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.

Table 1
HCC developed works (n.a. – not applicable). In spite of the fact that the first work illustrated in Section 2.2. is one of the pioneers in the HCC area [20], it has not been covered in this table since the evaluation metrics are not provided.

		Ho et al. [21]	Chiu et al. [22]	Shi et al. [23]
	Objective	Disease-free survival after hepatic reSection (1st year)	Mortality after hepatic reSection (1st year)	Mortality after HCC surgery
	Sample size	427	434	22.926
	MD	No	No	No
NN	Accuracy	n.a	n.a	96%
	AUC	0.777	0.991	0.82
	Sensitivity	0.787	0.997	0.784
	Specificity	0.542	0.962	0.946
LR	Accuracy	n.a	n.a	84%
	AUC	0.772	0.890	0.730
	Sensitivity	0.754	0.986	0.626
	Specificity	0.583	0.346	0.919

With respect to the HCC disease, several research works have been previously performed [20–23]. In the first work [20], the authors introduced a regression model to diagnose liver disorders, having as a base a 200 cirrhotic patients dataset. Each clinical trial was composed by different types of variables, including laboratory tests and histopathological data. Nevertheless, and due to the fact that the number of HCC patients was not significant (only 5% of the cirrhotic patients), their results were very preliminary and not sufficient to validate the system. Besides, they did not consider any treatment for missing input values. In the second work, Ho et al. [21] attempted to establish a model to describe free-disease survival after hepatic resection, regarding a particular temporal line (1, 3 and 5 years). They have reviewed a study population of 482 HCC patients, in order to collect each patient's demographics, risk factors and several other variables related to the laboratory tests, tumor stage and the resection procedure itself. Three prediction models were tested: NN, LR and Decision Trees (DT). This work's results showed that NN outperformed the other two models in the great majority of training and validation groups. Despite proving good results, this work only considered HCC patients who have received hepatic resection, discarding patients in other stages of HCC disease. Thus, this works neglects patient heterogeneity, which is an added factor of complexity, considered in this work. Furthermore, this work also completely ignores the missing data perspective, which does not accurately tackle the true reality of these contexts.

Following this research work, the same authors compared the performance of NN and LR models to predict mortality of HCC patients who underwent liver resection [22]. The only relevant difference between these two works is the response of the algorithms: one seeks to predict disease-free survival and the other intends to predict if the patient is alive or dead in the considered periods (may he be disease-free or not). In another recent work [23], Shi et al. evaluated the use of NN and LR models for predicting in-hospital mortality in HCC surgery patients: the analysis was limited to patients who underwent a HCC surgery and clinical records with missing data were directly discarded. For both latter works, the previously detected limitations can also be found: they neglect patient heterogeneity and missing data as well.

Table 1 resumes the developed works in the HCC area. Performing an analysis of the Table 1 and, in conclusion, despite the growing interest and recent advances in the study of HCC, none of the studies so far as considered such a focused and complete approach to HCC data as the one proposed in this work. We conduct a study of patients' survivability only for HCC disease, prior to any therapeutic constraint, regarding a context with heterogeneous and missing data, and accounting for patient's heterogeneity, thus traducing the reality of most clinical contexts.

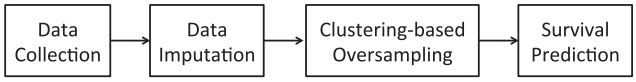

Fig. 1. Proposed methodology.

3. Methodology

This section describes the different four stages that compose the proposed methodology (see Fig. 1): Data collection, Data imputation, Cluster-based oversampling and Survival prediction. The main aspects of each stage are analyzed below.

3.1. Data collection

The first stage has been performed by the Service of Internal Medicine A of the Coimbra's Hospital and Universitary Centre (CHUC). It concerns the analysis of demographic, risk factor, laboratory and overall survival features from a set of $N = 165$ patients diagnosed with HCC. The resulting dataset comprises $n = 49$ features. They have been selected according to the EASL-EORTC (European Association for the Study of the Liver – European Organisation for Research and Treatment of Cancer) Clinical Practice Guidelines [3], the current state-of-the art on the management of HCC, in collaboration with a team of clinicians from CHUC's Service of Internal Medicine A. This dataset includes the clinical features considered to be the most significant to the clinicians' decision process, when choosing the most suitable therapeutic strategies and predicting its outcomes for each patient. A detailed description of the HCC dataset is presented in Table 2, which shows each feature's type/scale, range, statistics (mean/mode) and missing rate percentage. This is a heterogeneous dataset, with twenty-three quantitative variables (all ratio scaled) and twenty-six qualitative variables. Overall, the missing data represents 10.22% of the whole dataset and only eight patients have complete information in all the fields (4.85%).

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. This work is focused on the 1-year survivability prediction for HCC and, accordingly, the dataset's class distribution presents 63 cases labeled as 0 (dead) and the remaining 102 cases as 1 (alive).

3.2. Data imputation

In our methodology, this stage entails the process of ensuring that there are not inconsistencies in the collected data, i.e., missing

Table 2
Characterization of CHUC's hepatocellular carcinoma data. The dataset contains $N = 165$ records of $n = 49$ clinical variables, considered important to the clinicians decision process.

