J. Mantas et al. (Eds.)

doi:10.3233/978-1-61499-921-8-75

Phenotyping Obstructive Sleep Apnea Patients: A First Approach to Cluster Visualization

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

Abstract. The varied phenotypes of obstructive sleep apnea (OSA) poses critical challenges, resulting in missed or delayed diagnosis. In this work, we applied k-modes, aiming to identify groups of OSA patients, based on demographic, physical examination, clinical history, and comorbidities characterization variables (n=41) collected from 318 patients. Missing values were imputed with *k*-nearest neighbours (*k*-NN) and chi-square test was held. Thirteen variables were inserted in cluster analysis, resulting in three clusters. Cluster 1 were middle-aged men, while Cluster 3 were the oldest men and Cluster 2 mainly middle-aged women. Cluster 3 weighted the most, whereas Cluster 1 weighted the least. The same effect was described in increased neck circumference. The percentages of variables driving sleepiness, congestive heart failure, arrhythmias and pulmonary hypertension were very low (<20%) and OSA severity was more common in mild level. Our results suggest that it is possible to phenotype OSA patients in an objective way, as also, different (although not considered innovative) visualizations improve the recognition of this common sleep pathology.

Keywords. Categorical data, cluster analysis, data visualization, obstructive sleep apnea, phenotypes

1. Introduction

In obstructive sleep apnea (OSA), respiratory effort is preserved but ventilation decreases/fades due to partial/total occlusion of the upper airway. A diagnosis is established when a patient has an apnea-hypopnea index (AHI) $\geq$ 5 with associated symptoms or an AHI $\geq$ 15 regardless of associated symptoms [1]. OSA affects about 4% of men and at least 2% of women worldwide [1]; nonetheless the diverse phenotypes of OSA have not yet been formally characterized, posing critical challenges to its clinical recognition, resulting in missed or delayed diagnosis [2]. Nowadays, cluster analysis has been used to identify subtypes of patients who are diagnosed with a particular disorder, such as asthma [3]. The *k*-modes algorithm [4] extends the *k*-means paradigm to cluster categorical data by using a simple matching dissimilarity measure for categorical objects [5], modes instead of means for clustering, and a frequency-based method to update modes in the *k*-means fashion clustering process. The dissimilarity measure can be defined by the total mismatches of the corresponding variable categories of the two

[1] Corresponding Author, Rua Dr. Plácido da Costa, s/n, 4200-450, Porto, Portugal; E-mail: danielasantos@med.up.pt

objects: the smaller the number of mismatches, the more similar the two objects are [5]. As we know, the primary goal of data visualization is to communicate information clearly and efficiently via statistical or information graphics. Each visualization intends to help users analysing and reasoning about data and evidence, that is why our intention was to phenotype OSA patients, applying categorical cluster analysis (*k*-modes) to identify groups, based on risk and diagnostic factors, and performed two different clusters visualizations.

2. Methods

In a previous diagnostic test study [6], we included all patients who undertook polysomnography (PSG) at Vila Nova de Gaia/Espinho hospital centre. All administrative records were retrospectively collected between January-May, 2015 and inclusion criteria was patients aged more than 18 years old; patients already diagnosed, patients with severe lung diseases or neurological conditions and pregnant women were excluded. We performed a pre-processing analysis and continuous variables were categorized; *k*-nearest neighbours (*k*-NN) imputation was conducted aiming to preserve all cases and missing data was replaced with a value obtained from related cases from the complete set of records [7]. A literature review helped define the most relevant OSA variables to be collected, in a total of 41 variables: demographic variables (e.g., gender); physical examination (e.g., body mass index (BMI)); clinical history (e.g., snoring); and comorbidities (e.g., stroke). Our dataset portrayed only categorical variables, which rose a question: what is the best way to distinguish high influence variables from low or no influence variables? Variables were selected, after chi-square analysis, if presenting a univariate significant association with the outcome (AHI), considering a 5% significance level. We used R software to perform descriptive and associative analysis (packages *gmodels* and *epitools*) and *k*-modes categorical clustering (package *klaR*) and to create standard barplot (*ggplot2*) and heatmap (*gplots*).

