Knowledge-Based Systems 123 (2017) 75–87

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Synthetic semi-supervised learning in imbalanced domains: Constructing a model for donor-recipient matching in liver transplantation

M. Pérez-Ortiz [a,*], P.A. Gutiérrez [b], M.D. Ayllón-Terán [c], N. Heaton [d], R. Ciria [c], J. Briceño [c], C. Hervás-Martínez [b]

[a] Department of Quantitative Methods, Universidad Loyola Andalucía, Third Building, C/ Escritor Castilla Aguayo 4, Córdoba 14004, Spain
[b] Department of Computer Science and Numerical Analysis, University of Córdoba, Córdoba, Spain
[c] Liver Transplantation Unit, Reina Sofía Hospital, Córdoba, Spain
[d] Liver Transplantation Unit, King's College, London, United Kingdom

ARTICLE INFO

Article history:
Received 23 June 2016
Revised 13 February 2017
Accepted 14 February 2017
Available online 22 February 2017

Keywords:
Liver transplantation
Transplant recipient
Survival analysis
Machine learning
Support vector machines
Semi-supervised learning
Imbalanced classification

ABSTRACT

Liver transplantation is a promising and widely-accepted treatment for patients with terminal liver disease. However, transplantation is restricted by the lack of suitable donors, resulting in significant waiting list deaths. This paper proposes a novel donor-recipient allocation system that uses machine learning to predict graft survival after transplantation using a dataset comprised of donor-recipient pairs from the King's College Hospital (United Kingdom). The main novelty of the system is that it tackles the imbalanced nature of the dataset by considering semi-supervised learning, analysing its potential for obtaining more robust and equitable models in liver transplantation. We propose two different sources of unsupervised data for this specific problem (recent transplants and virtual donor-recipient pairs) and two methods for using these data during model construction (a semi-supervised algorithm and a label propagation scheme). The virtual pairs and the label propagation method are shown to alleviate the imbalanced distribution. The results of our experiments show that the use of synthetic and real unsupervised information helps to improve and stabilise the performance of the model and leads to fairer decisions with respect to the use of only supervised data. Moreover, the best model is combined with the Model for End-stage Liver Disease score (MELD), which is at the moment the most popular assignation methodology worldwide. By doing this, our decision-support system considers both the compatibility of the donor and the recipient (by our prediction system) and the recipient severity (via the MELD score), supporting then the principles of fairness and benefit.

1. Introduction

In the last decades, new trends in biomedicine have used machine learning as a useful tool for a wide range of problems, resulting in remarkable applications for science [32,33]. Nowadays, liver transplantation represents a promising and accepted treatment for patients with end-stage liver disease. Nevertheless, transplantation is greatly hampered by the unavailability of suitable liver donors. Several methods have been developed and applied to find a better allocation system, able to prioritise recipients on the waiting list.

The first developed system for this purpose is the Donor Risk Index (DRI) [12], that establishes the quantitative risk of the transplant considering only donor information. On the other hand, the Model for End-stage Liver Disease (MELD) [20] is a widely validated methodology, globally considered as the cornerstone of the current policy for transplant allocation. This index is based on the "sickest-first" principle and uses only information of the recipient. Fig. 1 graphically represents the current process for organ allocation (figure restructured from [28]). Note that computational models are used for this purpose as a decision support system. As previously mentioned, donors are generally assigned to the candidates at greatest-risk (computed by the MELD score), a policy that does not allow the transplant team to do the matching according to the principles of fairness and survival benefit (i.e. pre-transplant and post-transplant mortality), which could lead to a risk of unconscious gaming when trying to match marginal donors to urgent candidates [25]. The method proposed here for organ allocation seeks to minimize futile liver transplantation, giving primary at-

* Corresponding author.
E-mail addresses: mariaperez@uloyola.es, i82perom@uco.es (M. Pérez-Ortiz), pagutierrez@uco.es (P.A. Gutiérrez), chervas@uco.es (C. Hervás-Martínez).

http://dx.doi.org/10.1016/j.knosys.2017.02.020
0950-7051/

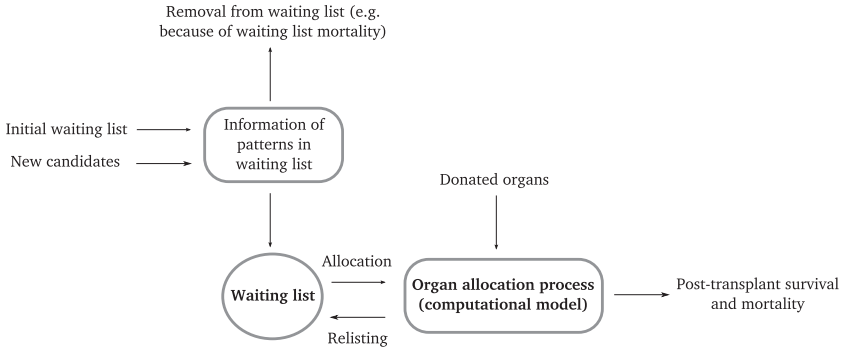

Fig. 1. Graphic representing the organ allocation process.

tention to patients with the best predicted lifetime gained due to transplantation. Under a survival benefit model, an allocated graft goes to the patient with the greatest difference between the predicted post-transplant life-time and the predicted waiting list life-time for this specific donor.