Prognostic factors	Type/scale	Range	Mean or mode	Missingness (%)
Gender	Qualitative/dichotomous	0/1	1	0
Symptoms	Qualitative/dichotomous	0/1	1	10.91
Alcohol	Qualitative/dichotomous	0/1	1	0
HBsAg	Qualitative/dichotomous	0/1	0	10.3
HBeAg	Qualitative/dichotomous	0/1	0	23.64
HBcAb	Qualitative/dichotomous	0/1	0	14.55
HCVAb	Qualitative/dichotomous	0/1	0	5.45
Cirrhosis	Qualitative/dichotomous	0/1	1	0
Endemic countries	Qualitative/dichotomous	0/1	0	23.64
Smoking	Qualitative/dichotomous	0/1	1	24.85
Diabetes	Qualitative/dichotomous	0/1	0	1.82
Obesity	Qualitative/dichotomous	0/1	0	6.06
Hemochromatosis	Qualitative/dichotomous	0/1	0	13.94
AHT	Qualitative/dichotomous	0/1	0	1.82
CRI	Qualitative/dichotomous	0/1	0	1.21
HIV	Qualitative/dichotomous	0/1	0	8.48
NASH	Qualitative/dichotomous	0/1	0	13.33
Esophageal varices	Qualitative/dichotomous	0/1	1	31.52
Splenomegaly	Qualitative/dichotomous	0/1	1	9.09
Portal hypertension	Qualitative/dichotomous	0/1	1	6.67
Portal vein thrombosis	Qualitative/dichotomous	0/1	0	1.82
Liver metastasis	Qualitative/dichotomous	0/1	0	2.42
Radiological hallmark	Qualitative/dichotomous	0/1	1	1.21
Age at diagnosis	Quantitative/ratio	20–93	64.69	0
Grams/day	Quantitative/ratio	0–500	71.01	29.09
Packs/year	Quantitative/ratio	0–510	20.46	32.12
Performance status	Qualitative/ordinal	0, 1, 2, 3, 4	0	0
Encefalopathy	Qualitative/ordinal	1, 2, 3	1	0.61
Ascites	Qualitative/ordinal	1, 2, 3	1	1.21
INR	Quantitative/ratio	0.84–4.82	1.42	2.42
AFP	Quantitative/ratio	1.2–1,810,346	19299.95	4.85
Hemoglobin	Quantitative/ratio	5–18.7	12.88	1.82
MCV	Quantitative/ratio	69.5–119.6	95.12	1.82
Leukocytes	Quantitative/ratio	2.2–13,000	1473.96	1.82
Platelets	Quantitative/ratio	1.71–459,000	113206.44	1.82
Albumin	Quantitative/ratio	1.9–4.9	3.45	3.64
Total Bil	Quantitative/ratio	0.3–40.5	3.09	3.03
ALT	Quantitative/ratio	11–420	67.09	2.42
AST	Quantitative/ratio	17–553	69.38	1.82
GGT	Quantitative/ratio	23–1575	268.03	1.82
ALP	Quantitative/ratio	1.28–980	212.21	1.82
TP	Quantitative/ratio	3.9–102	8.96	6.67
Creatinine	Quantitative/ratio	0.2–7.6	1.13	4.24
Number of nodules	Quantitative/ratio	0–5	2.74	1.21
Major dimension	Quantitative/ratio	1.5–22	6.85	12.12
Dir. bil	Quantitative/ratio	0.1–29.3	1.93	26.67
Iron	Quantitative/ratio	0–224	85.6	47.88
Sat	Quantitative/ratio	0–126	37.03	48.48
Ferritin	Quantitative/ratio	0–2230	439	48.48

values. In particular, this stage provides a cleaned complete database aimed at both minimizing the loss of clinical records and the distortion of the results in the later prediction stages. According to the literature [24,14,25], the most two conventional approaches used for managing missing data are to delete or impute values. By far, the most widely-used approach to missing data is the elimination of cases with unknown values. However, this procedure has been ruled out from the beginning, since 157 of 165 patients are incomplete. Then, an imputation-based approach had to be considered. Imputation is the process of replacing a missing datum with a substitute value [24], which is estimated using the available information in the database. This is an advantage compared to discarding incomplete cases, since imputing missing values provides additional information that can ease the later prediction stages and, thus, enhance the obtained results [9,10,26–28]. From the different imputation methods of the literature [14], we have chosen a nearest neighbor approach, which has shown its usefulness in many other clinical studies with missing values [29–31]. Initially, other simple statistical data imputation methods were tested, specifically mean/mode imputation and median imputation, which were found to add a distortion to the input data distribution. While these two methods ignore the relation between variables to perform imputation, KNN is a local approximation for imputation and, the, it is able to maintain the original input data distribution with a proper selection of K (in our case, $K = 1$). In this imputation approach, for each incomplete case $\mathbf{x}$, its closest neighbor $\mathbf{v}$ is chosen from those training samples with available information in the features to be imputed. This requires the computation of distances between each incomplete case and all the training samples, according to a similarity metric. We have used the Heterogeneous Euclidean-Overlap Metric (HEOM) distance [32], which efficiently handles both continuous and discrete variables in a missing data framework. Considering two input vectors, $\mathbf{x}_A$ and $\mathbf{x}_B$, the HEOM distance can be calculated by [32]:

$$d(\mathbf{x}_A, \mathbf{x}_B) = \sqrt{\sum_{j=1}^{n} d_j(x_{Aj}, x_{Bj})^2}, \quad (1)$$

being $d_j(x_{Aj}, x_{Bj})$ the distance between the two cases on its j-th attribute, where

$$d_j(x_{Aj}, x_{Bj}) = \begin{cases} 1, & \text{if } x_j \text{ is missing in } \mathbf{x}_A \text{ or } \mathbf{x}_B, \\ d_O(x_{Aj}, x_{Bj}), & \text{if } x_j \text{ is a discrete variable}, \\ d_N(x_{Aj}, x_{Bj}), & \text{if } x_j \text{ is a continuous variable}. \end{cases} \quad (2)$$

In Eq. (2), it is considered that the distance varies from 0 to 1 (the maximal distance value). If either one 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_O, for categorical attributes (Eq. (3)) and the normalized Euclidean distance, d_N, for continuous attributes (Eq. (4)):

$$d_O(x_{Aj}, x_{Bj}) = \begin{cases} 0, & \text{if } x_{Aj} = x_{Bj}; \\ 1, & otherwise; \end{cases} \quad (3)$$

$$d_N(x_{Aj}, x_{Bj}) = \frac{|x_{Aj} - x_{Bj}|}{max(x_j) - min(x_j)}. \quad (4)$$

Once the closest neighbor is found ($\mathbf{v}$), each unknown value in $\mathbf{x}$ is replaced by the corresponding available feature value of $\mathbf{v}$. At this point, it should be noted that (I) the closest neighbor imputation approach has been applied in order to maintain the variability of the dataset; and (II) there is not any previous research work about imputation for HCC databases with missing values. Finally, at the end of the data imputation stage, all features are standardized using the well-known Z-Score transformation [33].

3.3. Cluster-based oversampling

Once the data is cleaned, we try to find naturally occurring clusters (or groups) within our HCC database. Each group will be composed of several patient samples with similar feature values. As it is explained next, this work uses the GAP statistic [34] to automatically choose the number of groups (K) and, then, clustering is performed using the K-means algorithm.

3.3.1. Clustering patients using K-means

Due to the nature of the available data, its efficiency and success in several fields of pattern recognition [35], particularly for clustering cancer data [36,37], and its application potential to cluster-based sampling algorithms [38], we have chosen to apply K-means algorithm to cluster the HCC dataset.

K-means is a well-known unsupervised learning algorithm for data partition with low computational cost for high-dimensional datasets [39,35]. For a given number of groups (K), this method finds K centroids, $\{\mathbf{c}_k\}_{k=1}^{K}$, in order to place them as much as possible far away from each other. Those samples with the same nearest centroid are included in the same group: $\mathcal{C}_k$. K-means iteratively minimizes the sum of distances from each object to its centroid, over all clusters. In particular, the following error function is minimized [35]:

$$J = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in \mathcal{C}_k} d(\mathbf{x}_i, \mathbf{c}_k), \quad (5)$$

where $d(\mathbf{x}_i, \mathbf{c}_k)$ denotes the distance from the i-th input sample to the k-th centroid.

