Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-126

Impact of Imputing Missing Data in Bayesian Network Structure Learning for Obstructive Sleep Apnea Diagnosis

Daniela FERREIRA-SANTOS [a,b,1], Matilde MONTEIRO-SOARES [a,b] and Pedro Pereira RODRIGUES [a,b]
[a] *CINTESIS – Centre for Health Technology and Services Research, Portugal*
[b] *MEDCIDS-FMUP – Faculty of Medicine of the University of Porto, Portugal*

Abstract. Numerous diagnostic decisions are made every day by healthcare professionals. Bayesian networks can provide a useful aid to the process, but learning their structure from data generally requires the absence of missing data, a common problem in medical data. We have studied missing data imputation using a step-wise nearest neighbors' algorithm, which we recommended given its limited impact on the assessed validity of structure learning Bayesian network classifiers for Obstructive Sleep Apnea diagnosis.

Keywords. obstructive sleep apnea, Bayesian network, missing data imputation

1. Introduction

Numerous decisions are made every day by healthcare professionals based in an estimated probability that a specific disease or condition is present. In the diagnostic setting, the probability that a particular disease, such as obstructive sleep apnea (OSA), is present can be used to support further testing, request initiate treatment or reassure patients [1]. OSA is one of the most prevalent sleep disorders, affecting approximately 3-7% of men and 2-5% of women worldwide [2]. Despite its high-frequency, OSA remains underdiagnosed and underestimated, with 75-80% of cases remaining unidentified [2]. Its severity is assessed using the apnea-hypopnea index (AHI), stratifying into mild (5-15), moderate (15-30) and severe (higher than 30). Missing data is a relatively common problem in almost all types of studies, having a significant effect on the conclusions that can be drawn from the data. It is defined as the data value that is not stored for a variable in the observation of interest [1]. Its relevance is such that, when reporting a study development or validation of a diagnosis model, TRIPOD checklist has a specific topic for missing data, where it is mandatory to describe how missing data were handled with details [4]. This work objective was to study the impact of missing data imputation, using nearest neighbors (NN), on structure learning of Bayesian network classifiers for OSA diagnosis.

[1] Corresponding Author: Daniela Ferreira-Santos, CINTESIS and MEDCIDS-FMUP, Rua Dr. Plácido da Costa, s/n 4200-450 Porto, Portugal, E-mail: danielasantos@med.up.pt

2. Methods

2.1. Patients

We have included all patients that performed polysomnography at Vila Nova de Gaia/Espinho hospital center sleep laboratory. All medical and/or sleep laboratory records were retrospectively collected between the 1st of January to the 31st of May 2015. Included patients aged more than 18 years old, while patients already diagnosed, patients with severe lung diseases or neurological conditions and pregnant women were excluded. In case of duplicate exams, the best sleep efficiency was selected.

2.2. Variables and pre-processing

A literature review was previously conducted to define the most relevant OSA variables to be collected from administrative records. A total of 48 variables were collected: **demographic variables:** gender, age; **physical examination:** body mass index (BMI), neck (NC) and abdominal circumferences (AC), modified Mallampati, craniofacial/upper-airway abnormalities; **clinical history:** daytime sleepiness, snoring, witnessed apneas, gasping/choking, sleep fragmentation, non-repair sleep, behavior changes, decrease concentration, morning headaches, decreased libido, body position, sleep efficiency, vehicle crashes, drivers, driving sleepiness, nocturia, alcohol consumption, smoking, coffee, sedatives, family history/genetics, Epworth somnolence scale (ESS); **comorbidities:** atrial fibrillation, stroke, myocardial infarction (MI), pulmonary infarction, arterial and pulmonary hypertension, congestive heart failure (CHF), arrhythmias, pacemaker/cardiovector, respiratory alterations,, diabetes, dyslipidemia, renal failure, hypothyroidism, gastroesophageal reflux (GE), insomnia, glaucoma, bariatric surgery, depression/anxiety. The outcome measure was OSA clinical diagnosis, obtained from AHI, categorized into normal (AHI<5) or OSA (AHI≥ 5). We carry out a pre-processing analysis and continuous variables were categorized.

2.3. Imputing missing data

Instead of deleting any case that has missing data, *k*-NN imputation algorithms preserves all cases and replaces the missing data with a value obtained from related cases (*k* similar cases) in the whole set of records [3]. Our strategy followed systematic procedures: a) we observed the percentage of missing data, that ranged from 0% (e.g. gender) to 97% of missing data (e.g. bariatric surgery); b) variables were then ranked for data imputation, starting with outcome-wise statistically significant variables (with no quality problems suspected), followed by the remaining ordered in increasing percentage of missing data; c) 10-nearest neighbors imputation was done for each new included variable; d) odds ratio (OR) were computed to assess the impact of the referred *k*-NN imputation.

