Journal of Biomedical Informatics 58 (2015) 49–59

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients

Miriam Seoane Santos [a,b], Pedro Henriques Abreu [a,b,*], Pedro J. García-Laencina [c], Adélia Simão [d], Armando Carvalho [d]

[a] Centre for Informatics and Systems, University of Coimbra, Pólo II, Pinhal de Marrocos, 3030-290 Coimbra, Portugal
[b] Department of Informatics Engineering, Faculty of Sciences and Technology, University of Coimbra, Pólo II, Pinhal de Marrocos, 3030-290 Coimbra, Portugal
[c] Centro Universitario de la Defensa de San Javier (University Centre of Defence at the Spanish Air Force Academy), MDE-UPCT, Calle Coronel López Peña, s/n, 30720 Santiago de la Ribera, Murcia, Spain
[d] Internal Medicine Service, Hospital and University Centre of Coimbra, EPE, Rua Fonseca Pinto, 3000-075 Coimbra, Portugal

ARTICLE INFO

Article history:
Received 16 February 2015
Revised 13 August 2015
Accepted 20 September 2015
Available online 28 September 2015

Keywords:
Hepatocellular Carcinoma (HCC)
Clustering
K-means
Oversampling
SMOTE
Survival prediction

ABSTRACT

Liver cancer is the sixth most frequently diagnosed cancer and, particularly, Hepatocellular Carcinoma (HCC) represents more than 90% of primary liver cancers. Clinicians assess each patient's treatment on the basis of evidence-based medicine, which may not always apply to a specific patient, given the biological variability among individuals. Over the years, and for the particular case of Hepatocellular Carcinoma, some research studies have been developing strategies for assisting clinicians in decision making, using computational methods (e.g. machine learning techniques) to extract knowledge from the clinical data. However, these studies have some limitations that have not yet been addressed: some do not focus entirely on Hepatocellular Carcinoma patients, others have strict application boundaries, and none considers the heterogeneity between patients nor the presence of missing data, a common drawback in healthcare contexts. In this work, a real complex Hepatocellular Carcinoma database composed of heterogeneous clinical features is studied. We propose a new cluster-based oversampling approach robust to small and imbalanced datasets, which accounts for the heterogeneity of patients with Hepatocellular Carcinoma. The preprocessing procedures of this work are based on data imputation considering appropriate distance metrics for both heterogeneous and missing data (HEOM) and clustering studies to assess the underlying patient groups in the studied dataset (K-means). The final approach is applied in order to diminish the impact of underlying patient profiles with reduced sizes on survival prediction. It is based on K-means clustering and the SMOTE algorithm to build a representative dataset and use it as training example for different machine learning procedures (logistic regression and neural networks). The results are evaluated in terms of survival prediction and compared across baseline approaches that do not consider clustering and/or oversampling using the Friedman rank test. Our proposed methodology coupled with neural networks outperformed all others, suggesting an improvement over the classical approaches currently used in Hepatocellular Carcinoma prediction models.

1. Introduction

For the past few years, we have been witnessing an exponential growth of cancer incidence and related deaths worldwide. Solely in 2012, the World Health Organization (WHO) reported about 14.1 millions of new cancer cases and 8.2 millions of deaths [1]. Liver cancer was the sixth most frequently diagnosed cancer and the second cause of cancer-related deaths worldwide, accounting for 9.1% of all deaths [1,2]. Hepatocellular Carcinoma (HCC) represents more than 90% of primary liver cancers and it is a major global health problem [3]. In Portugal, liver cancer did not figure among the most frequently diagnosed cancers. Nevertheless, it was the seventh leading cause of cancer mortality, being responsible for 3.8% of cancer deaths [1]. Some studies regarding this pathology have emerged, attempting to define its dimension in Portugal. According to the work of Tato Marinho et al. [4], HCC hospital admissions tripled from 1993 to 2005, with the overall costs of admission rising proportionally. In 2010, the Portuguese Society

* Corresponding author at: Centre for Informatics and Systems, University of Coimbra, Pólo II, Pinhal de Marrocos, 3030-290 Coimbra, Portugal.
E-mail addresses: miriams@student.dei.uc.pt (M.S. Santos), pha@dei.uc.pt (P.H. Abreu), pedroj.garcia@cud.upct.es (P.J. García-Laencina), adeliasimao@gmail.com (A. Simão), aspcarvalho@gmail.com (A. Carvalho).

http://dx.doi.org/10.1016/j.jbi.2015.09.012
1532-0464/

of Hepatology (PSH) predicted an increasing number of liver cases by approximately 70% by the end of 2015, seeking a greater national awareness regarding liver diseases [5].

Data-driven statistical research has become an attractive complement for clinical research. Survival prediction is one of the most challenging tasks addressed by the medical research communities [6–10]. It consists in analyzing a substantial amount of clinical data, drawing patterns and conclusions from those data, and using them to determine the survivability of a particular patient suffering from a given disease over a certain period of time. However, modeling and predicting disease outcomes may turn to be a difficult quest due to two main reasons: one relates to the dataset's size, while the other concerns its complexity.