In the K-means algorithm, it is needed to perform an appropriate initialization of the centroids [35]. Bad initialization leads to suboptimal solutions with poor results. In order to avoid this drawback, our methodology uses the K-means++ procedure [40], which provides a robust initialization that leads to a competitive solution for the data partition. Another user-specified parameter is the number of clusters (K), which is a critical choice for the resulting partition [35]. Although there is not a theoretical criterion for selecting this parameter, the GAP statistic allows to find the proper K for K-means clustering [34]. It is a commonly used approach in practice which automatically provides very competitive results. According to Tibshirani et al. [34], and given that the sum of distances between the N_k points in $\mathcal{C}_k$ is $D_k = \sum_{\mathbf{x}_A, \mathbf{x}_B \in \mathcal{C}_k} d(\mathbf{x}_A, \mathbf{x}_B)$, the intra-cluster variance, $W_K = \sum_{k=1}^{K} \frac{1}{2N_k} D_k$, gives a measure of the compactness of our clustering. W_K can be used to heuristically determine the optimal K: considering a range of possible values for K, the evolution of W_K with respect to the number of clusters is plotted and, then, the most dramatic decrease ("elbow") in the plot is found for the optimal value of K. The gap statistic formalizes this heuristic procedure and it automatically provides the optimal K [34]. To assess the optimal number of clusters for the HCC dataset, the gap statistic was calculated for a range of 2–30 clusters. The optimal number of clusters was found for $K = 10$ clusters.

An important issue is that different runs (initializations of K centroids) of K-means give different partitions. For overcoming this inconvenient in practice, multiple different initializations are considered and, from its K-means solutions, the partition with the smallest error is selected as final [35]. In contrast, several works have implemented ensembles methods by combining multiple partitions to obtain an integrated partition using a consensus function [41–44]. Nevertheless, in our approach, the aim is not to achieve a unique clustering. As it is explained next, our proposed methodology exploits the diversity of the multiple obtained partitions for constructing an augmented dataset in a two-phase sampling procedure.

3.3.2. First sampling phase: balancing groups by synthetic samples

In this first stage, oversampling is applied in order to diminish the impact of underlying patient groups/profiles with reduced sizes on survival prediction. In most clinical databases, several patient profiles can be found and, naturally, with different sizes. In terms of groups, a database is imbalanced if its underlying clusters are not approximately equally represented. As it is shown in our experiments, a high imbalance in the sizes of the patient profiles hinders the design of survival prediction models. For most imbalanced datasets, the application of sampling techniques improves classifier accuracy. Random oversampling and undersampling are two of the most common sampling techniques [16]. Oversampling augments the original set of samples by randomly replicating minority class examples. Undersampling removes majority class examples from the original set. Both of these algorithms achieve class balance, but can also potentially hinder the learning task: oversampling does not incorporate any new information and may lead to overfitting, while undersampling may remove important examples to the learning step, causing the classifier to miss important concepts [38].

To avoid this drawback, this work takes advantage of the SMOTE algorithm [16], which is the most popular and applied oversampling procedure. SMOTE algorithm generates synthetic minority samples, based on the similarity between the available minority samples, considering it's K-nearest neighbors [38]. In particular, we have also implemented a cluster-based approach which follows the same principles of SMOTE. In the original version of SMOTE, this algorithm is used to oversample the minority class label, which means the class of the newly generated synthetic samples is already previously established. In our implemented approach, SMOTE has been adapted to oversample clusters with reduced sizes. Some clusters may contain different class labels. Thus, the assessment of the class label for each new synthetic sample is done according to a random number between 0 and 1, call it φ, used in the original SMOTE implementation to create each new synthetic sample.

In brief, this first sampling phase is composed of the following main steps:

1. *Selection of clusters with reduced sizes*. Instead of balancing all groups to the largest one, the size reference is established to the second largest cluster (this criteria was chosen after performing some preliminary experiments). Then, oversampling is performed in those clusters with lower sizes than this reference.
2. *Generation of synthetic samples*. Within each cluster $\mathcal{C}_k$ of reduced size:
   (a) Consider each sample $\mathbf{x}$ in the cluster. Note that if amount of oversampling (relationship between the samples to be generated and the existing ones) does not require the oversampling of all the existing samples, the samples to oversample are chosen randomly.
   (b) Choose one of its V nearest neighbors, $\mathbf{v}$. In this work, several different values of V were tested, from 1 to 5 nearest neighbors. $V = 3$ has provided the best results (considering the complete experimental setup), and thus it was used as the appropriate number of neighbors to SMOTE.
   (c) Create a new synthetic sample $\mathbf{s}$ according to SMOTE's equation:

$$\mathbf{s} = \mathbf{x} + \varphi(\mathbf{x} - \mathbf{v}); \tag{6}$$

   being φ a random number between 0 and 1.
   (d) The class label (survival status) of $\mathbf{s}$ is assigned according to φ. If φ is greater than 0.5, the class label of $\mathbf{s}$ is the same as $\mathbf{v}$. On the contrary, if φ is smaller or equal to 0.5, the class label of $\mathbf{s}$ is the same as $\mathbf{x}$.
   (e) Step (c) is repeated according to the amount of oversampling required.

It should be noted that the above procedure is repeated for each obtained partition in the K-means clustering. Fig. 2 depicts a scheme of this first sampling phase, where there is also shown the previous stages of data imputation and clustering of the dataset $\mathcal{D}$. Let us assume that R runs of K-means have been done, where the value of K has been previously determined using the GAP statistic. For the r-th run, with $r = 1, 2, \ldots, R$, its obtained partition is defined by K different clusters: $\{\mathcal{C}_{r,k}\}_{k=1}^{K}$. Those clusters with reduced sizes are oversampled:

$$\left\{\mathcal{C}^*_{r,k}\right\}_{k=1}^{K} = \{\mathcal{C}_{r,k} \cup S_{r,k}\}_{k=1}^{K}, \tag{7}$$

being $S_{r,k}$ the subset of generated samples for the k-th group in the r-th partition. Therefore, at the end of this first sampling phase, we have R different datasets: $\{\mathcal{D}^*_r\}_{r=1}^{R}$, where $\mathcal{D}^*_r = \cup_k C^*_{r,k}$.