3. Results

From the 318 patients covered, 211 had OSA (66%); from which, 115 (55%) were categorized as mild, 50 (24%) as moderate and 46 (22%) as severe. In total, we had 148 males (70%) with OSA, presenting a mean age of 61 (53-68) years old; the category 20-44 presented a lower percentage of patients in the OSA group, oppositely to categories 45-64 and 65-90 (p value <0.001). Focusing our attention only on the group with the pathology, BMI median value was 30 (27-30) kg/m^2 (p value 0.008); neck circumference (NC) and abdominal circumference (AC) had a mean of 42 (39-44) cm and 107 (100-113) cm. Modified Mallampati categories showed us that our patients have higher total percentages in the inferior levels (1 and 2 in a total of 64%). In those with craniofacial/upper airway abnormalities (CFA) we discovered higher number of patients in the OSA group (195 (92%)) without statistical significance; other variables such as snoring, nocturia, sleep fragmentation, insomnia, drivers, family history, myocardial infarction, arterial hypertension, pacemaker, stroke, renal failure, dyslipidaemia, and hypothyroidism had also no statistical significance. When analysing gasping/choking, we noticed a higher percentage of patients in the normal group (54 (50%)), without statistical significance; other variables present the same characteristic: behaviour changes, decreased libido, vehicle crashes, coffee, use of sedatives, respiratory alterations, and anxiety/depression. We included a total of 13 variables in the cluster analysis (gender, age, BMI, NC, modified Mallampati, witnessed apneas, non-repairing sleep, morning headaches, driving sleepiness, alcohol, congestive heart failure, arrhythmias, and pulmonary hypertension). Three distinct clusters were identified.

Table 1. Clinical characteristics of the cohort by the defined clusters, % [95%CI]

	Cluster 1	**Cluster 2**	**Cluster 3**	**p-value**	**P(C|F)**
	(n=122)	**(n=44)**	**(n=55)**		**(c_1, c_2, c_3)**
Male gender	85 [77-91]	20 [10-36]	80 [67-89]	**<0.001**	(.64, .06, .30)
Age				**<0.001***	
20-44	6 [3-13]	11 [4-25]	11 [5-23]		(.39, .28, .33)
45-64	65 [56-74]	57 [41-71]	27 [14-41]		(.65, .22, .33)
65-90	29 [21-38]	32 [19-48]	62 [48-74]		(.40, .18, .42)
Obese	21 [14-29]	43 [29-59]	76 [63-86]	**<0.001**	(.27, .23, .50)
Increased NC	30 [22-40]	77 [62-88]	96 [86-99]	**<0.001**	(.28, .28, .44)
Mallampati				**<0.001***	
1	22 [15-31]	11 [4-25]	36 [24-50]		(.50, .10, .40)
2	48 [39-58]	61 [46-75]	5 [1-16]		(.64, .32, .04)
3	26 [18-35]	23 [12-38]	47 [34-61]		(.45, .15, .40)
4	4 [1-9]	5 [1-17]	11 [5-23]		(.33, .17, .50)
Witnessed apneas	64 [55-73]	30 [17-45]	75 [61-85]	**<0.001**	(.57, .10, .33)
Non-repairing sleep	46 [36-55]	70 [55-83]	29 [18-43]	**<0.001**	(.52, .32, .16)
Morning headaches	45 [35-54]	84 [69-93]	18 [10-31]	**<0.001**	(.52, .38, .10)
Driving sleepiness	13 [7-20]	5 [1-17]	2 [0-11]	**0.045**	(.82, .12, .06)
Alcohol	85 [77-91]	20 [10-36]	91 [79-97]	**<0.001**	(.62, .06, .32)
Cong. Heart Fail.	10 [5-17]	14 [6-28]	18 [10-31]	0.309	(.41, .22, .37)
Arrhythmias	5 [2-12]	14 [6-28]	11 [5-23]	0.153	(.34, .33, .33)
Pulm. hypertension	4 [1-9]	9 [3-23]	13 [6-25]	0.073	(.27, .27, .46)
OSA				0.495	
Mild	56 [47-66]	61 [46-75]	45 [32-59]		(.55, .23, .22)
Moderate	24 [17-33]	20 [10-36]	25 [15-39]		(.54, .18, .28)
Severe	20 [13-28]	18 [9-33]	29 [18-43]		(.48, .17, .35)

*Fisher's exact test; CI: confidence interval; P(C|F): probability of belonging to a cluster given the presence of a factor; NC: neck circumference; Cong. Heart Fail.: congestive heart failure; Pulm. Hypertension: pulmonary hypertension