Regarding the first topic, several authors consider that small datasets limit the scope of data mining techniques, since they may not provide enough information to accomplish the learning task of some algorithms [11,12]. Nevertheless, in real-life problems, specially in healthcare contexts, relatively small datasets are normal, specifically for less common diseases.

Dataset complexity can be derived from the characteristics of the data that composes the dataset. For datasets with heterogeneous data, the assumptions of some data mining algorithms may not be verified, and thus they might not be applicable [13]. For datasets with Missing Data (MD) (i.e., with variables containing a percentage of missing values and/or with records where several variables are incomplete), data mining algorithms may produce biased models and estimates, which decreases their performance [14].

Furthermore, patient heterogeneity is also an important topic to consider. In HCC guidelines, as in general cancer research, patient survival and prognosis are related to tumor stage [3]. However, growing studies regarding other diseases have pointed out the need to expand staging systems for predicting the outcome of cancer patients [15]. A more robust approach to study heterogeneous groups is cluster analysis. The main advantage in this type of approaches is that they generate homogeneous groups, with similar prognostic features, that map onto similar survival patterns, thus allowing a more accurate prediction.

The aim of this work is to start from the previously published literature on the application of computational techniques for HCC disease and assess to what extent they could be generalized for HCC dataset with complex characteristics. These characteristics consist of a relative small dataset size (165 patients), an heterogeneous set of predictive variables (49 clinical variables, including ratio-scaled, dichotomous and ordinal variables), a high percentage of missing values (an overall MD rate of 10.22% with only eight patients have complete information) and an expected heterogeneity between patients, due to the range of values in the considered values and the class imbalance for the HCC dataset (as detailed in Section 3.1). The majority of works on HCC are based on Neural Networks (NN) and Logistic Regression (LR) models (please refer to Section 2.2). However, all of these works ignore patient heterogeneity and the presence of missing data. In this work, both NN and LR are applied to a real incomplete HCC dataset, addressing the limitations found in previous research works. These algorithms are combined with four different approaches. In the first approach, the prediction models directly use the obtained dataset after a data imputation phase, while in the second approach the obtained dataset (after the clean-up procedure) is oversampled using SMOTE (Synthetic Minority Over-sampling Technique) algorithm [16]. The other two approaches are based on a new methodology proposed in this article, which consists in using a dataset produced by a cluster-based oversampling method. The third approach generates R different datasets and properly merges them into a unique representative dataset, $\mathcal{M}$, which is used to build the prediction models. Finally, the fourth approach considers a combination of each R previously oversampled dataset with the representative dataset $\mathcal{M}$. This last approach constructs a survival prediction model for each combination of R datasets with the representative dataset, and achieves the final classification results through majority voting. These four approaches are tested for both data mining algorithms (NN and LR) using a Leave-One-Out Cross Validation (LOO-CV) approach, which is appropriate for small sample datasets. For more information, please consult Section 4.

To the best of authors' knowledge, this kind of methodology has never been proposed and applied for a HCC dataset presenting these characteristics. This topic is fully detailed in Section 3.

Regarding Accuracy, Area Under the ROC Curve (AUC) and F-measure as performance indicators, the obtained results for our cluster-based oversampling approaches revealed statistical significant improvements on the performance of the NN algorithm, in comparison to the other two most commonly used approaches, proving that our methodology is generally feasible to design survival prediction models for HCC disease.

The remainder of this paper is organized as follows: Section 2 presents a brief description about HCC disease and illustrates some related works in the area. Section 3 outlines the methodological steps used in this project concerning the four project phases: Data collection, Data imputation, Cluster-based oversampling and Survival prediction. Section 4 reports the collected results and, finally, Section 5 presents the conclusions and proposals for further studies.

2. Computational approaches for HCC

In order to predict 1-year survival of HCC patients, it is important to understand some underlying aspects of this pathology and to review the previous related works on the application of computational methods to HCC disease.

2.1. Notions of HCC disease

A Carcinoma is a type of cancer that arises when an epithelial cell undergoes a malignant transformation. In particular, when the source of cancer is an epithelial cell cancer of the liver, known as hepatocyte, the cancer is called hepatocellular carcinoma (HCC) [17,3]. HCC may have different growth patterns. Some malignant tumors begin as a single tumor that grows larger and only spread to other parts of the liver in later stages. A second pattern is described by the appearance of small cancerous nodules scattered throughout the liver. This pattern is particularly common in patients with cirrhosis, and the most frequently detected in Portugal.

Approximately 90% of HCCs are associated with a known underlying risk factor [17,3]. The most frequent factors include chronic viral hepatitis (types B and/or C) and cirrhosis. Regarding both hepatitis virus, their corresponding main markers involve the measurements of specific antigens and antibodies, while cirrhosis is usually assessed with Child-Pugh (CP) score [18], which employs five clinical measures of liver disease (Total Bilirubin, Albumin, Encephalopathy, Ascites and Prothrombin Time). Cirrhosis is present in over 80% of HCC cases, being clearly identified as the main precursor lesion of this pathology.

2.2. Previous related works

Machine learning algorithms are computational techniques particularly well-suited to cancer research [19]. They are frequently used to analyze the available data about the disease under study (i.e. existing clinical trials) and produce new conclusions regarding a particular patient.