3.3.3. Second sampling phase: construction of a representative dataset

In this second sampling phase, the goal is to exploit the diversity of the different generated sets, $\mathcal{D}^*_1, \mathcal{D}^*_2, \ldots, \mathcal{D}^*_R$, to obtain an augmented dataset, $\mathcal{M}$, which provides a better representation of the survival prediction problem. This work considers and evaluates two sampling schemes for merging the information from the

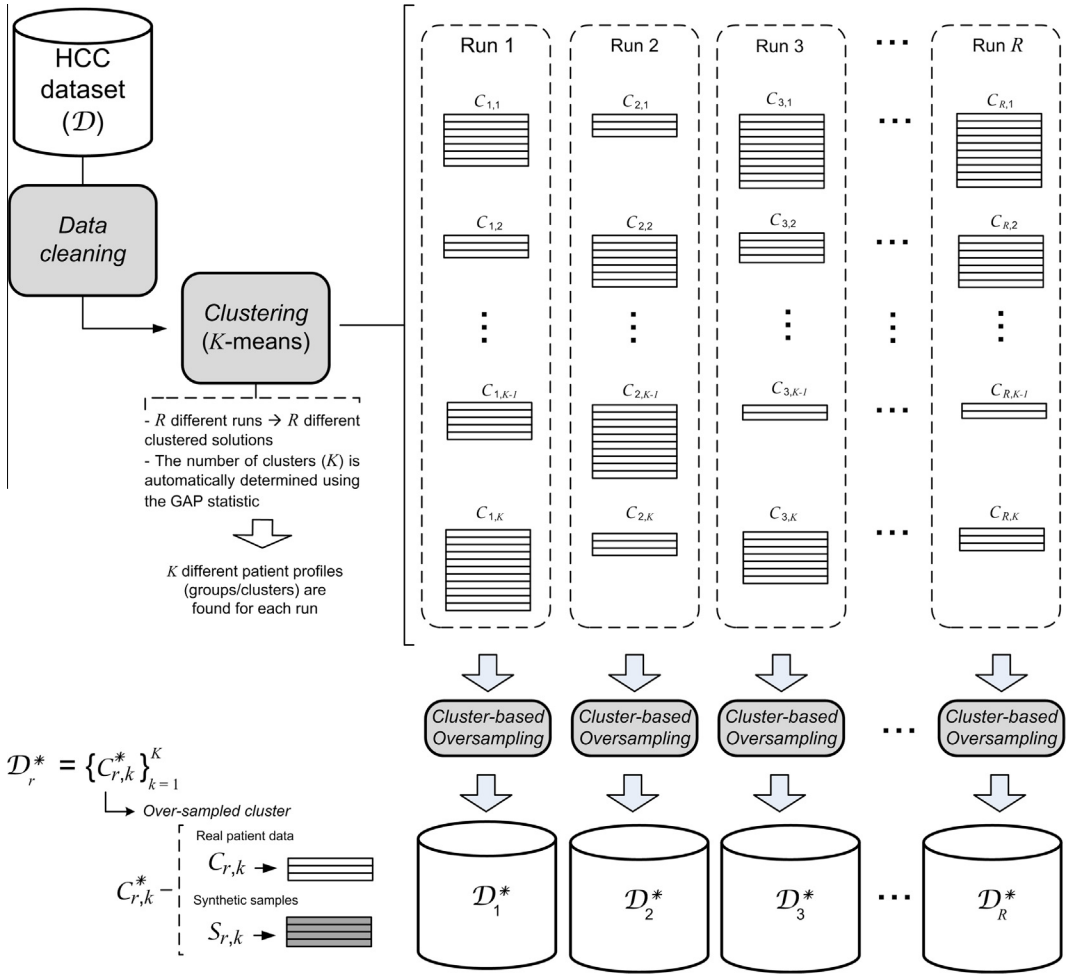

Fig. 2. First sampling phase: balancing groups by the generation of synthetic samples.

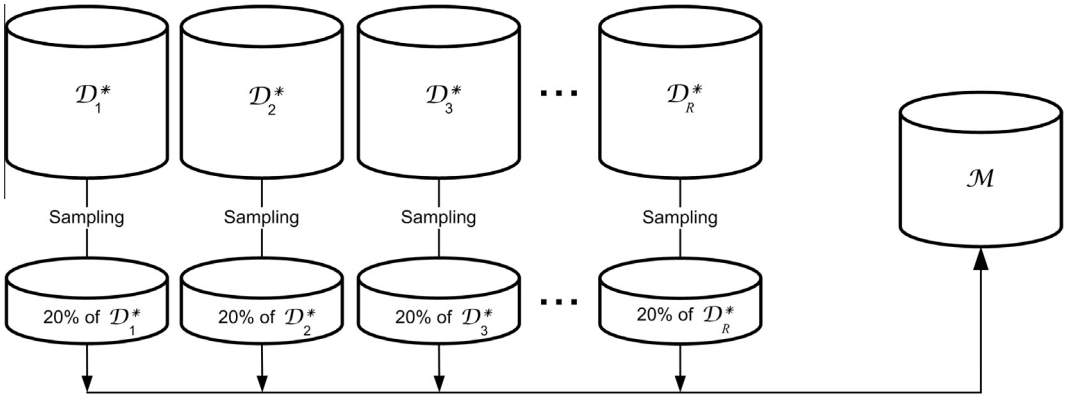

Fig. 3. Second sampling phase: construction of a representative dataset.

multiple oversampled datasets. In the second sampling scheme, $\mathcal{M}$ is defined by merging R data portions which have been sampled from $\{\mathcal{D}_r^*\}_{r=1}^R$. The resulting dataset, $\mathcal{M}$, is used to model survival prediction for HCC disease. In this work, each data portion is composed of 20% of samples from $\mathcal{D}_r^*$, following the principles of stratified random sampling method [45]. This ratio provides a representative contribution of each oversampled dataset and it has been chosen according to our experience in the HCC dataset. This sample scheme is illustrated in Fig. 3.

Based on this first sampling scheme, and instead of providing a single representative dataset ($\mathcal{M}$), we also implement another combination approach which finally produces R augmented datasets, $\{\mathcal{M}_r\}_{r=1}^R$. In particular, $\mathcal{M}_r$ is composed of $\mathcal{D}_r$ and $R-1$ portions of samples of the remaining oversampled datasets. Here, the same percentage of sampling is considered (20%) for each portion. With respect to survival prediction, in this second sampling scheme, R different models have to be designed using each representative dataset and, as it is explained next, their resulting R predictions are combined using majority voting.

3.4. Survival prediction

In this work, two well-known classification methods are applied [39]: Neural Networks (NN) and Logistic Regression (LR). These classifiers have shown their usefulness for survival prediction in previous research works with HCC data [21,23,22]. For each one of them, this work studies the impact of using the different generated datasets obtained with our cluster-based oversampling method on survival prediction for HCC disease.

4. Experiments

The proposed methodology has been experimentally evaluated using the HCC dataset, which has been previously described in SubSection 3.1. The experiments were performed in order to show that our proposed methodology is generally feasible to design survival prediction models for HCC disease and, also, that it outperforms other widely-used approaches. In this work, we have carried out a series of simulations, considering four different approaches. In the first approach, survival prediction models (for NN and LR) are developed using directly the dataset obtained from the data imputation stage, $\mathcal{D}$ (Without Cluster, No-Oversampling). For the second approach, the minority class in $\mathcal{D}$ is oversampled using the SMOTE algorithm (Without Cluster, Oversampling). Then, the same two classification algorithms are used. Note that, in this second approach, we analyze the impact of overcoming the class-imbalance in $\mathcal{D}$ on classification. In the third and fourth approaches, our methodology is applied. The third approach obtains a unique representative dataset $\mathcal{M}$ using the proposed cluster-based oversampling method, which has been described in SubSection 3.3 (With Cluster, Representative Set Approach). After that, $\mathcal{M}$ is used for constructing a survival prediction model using each classification algorithm. Finally, instead of providing a unique dataset $\mathcal{M}$, the fourth approach obtains R augmented datasets, $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_R$ (With Cluster, Augmented Sets Approach). For each one of them and for each classification algorithm, a survival prediction model is constructed. Then, the classification results obtained from the R models trained with the same classification method are combined by a majority voting scheme.

