Determination of Risk Factors for Hepatitis C by the Method of Random Forest

Girdhar G Agarwal¹, Ashok K Singh²*, Vimla Venkatesh³ and Neha Wal⁴

¹Department of Statistics, Lucknow University, India
²Department of Microbiology, William F. Harrah College of Hospitality, University of Nevada, Las Vegas, USA
³Department of Microbiology, King George Medical University, India
⁴Department of Microbiology, King George Medical University, India

Abstract

The data mining method of Random Forest is used on a dataset collected in a study of HIV infected patients in Lucknow, India for determination of risk factors for the Hepatitis C virus. The accuracy of prediction the Random Forest model from this study is 98.3%.

Introduction

The estimated prevalence of HCV infection worldwide is 170 million people or 2-3% of the world population, with lowest prevalence (0.01 - 0.1%) in the UK and Scandinavia [1] and the highest (10 - 20%) in Egypt [2]. In the Indian subcontinent, prevalence varies [3]: Sri Lanka 0.16%, India 0.33%, Bangladesh 0.6%, Nepal 0.6%, Afghanistan 1.0%, Bhutan 1.3%, Myanmar 0.34%–2.03%, Pakistan 6%–6.8%, and the prevalence is unknown for Maldives.

Investigated the different genotypes among patients with HCV related chronic liver disease using a dataset collected from a tertiary care hospital in south India during the 2002-2012 decade [4]. HCV genotype 3 and genotype 1 turned out to be the predominant genotypes in the entire Indian sub-continent, with Genotype 4 and genotype 6 showing up in some parts [5] provide a summary of status of HCV in India. Mahajan et al. [6] analyzed a dataset of 8035 patients in India and found that HCV is more common in men, in middle-aged people, rural backgrounds, and low to middle socioeconomic class [7].

Used a dataset collected in 2005-2009 from urban areas in Houston, TX; the subjects were predominantly African American drug users who tested negative for HIV and HBV; Cox proportional hazard regression analyses showed daily drug use via injection to be a significant predictor [8]. Used the decision tree approach from data mining to classify characteristics of the genotype a (1 to 6) and genotype 1b. [9] used binary logistic regression for predicting HCV [10].

Used decision trees to predict success of antiviral therapy in chronic Egyptian patients and found alpha-fetoprotein (AFP) level to be an important predictor [11]. Provide a survey of AI applications in diagnosis of HCV.

In the present article, we use the ensemble method of random forest [12] to determine the risk factors of HCV.

Data

The study was conducted in the Department of Microbiology of King George’s Medical University, Lucknow. This is a tertiary care teaching hospital located in the capital city of India’s most populous state - Uttar Pradesh (UP). The hospital caters to the poor and seriously ill patients from the city and several surrounding districts. Patients are mostly referred from primary health centers, district hospitals, nursing homes, private doctors and community health workers. The hospital has an Integrated Counselling and Testing Centre (ICTC) facility which is well attended and provides counselling and HIV testing and linkages for medical and psychosocial care for persons living with HIV infection. The hospital also has an active Anti-Retroviral Therapy (ART) Center. Ethics Committee of K.G.M.U. provided ethical clearance for the study. A total of 350 HIV-infected adults attending the ART Centre and the ICTC were enrolled in the study between January 2007 and July 2008, after obtaining informed consent. The subjects were interviewed using a pre-designed proforma. Current clinical symptoms were elicited by direct questioning by the
interviewer, using a checklist. The questions were designed to have yes or no answers. Questions were asked in the local language and responses recorded. Case records at the ART Centre were reviewed to obtain data for documented opportunistic infections in the subject around the interview date.

The dataset used in this study has 350 observations on a total of 90 variables, with presence of HCV as a binary response variable. Twenty five of the potential predictors of HCV have missing values (see Table 1); these variables are not being used in our analyses.

**Methods**

The method of Random Forest is a predictive model that can be used for regression or classification. Random Forest involves building a large number of decision trees, and outputting the mode of the classes predicted by individual trees (for classification), or the mean of predicted values (for regression) obtained for individual trees. The method of Random Forest is known for its high accuracy and efficiency [13-15]. A detailed description of the method of Random Forests is provided in Hastie, [12].

All of the computations reported in this article are done in the statistical software environment R (2017). The association between the response variable and each individual predictor was first tested by the method of chi-square test of independence; in many of the cases, the expected frequencies of several cells turned out to be less than 5, and the p-values for the chi-square test were evaluated by bootstrap [16].

The R-package random Forest was used for fitting a predictive model to the binary response HCV satisfaction score as a function of the selected predictors. The package random Forest outputs ‘Out of Bag’ (i.e., out of the training sample) estimates of prediction accuracy.
as well as a plot showing the importance of predictors in the model. The package was iteratively used by adding and dropping predictors until a final model with good prediction accuracy was obtained. The statistical significance of the model was tested using the method of Evans, Murphy, Holden, and [17].

**Results**

The random forest model was run several times, starting will all potential predictor and eliminating less important predictors until the final model was found. The statistical significance of the final model was then tested using the R-function rf.significance of the R-package rfUtilities developed by [18]. The overall accuracy of the final random forest model turned out to be 98.3%. The significance test using 1000 permutations for the fitted random forest model yielded a P-value of 0.000, which implies that the random forest model significantly fits the data; the out-of-bag (OOB) accuracy obtained from running the significance test is 98%.

Figure 1 is the variance importance plot of the final random forest model fitted to the dataset. It can be seen from Figure 2 that the important predictors include some of the known risk factors: Age, Jaundice, TFI, Depression, Education, IDU, BLDTR, Mdtrsb, HVW, Gender, infstp, Hdach, Unprs, mighpa, Wkns, freinj, Drguse, and VCPDC (see Table 2 for a brief description of these predictors).

All of the above predictors except Age and TFI are categorical. The chi-square test of independence was run for the response variable HCV and each of the above categorical predictors (Table 3). It can be seen from Table 3 that Jaundice, Depression, IDU, and HVW are significant (Figure 3).

**Discussion of Results**

Logistic regression [12,19] is the most commonly used method for predicting a binary response variable, but when applied to the dataset at hand, it yielded a model in which no predictor was significant and prediction accuracy was low (Figure 4,5). We have shown in this
article that the method of random forest is applicable in such cases, and can be used to predict HCV and also determine the risk factors of a disease.

References