2.4. Naïve Bayes and Tree Augmented Naïve Bayes

Globally, a Bayesian network represents a joint distribution of one set of variables, specifying the assumption of independence between them with the interdependence

between variables being represented by a directed acyclic graph. Each variable is represented by a node in the graph, and its dependence on the set of variables is represented by its ascendant nodes. This dependence is represented by a conditional probability table that describes the probability distribution of each variable, given their ascendant variables [5]. Naïve Bayes (NB, which assumes conditional independence among factors) and Tree Augmented Naïve Bayes (TAN, which allows for an optional dependence for each factor) were the Bayesian network classifiers used in this work. Both classifiers structure learning algorithm requires complete cases, so they were built with the imputed dataset, and we assessed also the impact for different number of selected variables. In the first approach, we used the 10 variables that were statistical significant with or without imputation; in the second approach, we augmented the variable set with 6 more variables found significant in the imputed OR calculation.

2.5. Statistical analysis

Variables were selected after performing Chi-square test or Fisher's exact test for categorical variables and student's t-test or Mann-Whitney U test for continuous variables. Variables were selected if presenting an univariate significant association with the outcome, considering a 5% significance level and for which no quality problems were suspected. Model parameters (NB10 and NB16; TAN10 and TAN16) were validated by comparing the AUC in the imputed derivation cohort with those calculated from a leave-one-out, 10 times 2-fold cross validation (for variability assessment with independent training and testing) and the original derivation cohort. We used R software for: a) missing data analysis; b) imputing missing data (package DMwr); c) descriptive and comparative analysis (packages gmodels and epitools); and d) analyzing AUC (package pROC).

3. Results

In the 318 patients included, 211 had OSA. In total, we had 198 males (62%, crude and imputed OR 2.58 [1.65-4.30)), where 148 (70%) had OSA. In 211 patients with OSA, 115 (55%) were categorized as mild, 50 (24%) as moderate and 46 (22%) as severe. Participants had a mean age of 58 (49-67) years old, being higher in the OSA group (61 (53-68), p value ≤ 0.001); age above 45 years presented a higher risk for OSA (crude and imputed OR 3.29 [1.80-6.72]) and 5.93 [3.02-13.44], respectively). BMI median value was 29 (27-32) Kg/m^2; when categorized into normal and obese, we observed higher number of obese patients in the OSA group (83 (54%), crude OR 2.18 [1.30-3.97], imputed OR 1.91 [1.19-3.30]). NC and AC had a mean of 42 (39-44) cm and 106 (100-113) cm in the OSA group, with AC not having statistical significance (p value 0.052). Crude OR of modified Mallampati in the category 4 was significant (4.50 [1.09-38.83]) but when imputing (32% of missing data) it lost significance (3.36 [0.83-27.75]). The same occurred with nocturia (crude OR 2.05 [1.13-4.16], imputed OR 1.32 [0.79-2.40]). In those with craniofacial/upper airway abnormalities we discovered higher number of patients in OSA group (64 (82%), crude OR 1.24 [0.57-3.26], imputed OR 1.31 [0.77-2.42]), without statistical significance; other variables such as snoring, drivers, smoking, use of sedatives, sleep efficiency, gasping/choking, respiratory changes, sleep fragmentation, MI, pulmonary and arterial hypertension, dyslipidemia, anxiety or

depression, pacemaker or cardiovector, vehicle crashes, genetics/family history, hypothyroidism, renal failure, stroke, decreased libido and concentration, behavior changes, pulmonary infarction, glaucoma and bariatric surgery had no statistical significance, also. Subjects in-taking coffee had a higher risk of OSA (133 (86%)) with statistical significance in the imputed OR (0.48 [0.27-0.96]). The same effect was described for CHF, arrhythmias, diabetes, GE and insomnia. OSA group was a higher number of patients with witnessed apneas (109 (64%), crude OR 1.92 [1.18-3.34], imputed OR 2.14 [1.37-3.53]). Additionally, with statistical significance in crude and imputed OR, we found non-repairing sleep, morning headaches, driving sleepiness, alcohol consumption and body position. Daytime sleepiness and ESS presented contradictory results raising data collection quality suspicion, and were not considered for analysis. Impact of imputing missing data was assessed with ROC curves for each model, along with their 95% confidence interval (CI), being presented in Figure 1, demonstrating imputed and original in-sample AUC. Imputed leave-one-out and 10 times 2-fold cross validation values are presented in Table 1. Specific cut-off values were chosen after assessing the AUC of the imputed derivation cohort, aiming at a sensitivity of 95%, to allow a rule-out approach aiming to avoid false negatives. The AUC values of the original and imputed derivation cohort in the four models overlapped, as did imputed leave-one-out and cross-validation.

