X CONGRESSO NAZIONALE SISMEC

Società Italiana di Statistica Medica ed Epidemiologia Clinica

Nuovi disegni nella ricerca clinica: la sfida della complessità tra etica e salute

ATTI

sismec

2019 | X CONGRESSO NAZIONALE SISMEC

Nuovi disegni nella ricerca clinica: la sfida della complessità tra etica e salute.

Segretaria di redazione: *Valentina Apostoli*

Edizione: *Marzo 2020*

A cura di: *Consiglio Direttivo SISMEC*

ISBN: 978-88-943456-2-9

sismec

A MACHINE LEARNING APPROACH FOR INSPECTING RELATIONSHIPS BETWEEN PARKINSON DISEASE AND METAL EXPOSURE

Vezzoli Marika[1], Renzetti Stefano[1], Parrinello Giovanni[1], Calza Stefano[1]

1 Department of Translational Medicine, University of Brescia, Italy

Keywords: PD, Ensemble Methods, Random Forest, Variable Importance, Accuracy

Introduction

Parkinson's disease (PD) is a severe neurodegenerative disorder featured by motor dysfunctions such as bradykinesia, tremor, rigidity, as well as other symptoms including chronic fatigue, depression and sleep disturbances impacting on the quality of patients' life. Environmental influence, especially heavy metals, is a solid etiology of PD, widely investigated in many studies.
From a methodological point of view, many researches involving PD deal with fat matrixes (more covariates than subjects), containing variables of different nature (qualitative and quantitative), some of them with a high number of missing values or strongly correlated (multicollinearity).
For analysing this dataset and addressing with the abovementioned problems, we need robust methods able also to model non-linear relationships between the outcome (typically the diagnosis) and a high number of covariates.

Aim

The aim of this research is to model the relationships between the exposure to heavy metals and PD using robust methods belonging to machine learning approach. In detail, we apply the Random Forests [1] algorithm which reduce the instability of single trees [2], providing accurate results in terms of sensitivity and specificity and, more important, identifying, among a huge number of covariates, which are the predictors that have a major impact on the diagnosis.

Methods

Valcamonica is an Italian valley where ferroalloy industries have been active for a century and with an increased prevalence of parkinsonism. In this research, a case-control study containing 169 observations (93 cases, 55.03% and 76 controls, 44.97%) compares patients with PD and control subjects coming from Valcamonica and other areas around Brescia town (Italy). The protocol includes information on age, sex, occupational, residential history, life habits, neuro-psychological testing, and the assessment of genetic polymorphism. Subjects are screened also for serum Cu, Zn, Fe, Mn in blood (MnB) and urine (MnU), transferrin, peroxides, alanine (ALT) and aspartate (AST) transaminases and direct bilirubin. All the PD patients underwent Unified Parkinson's Disease Rating Scale (UDPRS) and metal deterioration battery, namely a neuropsychological assessment using for epidemiological and clinical purposes in Italy. Some genetic analysis was also performed.
Since the dataset is affected by the following problems, (i) high number of missing values, (ii) variables are strongly correlated, and (iii) n<<p (number of observations is less than number of variables), we apply the Random Forests (RF) algorithm for modelling the diagnosis: such an algorithm is indeed an ensemble method extremely flexible and well suited to handle these complications.
RF are an extension of Classification/Regression trees (Breiman et al., 1984) which combine the predictions obtained from many trees grown on perturbed versions of the dataset. RF give also a variable importance measure, the Mean Decrease in Accuracy (MDA), which identifies the covariates which exert the higher impact on the prediction of the outcome (diagnosis). Analytically, within each tree of the RF all the values of the r-th variable are randomly permuted and new predictions are obtained on this new data set (Y, Xr); next,

a loss function Lr is computed and compared with the original loss function of the ensemble L. This procedure is repeated m times on different bootstrap samples and the MDA measure for the r-th variable is given by the following average (av) on m:

$$MDA_r = av_m(L_r - L).$$

The measure obtained is relativized, namely each value is divided by the maximum importance obtained and then multiplied by 100. If the measure assumes a large value (≥ 60), variables have good predictive ability; on the contrary, low values (<60), zero or negative values identify non-informative variables.
By repeating this approach for each variable, we then obtain an importance measure for all potential predictors in the dataset.

Results

In preliminary results, we grow a forest of 10,000 classification trees where we model the diagnosis using heavy metals as covariates. From the forest obtained, we extract the variable importance measure identifying, as determinant of the PD, two heavy metals: copper (CU) and zinc (Zn). These results are in line with other studies that address the same topic ([3], [4]) but based on other methodologies. The main advantage of RF approach is the overperformance of the predictions obtained respect other parametric methods, producing a reliable model.

Conclusions

Compared to a classification tree or to logistic regression, RF increase the accuracy of the predictions by solving the problem of instability (small changes in the data induce big changes in the results) typical of classification trees.

Bibliography

[1] Breiman L., Random Forest. Mach. Learn., 2001; 45(1):5-32.

[2] Breiman L., Friedman J., Olshen R., Stone C., Classification and Regression Trees, Wadsworth Inc., California, 1984.

[3] Lucchini R., Albini E., Benedetti L., et al., Neurological and neuropsychological features in Parkinsonian patients exposed to neurotoxic metals. G Ital Med Lav Ergon, 2007, 29(3 Suppl):280-281.

[4] Squitti R., Gorgone G., Panetta V., et al., Implications of metal exposure and liver function in Parkinsonian patients resident in the vicinities of ferroalloy plants. J Neural Transm, 2009, 116(10):1281-1287.