As shown in previous research, there are different donor characteristics which result in an increased risk and/or graft losses [4]. These risks and characteristics should be carefully considered and included in the decision support system, since the combination of several of these risk factors can result in graft loss [3]. Moreover, it has been noted that there are some factors (concerning both the donor and the recipient) that influence the outcome of transplantation to a great extent [25]. Because of this, these first approaches cannot be considered good predictors of graft failure after transplantation, since they only take into consideration either characteristics of donors or recipients (but not both). New methods have emerged in the last years to deal with these issues. Rana et al. [27] devised a scoring system (named SOFT) that predicts survival 3 months after liver transplantation, to complement MELD waiting list mortality by making use of both donor and recipient characteristics. Dutkowski et al. [11] recently proposed a balance of risk (BAR) score based on donor and recipient characteristics. Finally, in [2,10], a rule-based system was developed to determine graft survival one year after the transplant using data from different Spanish liver transplantation units, showing that the use of machine learning substantially improves the prediction capabilities of all previous approaches. In this case, the input of this rule-based system was the response of two artificial neural networks trained with donor, recipient and transplant organ characteristics (all of these sources of information being used in this paper) and using evolutionary computation. One of the main limitations of the approaches developed in [2,10] is that, in order to approach the imbalanced nature of the data, specific fitness functions are applied for tuning the neural network weights and structure through the use of multi-objective evolutionary algorithms, thus the corresponding computational cost is very high.

Although the good performance of machine learning methods has been assessed for donor-recipient matching, the imbalanced nature of the data is still a handicap, as the results for the minority class tend to be worse (with respect to the majority one). Class imbalance is indeed one of the most common problems found in medical applications (and also in machine learning in general [15,24]), where one or several classes have a much lower prior probability in the training set (in the context of this paper the less frequent class is graft loss, although correctly predicting a failure is the main objective). This fact needs to be taken into account in the model construction phase, because, otherwise, one could obtain accurate but trivial models (i.e. that always predict the majority class). The approaches developed over the years for tackling the class imbalance problem can be categorised in two groups: sampling [7,15] and algorithmic approaches [6]. Sampling concerns those methods that rely on a modification of the dataset (e.g. by over-sampling new data or by under-sampling) and algorithmic approaches modify the classifier (e.g. using a cost-sensitive method). Although both over-sampling and under-sampling approaches have been shown to improve classifier performance over imbalanced datasets, it has been shown in different studies that over-sampling is more useful than under-sampling [18], especially for highly imbalanced and small datasets. Concerning cost-sensitive approaches, several works have shown that a replication of data or an imposition of higher weights for some patterns could result in over-fitting [14,26].

In this paper, our main focus is to develop different strategies to improve the classification of the minority class, based on simpler implementations than the ones used in previous research [2,10]. At the same time, we evaluate the applicability of this strategy to other transplant units, by considering a liver transplant dataset obtained from the King's College Hospital in the United Kingdom. Specifically, we tackle the imbalanced nature of the dataset by taking advantage of virtual donor-recipient pairs to improve the accuracy on the minority class. This new perspective for alleviating the imbalance problem in transplantation datasets is based on the use of semi-supervised learning. Important unsupervised information is available at the hospital and can be introduced during model construction by two ideas: exploiting very recent transplants (those whose follow-up time is not completed) and generating non-real or virtual matchings from other pairs that have been already transplanted (i.e. using potential organ transplantations that could have occurred in the past but did not). Most existing semi-supervised learning methods assume a balance between negative and positive samples in both the labelled and unlabelled data [21]. Unfortunately, the issue of semi-supervised learning with imbalanced data has been barely studied in the literature [17,21,23], only mainly from the under-sampling and ensemble points of view. The proposed unsupervised data generation (virtual or real) can reduce the bias of the obtained classifiers towards the majority class. Although the number of successful techniques to approach class imbalance is large, we compare our proposals to two well-known ideas: over-sampling and cost-sensitive learning [7,36], which are also the techniques that have been seen to perform better with Support Vector Machines [22] (the classification paradigm used in this paper).

In summary, this paper studies different hypotheses concerning imbalanced data in semi-supervised scenarios: (1) whether a large amount of unlabelled data could help tackling the imbalanced classification problem, (2) whether the ratio of positive/negative unlabelled patterns affects the results of the semi-supervised method, and (3) whether it is possible to successfully label unlabelled data and balance the class distribution. Therefore, apart from considering two sources of unlabelled data, this paper also explores