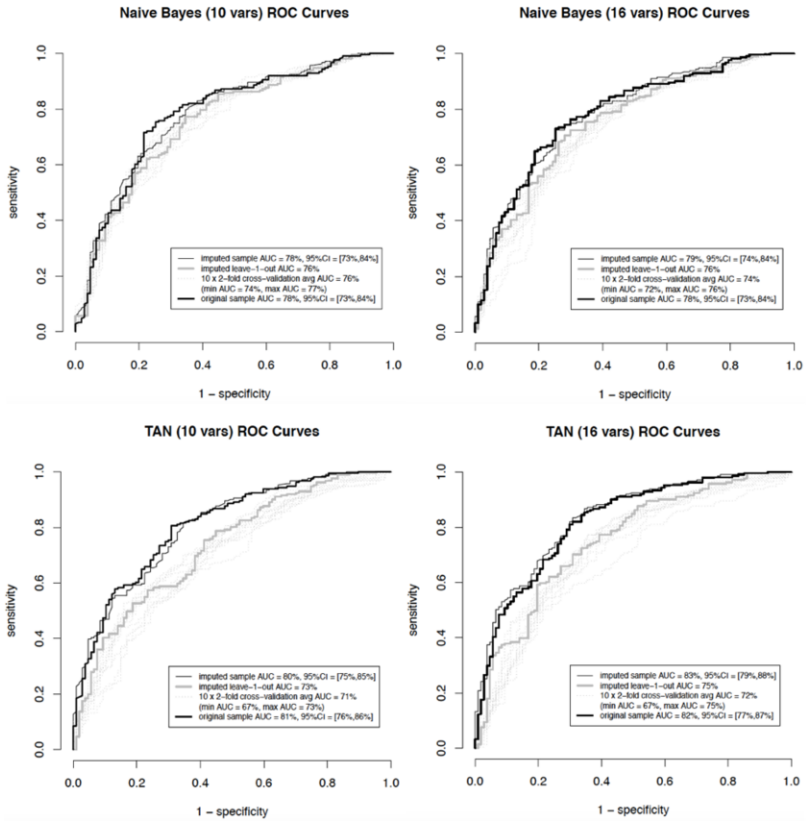

Figure 1. Receiver operating characteristics analyses and area under the curve values for NB10, TAN10, NB16 and TAN16, as well as for the internal validation procedures and the original derivation cohort.

Table 1. Validity assessment [%] estimated from 10 times 2-fold cross validation.

Model	Cut	Accuracy [CI 95%]	Sensitivity [CI 95%]	Specificity [CI 95%]	Precision (+) [CI 95%]	Precision (-) [CI 95%]
NB10	19.56%	70.85% [69.63,72.07]	95.07% [93.72,96.42]	23.09% [21.01,25.17]	70.91% [70.23,71.59]	71.69% [65.29,78.08]
TAN10	34.18%	68.71% [67.51,69.91]	89.05% [87.31,90.80]	28.6% [26.19,31.00]	71.1% [70.41,71.8]	57.70% [53.78,61.61]
NB16	13.17%	70.79% [70.00,71.57]	94.36% [93.00,95.72]	24.29% [22.44,26.13]	71.09% [70.67,71.51]	70.20% [65.62,74.78]
TAN16	23.61%	69.78% [68.39,71.17]	90.33% [88.55,92.11]	29.27% [26.36,32.18]	71.6% [70.73,72.47]	61.48% [56.51,66.46]

NB10, NB16: Naïve Bayes with 10 or 16 variables; **TAN10, TAN16:** Tree Augmented Naïve Bayes with 10 or 16 variables

4. Discussion and Conclusion

The occurrence of missing data is a major concern in several areas, including medical domains such as OSA diagnosis. The work of Hernández-Pereira *et al.* [6] tried to improve detection of apneic events by treating missing data; however, it only addresses numeric values. We have proposed a step-wise *k*-NN imputation approach (instead of more common list-wise deletion), proving to be a far better and valuable solution, with limited impact in structure learning Bayesian network classifiers. Main advantages include: a) imputed values are actually occurring values and not constructed values; b) it makes use of auxiliary information provided by the independent variables, preserving thus the original structure of the data; and c) it is fully non-parametric and does not require explicit models to relate factors and outcomes, being thus less prone to model misspecification.

Acknowledgments

This work has been developed under the scope of project NanoSTIMA [NORTE-01-0145-FEDER-000016], which was financed by the North Portugal Regional Operational Programme [NORTE 2020], under the PORTUGAL 2020 Partner- ship Agreement, and through the European Regional Development Fund [ERDF].

References

[1] H. Kang, The prevention and handling of the missing data, Korean J. Anesthesiol., vol. 64, n. 5, pp. 402-406, 2013.
[2] N. M. Punjabi, The epidemiology of adult obstructive sleep apnea, Proc. Am. Thorac. Soc., vol. 5, n. 2, pp. 136-143, 2008.
[3] L. Beretta e A. Santaniello, Nearest neighbor imputation algorithms: a critical evaluation, BMC Med. Inform. Decis. Mak., vol. 16, n. S3, p. 74, 2016.
[4] G. S. Collins, J. B. Reitsma, D. G. Altman, e K. G. M. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement, Eur. Urol., vol. 67, n. 6, pp. 1142-1151, 2015.
[5] T. Mitchell, Machine Learning. McGraw-Hill, 1997.
[6] E. Hernández-Pereira, D. Álvarez-Estévez, e V. Moret-Bonillo, Improving detection of apneic events by learning from examples and treatment of missing data, Stud. Health Technol. Inform., vol. 207, pp. 213-224, 2014.