Computers in Biology and Medicine 59 (2015) 125–133

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

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 Afonoso [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

ARTICLE INFO

Article history:
Received 20 October 2014
Accepted 9 February 2015

Keywords:
Breast cancer
5-year survival prediction
Missing data
Imputation
Discrete data

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.

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

* Corresponding author at: Department of Informatics Engineering, Faculty of Sciences and Technology, University of Coimbra, Pólo II, Pinhal de Marrocos, 3030-290 Coimbra, Portugal.
E-mail addresses: pedroj.garcia@cud.upct.es (P.J. García-Laencina), pha@dei.uc.pt (P.H. Abreu), p_abreu@sapo.pt (M.H. Abreu), noemia.afonso@netcabo.pt (N. Afonoso).

http://dx.doi.org/10.1016/j.compbiomed.2015.02.006
0010-4825/

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., SEER[1]), 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 clinic 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 this 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 vector[2] 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.