Table 1 summarizes the clinical characteristics of the total cohort in the selected variables by cluster. Figure 1 and 2 visually synthesizes the information obtained from Clusters 1 to 3. As shown, patients in Cluster 1 were middle-aged men, weighted the least, with non-increased NC and lower percentages in the lowest levels of modified Mallampati. This cluster had the second higher percentage of witnessed apneas, non-repairing sleep, morning headaches, and alcohol consumption. Driving sleepiness had the high percentage in this cluster; oppositely to congestive heart failure, arrhythmias and pulmonary hypertension. These patients presented a severity category of mild, the second higher percentage. Cluster 2 was comprised mostly of middle-aged women, with the second higher percentage of increased NC and BMI. Most of the patients had a level 2 modified Mallampati. This cluster had the lowest percentage of witnessed apneas, alcohol consumption, and very low percentages of comorbidities. Regarding OSA severity they were also mostly categorized in the mild stage. Cluster 3 had the oldest men, weighted the most and presented an increased NC in 96% of the patients. This cluster showed a higher percentage of the modified Mallampati in the higher levels (3 and 4), and a higher number of patients reporting witnessed apneas. Regarding non-repairing sleep, morning headaches, and driving sleepiness, Cluster 3 had the lower percentages, in contrast with alcohol consumption, congestive heart failure, and pulmonary hypertension. This cluster presented a lower value in the mild severity category and a higher value in the severe level.