Experiments have been performed using a Leave-One-Out Cross Validation (LOO-CV) process for performance evaluation [39], which is appropriate since the amount of available data is not large. This evaluation scheme also avoids undesirable shifts from the random selection of training and test sets. Specifically, for the N total number of samples involved in the study, one is left out for testing, and the remaining $N-1$ are used for designing the survival prediction models. For each iteration of the LOO-CV procedure, thirty runs of the cluster-oversampling were performed.

All simulations have been carried out in MATLAB 8.2 (R2013b) environment running in the same machine.

To perform the evaluation of each classifier (NN and LR), three different measures were used: Accuracy, AUC and F-Measure. Traditionally, the most widely-used performance measure in classification problems is Accuracy. However, it ignores the probability estimations of classification in favor of class labels. In many research areas, and particularly biomedical applications, two additional performance measures based on the confusion matrix and the ROC (Receiver Operator Curve) are usually applied: AUC (area under the ROC curve) and F-measure. In one hand, the area under the ROC curve (AUC) is a measure of how well a classification model can distinguish between two diagnostic groups (diseased/normal). In practice, the AUC is often used when a representative measure of discrimination is needed and it can even replace the Accuracy as a performance measure [46]. In the other hand, the F-measure is defined as the harmonic mean of Precision (or the Positive Predictive Value) and Recall (or Sensitivity or True Positive Rate) and, then, it provides a balance between both performance metrics. For each one of these metrics, three indicators were used: Mean, Standard deviation (Std) and Rank. The first two indicators, Mean and Std, are computed from the obtained experimental results with the different configurations of the classification methods (i.e. different hidden layer sizes for the NN classifier and different thresholds for the LR classifier). According to the work proposed by Demsar [47], the third indicator is the Rank obtained by a Friedman rank test, which was used to compare the obtained performance results (Accuracy/AUC/F-measure) in the four tested approaches regarding both classifiers. Tables 3 and 4 respectively

Table 3
Neural Networks (NN) performance evaluation using Accuracy, AUC and F-Measure. For each measure, three indicators were used: mean and standard deviation (std) for the best configuration of each approach and the Rank of the Friedman rank test.

Approach	Accuracy			AUC			F-Measure		
	Mean	Std	Rank	Mean	Std	Rank	Mean	Std	Rank
Without cluster									
No-oversampling	0.687	0.043	4	0.650	0.068	3.8	0.550	0.075	4
Oversampling	0.717	0.038	2.91	0.661	0.034	3.2	0.645	0.027	2.73
With cluster									
Representative set approach	0.737	0.023	2.09	0.689	0.021	2	0.640	0.034	2.27
Augmented sets approach	0.752	0.011	1	0.700	0.015	1	0.665	0.018	1

Table 4
Logistic Regression (LR) performance evaluation using Accuracy, AUC and F-Measure. For each measure, three indicators were used: mean and standard deviation (std) for the best configuration of each approach and the Rank of the Friedman rank test. It should be noted that for Without Cluster and No-Oversampling, std values are not applicable, n.a.

Approach	Accuracy			AUC			F-Measure		
	Mean	Std	Rank	Mean	Std	Rank	Mean	Std	Rank
Without cluster									
No-oversampling	0.721	n.a	2.4	0.658	n.a	2	0.651	n.a	2.6
Oversampling	0.706	0.010	3	0.649	0.007	3	0.639	0.012	3.2
With cluster									
Representative set approach	0.725	0.016	2.6	0.668	0.014	3	0.648	0.020	2.4
Augmented sets approach	0.730	0.014	2	0.673	0.012	2	0.652	0.016	1.8

show the obtained experimental results of the four approaches for each classifier, NN and LR. Due to space limitations and to make easier the experimentation analysis for the reader, in Tables 3 and 4, the first two indicators, Mean and Std, are related to the best results achieved by each of the classifier configurations; and, meanwhile, the third indicator, Rank, is related to the final ranking of the Friedman test. For more detailed information about simulation results, please consult Appendix A.

Table 3 illustrates the obtained results by NN classifier. From the analysis of the table, it is easy to note that using any of the evaluation measures (Accuracy, AUC and F-Measure) and only the first two indicators (mean and std), Augmented Sets Approach presented the best results in comparison to the other three approaches.

As concerns the NN classifier, eleven distinct network configurations were used in the experience (5–55 hidden neurons in a step of 5) and for each hidden layer size, 30 runs were performed. For Friedman rank test, each group represents the average Accuracy of each considered configuration of neurons for each of the four used approaches. The average of the results of each approach is compared to all other approaches and, following the work presented by Demsar [47], $Ff = 7.691$ was calculated and compared to the F distribution $(3, 30) = 2.92$ with a significance level of $\alpha = 0.5$. As consequence, the null hypothesis of equivalence between the four approaches is rejected. Comparing all the four approaches for a 5% significance level using the Nemenyi test [47], it was possible to obtain CD = 1.4142. CD is the critical value for the difference of mean ranks between the four approaches. Attending to Table 3, Augmented Sets Approach performed better than the approaches that did not use cluster strategy. Also, Representative Set Approach performed better than the first two approaches that can be considered widely-used approaches. These findings follow the ones previously detected for the other indicators (mean and std).

Regarding the other classification method (LR), the same analysis was performed. Table 4 illustrates the Accuracy results for LR classifier. Mean and std represent the best results presented by each approach related to a specific threshold. Once again, Augmented Sets Approach presented better results than the other three approaches. It should be noted that for the first approach (Without Cluster, No-Oversampling), the LR model always produces the same results, since there are not any random factors considered in this method (the input data is always the same). Thus, only one run was performed and, then, std measure is not applicable (n.a.) for this first approach. The other three approaches consider oversampling of cases, which implies adding random factors to the procedure (chosen nearest neighbor, random number φ), and thus, similarly to the NN case, 30 runs of these approaches were also performed. To produce the rank mean value, each group corresponds to one of the five different thresholds used (0.5–0.9) and, for each configuration, 30 runs were performed (as in Table 3, only the final rank mean was displayed). Again, for this scenario $Ff = 7.800$ was calculated and compared to the F distribution $(3, 12) = 3.4903$ with a significance level of $\alpha = 0.5$. Consequently, the null hypothesis of equivalence between the four approaches is rejected. Comparing all the four approaches for a 5% significance level using the Nemenyi test [47], it was possible to obtain CD = 2.0976. With respect to Table 4, none of the approaches proved to be better than the others using Rank as indicator.

