remarked that Delen et al. [3] performed a complete data analysis: instances with missing data (40% of the sample size) were removed from their dataset. Besides that, they designed the prediction models considering the results in the test set, which leads to overestimating the performance. In 2006, Cruz and Wishart [15] conducted a comprehensive survey of machine learning applications in cancer prognosis problems and they suggested to consider SVM classifiers instead of the traditional neural networks. Despite its relevance in cancer datasets (more than 80% of breast cancer studies deal with MD [5]), the treatment of incomplete data was not analyzed in that survey [15].

In spite of the fact that the most commonly reported approaches in breast cancer survival are based on complete data analysis [5], recent works have used more appropriate MD treatments. Note that MD is an area which has attracted an increasing attention [14,41–43] and, then, several techniques, derived from statistics and machine learning, have been developed and applied for breast cancer datasets. Herring et al. [44] combined the use of regression models with an EM-based approach in a breast cancer dataset with 20% of incomplete data. Jerez et al. [6,45] analyzed the impact of different imputation-based approaches on the obtained survival prediction by a neural network. Both research works studied a breast cancer dataset with a low percentage of MD (less than 6%), most of its missing values being in an ordinal variable. Also, there was an enough number of complete cases (more than 50%) and there were no cases with more than three missing values. In view of their experiments [6,45], they recommended applying machine learning approaches for MD imputation due to their improvement in prediction accuracy.

In this research work, a real breast cancer dataset is analyzed by considering several MD treatments with different decision methods for survival prediction. It should be remarked that there are several aspects that make the survival prediction in this clinical dataset harder. Firstly, there is a significant percentage of missing values (18%); from the sixteen variables, only two features are complete and three features have more than 40% of MD. Secondly, most cases are incomplete (only 3% of the samples are completely observed) and more than 40% of the samples have more than three values missing. Thirdly, most variables are discrete (binary or categorical), which restricts the imputation and decision methods to be applied. Due to this, and comparing with other breast cancer datasets that we have reviewed in the literature, this problem is a challenge in terms of complexity.

As shown in our experimental results, MD imputation is required in order to provide accurate survival models: prediction methods that avoid explicit MD imputations completely fail. Then, we have applied three representative imputation techniques (Mimp, EMimp, KNNimp) and, once the incomplete cases are imputed, four decision methods (KNNclas, C Tclas, LRclas, SVMclas) are constructed to predict breast cancer survival. Performance evaluation has been done by considering four widely known measures: Accuracy, Sensitivity, Specificity and AUC. For each decision method, we have analyzed the impact of each imputation technique and, also, determined the imputation technique which entails a better improvement of the survival prediction, mainly in terms of the AUC value. The best results are obtained for the prediction model based on the KNN algorithm in which the AUC is 0.7845 ± 2.88e–2 (mean and standard deviation from a nested 10-fold CV process), the MD imputation also being done by the KNN algorithm, which constitutes very good results in this scenario. For all prediction models, the Mimp provides the worst performance. Another issue is that, although the CT approach has shown its usefulness in other related studies, its obtained results

| Table 5 | Non-trainable parameters chosen by validation for each combination of classification method – K Nearest Neighbors (KNNclas), Classification Trees (C Tclas), Logistic Regression (LRclas) and Support Vector Machines (SVMclas) – and imputation approach – Mean imputation (Mimp), Expectation–Maximization imputation (EMimp) and K–Nearest Neighbors imputation (KNNimp). The first value is associated with the parameter of the classification method and the second value to the parameter of the imputation approach. The term n.a. (not applicable) is used when there is not any parameter value to be selected. |
|---|---|---|---|---|---|---|---|---|---|
| Classification Method | KNNclas | C Tclas | LRclas | SVMclas |
| | Mimp | EMimp | KNNimp | Mimp | EMimp | KNNimp | Mimp | EMimp | KNNimp |
| All features (Table 3) | (15, n.a.) | (30, 8) | (30, 10) | (n.a., n.a.) | (n.a., 6) | (n.a., 10) | (n.a., 10) | (n.a., n.a.) | (n.a., n.a.) |
| Discarding features (Table 4) | (15, n.a.) | (20, 5) | (35, 10) | (n.a., n.a.) | (n.a., 10) | (n.a., 5) | (n.a., 10) | (n.a., 12) | (n.a., 3) |
References


Conflict of interest statement

None declared.

Acknowledgments

The first author is supported by the 2014 Santander Ibero-American Universities Programme for Young Teachers and Researchers. This work is also partially supported by iCOS project (CENTRO-07-ST24-FEDER-002003), which is co-financed by QREN, in the scope of the Mais Centro Program and FEDER.