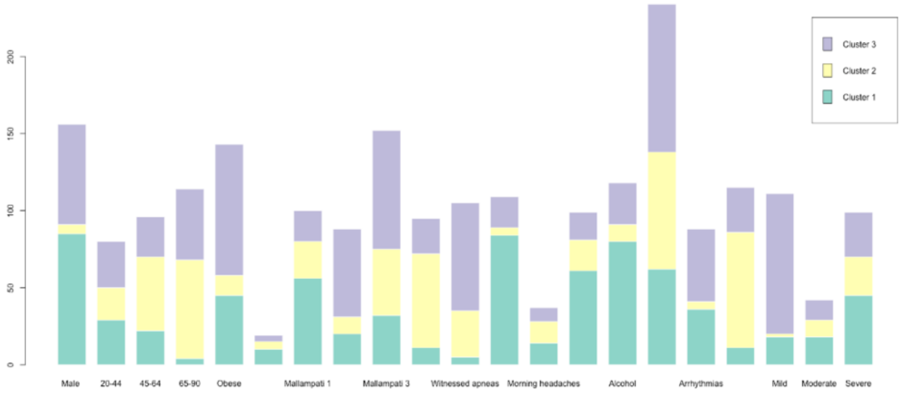

Figure 1. Clinical characteristics of the cohort in Cluster 1, 2 and 3

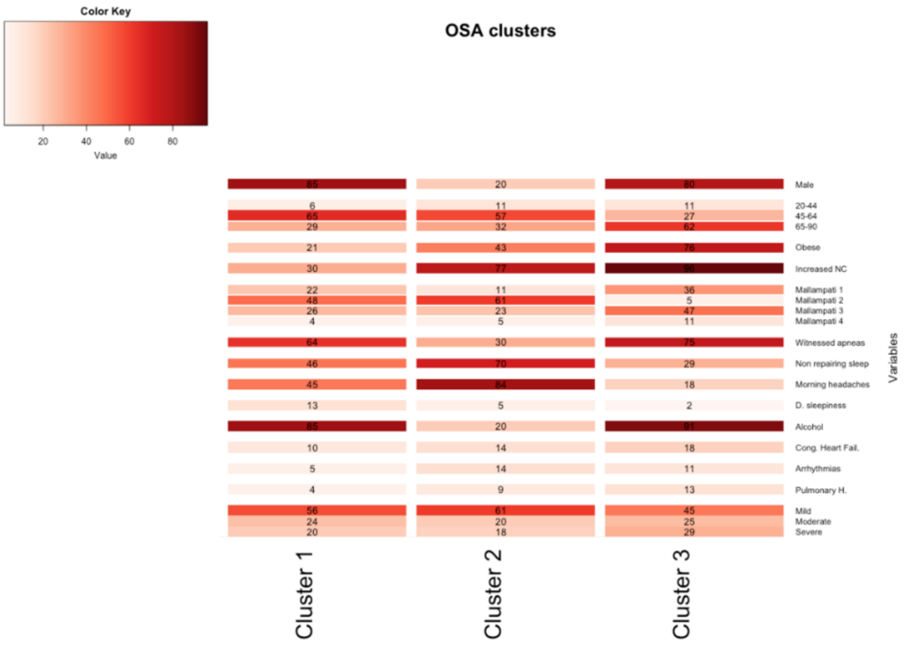

Figure 2. Percentages of each clinical characteristics by cluster

4. Discussion and Conclusion

Understanding different phenotypes in OSA diagnosis is particularly important. Patients in Cluster 1 and 3 were males with different categorized age (middle-aged adults vs.

elderly). Analysing physical examination aspects, we verified that Cluster 1 had lower BMI and NC, oppositely to Cluster 2 and 3. Regarding clinical history, the overall percentage of modified Mallampati was in the lower levels. Cluster 2 reported lower percentage of witnessed apneas, but higher values in non-repairing sleep and morning headaches. The global percentages of driving sleepiness and all selected comorbidities (congestive heart failure, arrhythmias, and pulmonary hypertension) were very low and without statistical significance. Moreover, the percentage of the outcome measure was demonstrated in each cluster; Cluster 2 had 61% [46%-75%] of mild severity, followed by Cluster 1 (56% [47%-66%]) and Cluster 3, with significantly smaller proportion (45% [32%-59%]). To the best of our knowledge, this is the first attempt to phenotype OSA patients using k-modes categorical clustering. Our results suggest that there are different clinical sub-types of OSA, helping focus our attention on a detailed description of OSA diagnosis. The major strengths of this study are newly data analysis, applying standard visualizations to the data, and the clinical cohort representing OSA patients (all levels of severity) that performed PSG. Furthermore, the inclusion of a comprehensive number of risk and diagnostic factors enhances our understanding of OSA diagnosis. This first approach to cluster visualization only intended to summarize the available options and prepare the path for future and more complex visualizations, like circle packing or sunburst.

Acknowledgements

DFS acknowledges Fundação para a Ciência e Tecnologia (FCT) under grant number PD/BD/13553/2018. The work has been developed under the scope of project NORTE-01-0145- FEDER-000016 (NanoSTIMA), financed by the North Portugal Regional Operational Programme (NORTE-2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

References

[1] American Academy of Sleep Medicine, Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research. The report of an American Academy of Sleep Medicine Task Force, *Sleep* **22** (1999), 667–689.

[2] L. Ye, G. Pien, S. Ratcliffe, E. Björnsdottir, E. Arnardottir, A. Pack, B. Benediktsdottir, T. Gislason, The different clinical faces of obstructive sleep apnoea: A cluster analysis, *European Respiratory Journal* **44** (2014), 1600–1607. doi:10.1183/09031936.00032314

[3] W. Moore, D. Meyers, S. Wenzel, W. Teague, H. Li, X. Li, R. D'Agostino, M. Castro, D. Curran-Everett, A. Fitzpatrick, B. Gaston, N. Jarjour, R. Sorkness, W. Calhoun, K. Chung, S. Comhair, R. Dweik, E. Israel, S. Peters, W. Busse, S. Erzurum, E. Bleecker, Identification of asthma phenotypes using cluster analysis in the severe asthma research program, *American Journal of Respiratory and Critical Care Medicine* **181** (2010), 315–323.

[4] Z. Huang, A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining, *Research Issues on Data Mining and Knowledge Discovery* (1997), 1–8. doi:10.1.1.6.4718

[5] L. Kaufman, P. J. Rousseuw, Finding Groups in Data: An Introduction to Cluster Analysis, *Wiley* **47** (1990), 788. doi:10.2307/2532178

[6] D. Ferreira-Santos, P. P. Rodrigues, Improving Diagnosis in Obstructive Sleep Apnea with Clinical Data: A Bayesian Network Approach, *2017 IEEE 30th International Symposium on Computer-Based Medical Systems* (2017), 612–617. doi:10.1109/CBMS.2017.19

[7] D. Ferreira-Santos, M. Monteiro-Soares, P. P. Rodrigues, Impact of Imputing Missing Data in Bayesian Network Structure Learning for Obstructive Sleep Apnea Diagnosis, *Studies in Health Technology and Informatics* **247** (2018), 126–130.