Finally, in spite of the fact that the related works illustrated on Section 2.2 did not cover the same topic as our proposed approach, and due to that they cannot be directly compared, for our best approach (Augmented Sets Approach), the Sensitivity and Specificity values for both algorithms (NN and LR) are in the range of the previously published results. For NN and LR, the sensitivity results were 0.647 and 0.741 and, for specificity, the results were 0.827 and 0.777 respectively.

5. Conclusions and future work

In this work, a new methodology capable of predicting the 1-year survival for patients with HCC has been presented. To achieve that, a HCC dataset composed by 165 patients followed in an university hospital center was used. At the beginning of the study, this dataset presented three main challenges: its heterogeneity concerning the type of considered variables (49 features were used encompassing dichotomous, ratio scaled and ordinal features),

the percentage of MD present in the dataset (overall, MD constitutes 10.22% of the total data with only eight patients having complete information) and, finally, the data unbalance observed, that made even more difficult to create a valid methodology to predict 1-year survival for patients with HCC. The proposed methodology relied on a cluster-based oversampling approach where two classifiers (NN and LR) were separately used in two novel approaches, referred to as Representative Set Approach and Augmented Sets Approach, and compared with two widely-used approaches (explained in the previous sections). The main difference of these two sets of approaches consists in using a new cluster-based methodology that addresses the challenges previously detected at the beginning of the study.

The achieved results were assessed using three performance measures: Accuracy, AUC and F-measure. To compare the obtained results of the four used approaches for each classifier, Friedman rank test was used as proposed in Demsar's work [47]. The proposed methodology coupled with NN classifier presented better results than the other two widely-used approaches regarding all the performance measures previously defined, proving that our methodology provides a more appropriate approach to design survival prediction models in a HCC context with the discussed characteristics.

There are two possible directions for future works: the application of the proposed methodology to other medical and non-medical classification problems; and its extension to estimate missing values in the input data.

To the extent of authors' knowledge, this methodology has never been proposed or applied in HCC dataset in particular, or other diseases or subjects in general. Thus, the issue of reproducibility and generalization for other contexts has not yet been addressed. This topic could be a possibility for future work: extending our methodology to other contexts besides HCC disease, whether they are healthcare contexts or not.

Another ongoing work is the application of the proposed methodology to MD imputation. The implemented cluster-based oversampling algorithm could be adapted to estimate missing values. In this work, the data in each cluster is complete and the new synthetic samples are generated as explained in Section 3.3.2. Considering incomplete data in the clusters, a modification of our algorithm would be used to generate new samples, where each missing value is replaced. Then, these newly generated samples could be used in our cluster-based approach to impute the missing observation in the original dataset. Following the LOO-CV procedure, each missing value could be replaced $\mathcal{M}$ times, and for each one of these $\mathcal{M}$ times, a classifier would be used. The proper values to be imputed should be chosen according to the set that allows the best classification performance.

Acknowledgments

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

Appendix A. Simulation results

Tables 5–10

Table 5
Obtained Accuracy results (mean ± standard deviation) in the HCC dataset using NN architectures with different hidden layer sizes (from 5 to 55).

Number of neurons	Without cluster		With cluster	
	No-oversampling	Oversampling	Representative set approach	Augmented sets approach
5	0.6525 ± 0.0314	0.6927 ± 0.0211	0.7131 ± 0.0208	0.7477 ± 0.0160
10	0.6709 ± 0.0210	0.7016 ± 0.0257	0.7087 ± 0.0186	0.7396 ± 0.0251
15	0.6659 ± 0.0228	0.7077 ± 0.0282	0.7117 ± 0.0263	0.7493 ± 0.0106
20	0.6869 ± 0.0432	0.7006 ± 0.0228	0.7238 ± 0.0245	0.7360 ± 0.0220
25	0.6814 ± 0.0450	0.7170 ± 0.0379	0.7133 ± 0.0196	0.7436 ± 0.0126
30	0.6618 ± 0.0283	0.6956 ± 0.0289	0.7279 ± 0.0202	0.7519 ± 0.0105
35	0.6602 ± 0.0230	0.6842 ± 0.0330	0.7368 ± 0.0225	0.7461 ± 0.0153
40	0.6632 ± 0.0397	0.6947 ± 0.0195	0.7248 ± 0.0226	0.7420 ± 0.0207
45	0.6352 ± 0.0270	0.6768 ± 0.0326	0.7119 ± 0.0245	0.7392 ± 0.0193
50	0.6406 ± 0.0273	0.6768 ± 0.0240	0.7220 ± 0.0227	0.7360 ± 0.0232
55	0.6412 ± 0.0243	0.6788 ± 0.0272	0.7236 ± 0.0178	0.7453 ± 0.0202

Table 6
Obtained AUC results (mean ± standard deviation) in the HCC dataset using NN architectures with different hidden layer sizes (from 5 to 55).

Number of neurons	Without cluster		With cluster	
	No-oversampling	Oversampling	Representative set approach	Augmented sets approach
5	0.6179 ± 0.0332	0.6377 ± 0.0179	0.6626 ± 0.0215	0.6941 ± 0.0161
10	0.6353 ± 0.0209	0.6465 ± 0.0200	0.6573 ± 0.0205	0.6871 ± 0.0248
15	0.6267 ± 0.0259	0.6467 ± 0.0252	0.6633 ± 0.0239	0.6983 ± 0.0125
20	0.6473 ± 0.0397	0.6452 ± 0.0180	0.9731 ± 0.0203	0.6883 ± 0.0214
25	0.6435 ± 0.0434	0.6610 ± 0.0343	0.6652 ± 0.0199	0.6943 ± 0.0163
30	0.6279 ± 0.0300	0.6432 ± 0.0230	0.6754 ± 0.0221	0.6998 ± 0.0122
35	0.6729 ± 0.0287	0.6310 ± 0.0271	0.6887 ± 0.0210	0.7002 ± 0.0154
40	0.6502 ± 0.0682	0.6391 ± 0.0157	0.6757 ± 0.0195	0.6935 ± 0.0246
45	0.6022 ± 0.0369	0.6257 ± 0.0260	0.6646 ± 0.0256	0.6926 ± 0.0241
50	0.6046 ± 0.0323	0.6265 ± 0.0196	0.6754 ± 0.0231	0.6880 ± 0.0269
55	0.6061 ± 0.0287	0.6304 ± 0.0233	0.6762 ± 0.0163	0.6982 ± 0.0229

Table 7
Obtained F-Measure results (mean ± standard deviation) in the HCC dataset using NN architectures with different hidden layer sizes (from 5 to 55).

Number of neurons	Without cluster		With cluster	
	No-oversampling	Oversampling	Representative set approach	Augmented sets approach
5	0.4780 ± 0.0548	0.6134 ± 0.0277	0.6184 ± 0.0301	0.6620 ± 0.0199
10	0.5194 ± 0.0372	0.6178 ± 0.0363	0.6169 ± 0.0224	0.6517 ± 0.0308
15	0.5297 ± 0.0463	0.6446 ± 0.0270	0.6107 ± 0.0394	0.6604 ± 0.0141
20	0.5496 ± 0.0754	0.6185 ± 0.0322	0.6279 ± 0.0395	0.6384 ± 0.0324
25	0.5345 ± 0.0803	0.6361 ± 0.0457	0.6128 ± 0.0250	0.6515 ± 0.0138
30	0.4940 ± 0.0551	0.6059 ± 0.0391	0.6385 ± 0.0315	0.6650 ± 0.0182
35	0.4876 ± 0.0412	0.6056 ± 0.0378	0.6399 ± 0.0340	0.6490 ± 0.0228
40	0.4590 ± 0.0452	0.6160 ± 0.0261	0.6255 ± 0.0469	0.6499 ± 0.0208
45	0.4473 ± 0.0373	0.5942 ± 0.0375	0.6110 ± 0.0291	0.6433 ± 0.0198
50	0.4598 ± 0.0507	0.5883 ± 0.0322	0.6202 ± 0.0343	0.6415 ± 0.0256
55	0.4619 ± 0.0360	0.5823 ± 0.0412	0.6220 ± 0.0333	0.6508 ± 0.0259

Table 8
Obtained accuracy results (mean ± standard deviation) in the HCC dataset using LR architectures with different thresholds (from 1 to 5). For without cluster, no-oversampling, mean and std values are not applicable, n.a (4), but the corresponding values for this approach are 0.7030, 0.7091, 0.7030, 0.7212, 0.7152.

Number of neurons	Without cluster		With cluster	
	No-oversampling	Oversampling	Representative set approach	Augmented sets approach
1	n.a.	0.7018 ± 0.0093	0.7248 ± 0.0158	0.7301 ± 0.0139
2	n.a.	0.7030 ± 0.0090	0.7220 ± 0.0163	0.7267 ± 0.0147
3	n.a.	0.7053 ± 0.0091	0.7109 ± 0.0166	0.7174 ± 0.0136
4	n.a.	0.7055 ± 0.0095	0.6986 ± 0.0169	0.6966 ± 0.0118
5	n.a.	0.7057 ± 0.0100	0.6721 ± 0.0163	0.6733 ± 0.0124

Table 9
Obtained AUC results (mean ± standard deviation) in the HCC dataset using LR architectures with different thresolds (from 1 to 5). For without cluster, no-oversampling, mean and std values are not applicable, n.a (4), but the corresponding values for this approach are 0.6520, 0.6535, 0.6476, 0.6585, 0.6528.

Number of neurons	Without cluster		With cluster	
	No-oversampling	Oversampling	Representative set approach	Augmented sets approach
1	n.a.	0.6489 ± 0.0074	0.6683 ± 0.0137	0.6725 ± 0.0124
2	n.a.	0.6483 ± 0.0072	0.6613 ± 0.0139	0.6662 ± 0.0135
3	n.a.	0.6485 ± 0.0069	0.6477 ± 0.0130	0.6541 ± 0.0110
4	n.a.	0.6470 ± 0.0074	0.6351 ± 0.0122	0.6340 ± 0.0085
5	n.a.	0.6445 ± 0.0081	0.6137 ± 0.0109	0.6144 ± 0.0082

Table 10
Obtained F-Measure results (mean ± standard deviation) in the HCC dataset using LR architectures with different thresolds (from 1 to 5). For without cluster, no-oversampling, mean and std values are not applicable, n.a (4), but the corresponding values for this approach are 0.6080, 0.6250, 0.6202, 0.6515, 0.6466.

Number of neurons	Without cluster		With cluster	
	No-oversampling	Oversampling	Representative set approach	Augmented sets approach
1	n.a.	0.6119 ± 0.0117	0.6417 ± 0.0211	0.6489 ± 0.0187
2	n.a.	0.6181 ± 0.0127	0.6483 ± 0.0196	0.6518 ± 0.0154
3	n.a.	0.6253 ± 0.0133	0.6471 ± 0.0196	0.6509 ± 0.0164
4	n.a.	0.6305 ± 0.0129	0.64445 ± 0.0182	0.6407 ± 0.0133
5	n.a.	0.6385 ± 0.0142	0.6326 ± 0.0168	0.6339 ± 0.0127

References

[1] W.H. Organization, Globocan 2012: estimated cancer incidence, mortality and prevalence worldwide in 2012. <http://globocan.iarc.fr/>.

[2] W.H. Organization, Cancer fact sheet, 2014. <http://www.who.int/mediacentre/factsheets/fs297>.

[3] Anon., European association for the study of the liver, European organisation for research and treatment of cancer, EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma, J. Hepatol. 56 (4) (2012) 908–943.

[4] R.T. Marinho, J. Giria, M.C. Moura, Rising costs and hospital admissions for hepatocellular carcinoma in portugal (1993–2005), World J. Gastroenterol. 13 (10) (2007) 1522–1527.

[5] L.P.C. Cancro, Cancro do fígado pode aumentar 70 por cento até, 2015. <http://www.ligacontracancro.pt/noticias/detalhes.php?id=115>.

[6] H.B. Burke, P.H. Goodman, D.B. Rosen, D.E. Henson, J.N. Weinstein, F.E. Harrell, J.R. Marks, D.P. Winchester, D.G. Bostwick, Artificial neural networks improve the accuracy of cancer survival prediction, Cancer 79 (4) (1997) 857–862.

[7] J. Thongkam, G. Xu, Y. Zhang, F. Huang, Toward breast cancer survivability prediction models through improving training space, Expert Syst. Appl. 36 (10) (2009) 12200–12209.

[8] N. Esfandiari, M.R. Babavalian, A.-M.E. Moghadam, V.K. Tabar, Knowledge discovery in medicine: current issue and future trend, Expert Syst. Appl. 41 (9) (2014) 4434–4463.

[9] P.H. Abreu, H.A. Amaro, D. Castro-Silva, P. Machado, M.H. Abreu, N. Afonso, A. Dourado, Overall survival prediction for women breast cancer using ensemble methods and incomplete clinical data, in: L.M. Roa Romero (Ed.), XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013, IFMBE Proceedings, vol. 41, Springer, 2014, pp. 1366–1369.

[10] P.H. Abreu, H.A. Amaro, D. Castro-Silva, P. Machado, M.H. Abreu, N. Afonso, A. Dourado, Personalizing breast cancer patients with heterogeneous data, in: Y.-T. Zhang (Ed.), International Conference on Health Informatics, IFMBE Proceedings, vol. 42, Springer, 2014, pp. 39–42.

[11] J. Yuan, T. Fine, Neural-network design for small training sets of high dimension, IEEE Trans. Neural Netw. 9 (2) (1998) 266–280.

[12] R. Andonie, Extreme data mining: Interference from small datasets, Int. J. Comput. Commun. Control 5 (3) (2010) 280–291.

[13] F. Harrell, K. Lee, D. Mark, Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, Stat. Med. 15 (4) (1996) 361–387.
[14] P.J. García-Laencina, J.L. Sancho-Gómez, A. Figueiras-Vidal, Pattern classification with missing data: a review, Neural Comput. Appl. 19 (2010) 263–282.
[15] K. Qi, D. Wu, L. Sheng, D. Henson, A. Schwartz, E. Xu, K. Xing, D. Chen, On an ensemble algorithm for clustering cancer patient data, BMC Syst. Biol. 7 (Suppl. 4) (2013) S9.
[16] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (1) (2002) 321–357.
[17] A. Forner, J.M. Llovet, J. Bruix, Hepatocellular carcinoma, Lancet 379 (9822) (2012) 1245–1255.
[18] F. Durand, D. Valla, Assessment of the prognosis of cirrhosis: childpugh versus meld, J. Hepatol. 42 (2005) S100–S107.
[19] J.A. Cruz, D.S. Wishart, Applications of machine learning in cancer prediction and prognosis, Cancer Informat. 2 (2006) 59–78.
[20] H.A. Wasyluk, J. Cianciara, L. Bobrowski, A. Drapato, Founding of database for cirrhotic patients for early detection of hepatocellular carcinoma, Hepatology 6 (3) (2010) 13–16.
[21] W.-H. Ho, K.-T. Lee, H.-Y. Chen, T.-W. Ho, H.-C. Chiu, Disease-free survival after hepatic resection in hepatocellular carcinoma patients: a prediction approach using artificial neural network, PLoS ONE 7 (1) (2012) e29179.
[22] H.C. Chiu, T.W. Ho, L.K. T., H.Y. Chen, W.H. Ho, Mortality predicted accuracy for hepatocellular carcinoma patients with hepatic resection using artificial neural network, Sci. World J. 2013 (2013) 201976–10.
[23] H.-Y. Shi, K.-T. Lee, H.-H. Lee, W.-H. Ho, D.-P. Sun, J.-J. Wang, C.-C. Chiu, Comparison of artificial neural network and logistic regression models for predicting in-hospital mortality after primary liver cancer surgery, PLoS ONE 7 (4) (2012) e35781.
[24] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, second ed., Wiley, Hoboken, NJ, 2002.
[25] F. Cismondi, A.S. Fialho, S.M. Vieira, S.R. Reti, J.M. Sousa, S.N. Finkelstein, Missing data in medical databases: impute, delete or classify?, Artif Intell. Med. 58 (1) (2013) 63–72.
[26] P.J. García-Laencina, J.-L. Sancho-Gómez, A.R. Figueiras-Vidal, M. Verleysen, K nearest neighbours with mutual information for simultaneous classification and missing data imputation, Neurocomputing 72 (7–9) (2009) 1483–1493.
[27] P.J. García-Laencina, J.-L. Sancho-Gómez, A.R. Figueiras-Vidal, Classifying patterns with missing values using multi-task learning perceptrons, Expert Syst. Appl. 40 (4) (2013) 1333–1341.
[28] P.J. García-Laencina, P.H. Abreu, M.H. Abreu, N. Afonso, Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values, Comput. Biol. Med. 59 (2015) 125–133.
[29] R.J.A. Little, Methods for handling missing values in clinical trials, J. Rheumatol. 26 (8) (1999) 1654–1656.
[30] O. Troyanskaya, M. Cantor, O. Alter, G. Sherlock, P. Brown, D. Botstein, R. Tibshirani, T. Hastie, R. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics 17 (2001) 520–525.
[31] J.M. Jerez, I. Molina, P.J. Garcia-Laencina, E. Alba, N. Ribelles, M. Martin, L. Franco, Missing data imputation using statistical and machine learning methods in a real breast cancer problem, Artif. Intell. Med. 50 (2) (2010) 105–115.
[32] G.E. Batista, M.C. Monard, An analysis of four missing data treatment methods for supervised learning, Appl. Artif. Intell. 17 (2003) 519–533.
[33] M.M. Suarez-Alvarez, D.-T. Pham, Y. Mikhail, Y.I. Prostov, Statistical approach to normalization of feature vectors and clustering of mixed datasets, Proc. Roy. Soc. London A: Math. Phys. Eng. Sci. 468 (2145) (2012) 2630–2652.
[34] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, J. Roy. Stat. Soc.: Ser. B (Stat. Methodol.) 63 (2) (2001) 411–423.
[35] A.K. Jain, Data clustering: 50 years beyond k-means, Pattern Recogn. Lett. 31 (8) (2010) 651–666.
[36] R. Chauhan, H. Kaur, M. Alam, Data clustering method for discovering clusters in spatial cancer databases, Int. J. Comput. Appl. 10 (6) (2010) 9–14.
[37] S.M. Winkler, M. Affenzeller, H. Stekel, An integrated clustering and classification approach for the analysis of tumor patient data, in: R. Moreno-Daz, F. Pichler, A. Quesada-Arencibia (Eds.), Computer Aided Systems Theory – EUROCAST 2013, Lecture Notes in Computer Science, vol. 8111, Springer, Berlin Heidelberg, 2013, pp. 388–395.
[38] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.
[39] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[40] D. Arthur, S. Vassilvitskii, K-means++: the advantages of careful seeding, in: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '07, 2007, pp. 1027–1035.
[41] S. Dudoit, J. Fridlyand, Bagging to improve the accuracy of a clustering procedure, Bioinformatics 19 (9) (2003) 1090–1099.
[42] S. Vega-Pons, J. Ruiz-Schucloper, A survey of clustering ensembles, Int. J. Pattern Recogn. Artif. Intell. 25 (3) (2011) 337–372.
[43] F. Yang, X. Li, Q. Li, T. Li, Exploring the diversity in cluster ensemble generation: random sampling and random projection, Expert Syst. Appl. 41 (10) (2014) 4844–4866.
[44] Z. Yu, L. Li, H.-S. Wong, J. You, G. Han, Y. Gao, G. Yu, Probabilistic cluster structure ensemble, Inform. Sci. 267 (0) (2014) 16–34.
[45] P.G. de Vries, Stratified random sampling, in: Sampling Theory for Forest Inventory, Springer, Berlin, Heidelberg, 1986, pp. 31–55.
[46] J. Huang, Using auc and accuracy in evaluating learning algorithms, IEEE Trans. Knowl. Data Eng. 17 (3) (2005) 290–310.
[47] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Machine Learning Res. 7 (2006) 1–30.