A Conditional Mutual Information-based Feature Selection Method for Gender Classification

Marta Iovino
Department of Engineering
University of Palermo
Palermo, Italy
marta.iovino@unipa.it

Ivan Lazic
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
ivan.lazic@uns.ac.rs

Chiara Barà
Department of Engineering
University of Palermo
Palermo, Italy
chiara.bara@unipa.it

Luca Faes
Department of Engineering
University of Palermo
Palermo, Italy
luca.faes@unipa.it

Riccardo Pernice
Department of Engineering
University of Palermo
Palermo, Italy
riccardo.pernice@unipa.it

***Abstract*—This study proposes a feature selection approach exploiting Conditional Mutual Information to identify the most relevant features to perform gender classification. The approach, applied with features extracted from cardiovascular time series, is combined with a Linear Discriminant Analysis classifier. The feature selection method allowed to noticeably reduce the number of used features, achieving at the same time comparable and acceptable accuracy (around 62%) and overall good recall and F1-scores for females (around 71% and 63%, respectively).**

***Index Terms*—Gender Classification, Information Theory, Cardiovascular Time Series**

I. Introduction

The topic of sex differences in cardiovascular regulation, marked by considerable debate and conflicting findings [1], remains elusive due to the mixed scenery presented in studies exploring heart rate variability (HRV) and blood pressure variability (BPV). Although some research works suggest increased HRV in females than in males, the results are still conflicting and there is also a lack of information regarding sex differences in BPV [1]. When dealing with large and intricate datasets, Artificial Intelligence emerges as a main tool to extract useful information. In this context, Feature Selection (FS), also known as variable elimination, has been proven to improve classification model performance by identifying relevant variables [2]. Information theory (IT) is widely used for feature selection criteria, but current definitions of feature relevancy and redundancy often overlook interactions among features which can affect the selection procedures [3]. This limitation stems from the definition of mutual information (MI), which fails in fully describing interactions between multiple variables but can become possible through the use of Conditional Mutual Information (CMI) [3].

In this context, we propose an approach for the FS phase for gender classification using CMI to determine the most informative features from cardiovascular time series. The main aim of this work is to demonstrate that CMI-based FS, combined with the Linear Discriminant Analysis (LDA) classifier [4], helps in differentiating physiological regulation between males and females.

II. Materials and Methods

A. Description of the proposed FS approach

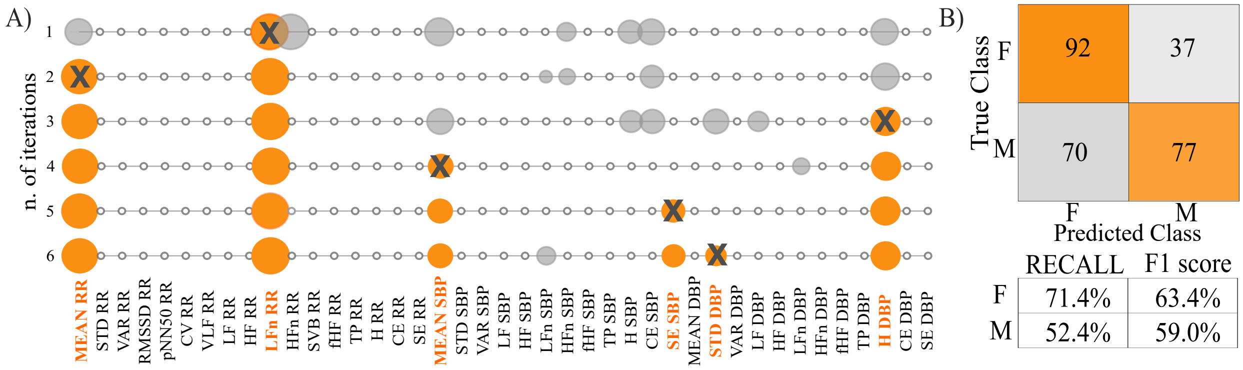
In the IT framework, the CMI quantifies the shared information between a single variable $X_i \in \mathbf{X}$ and a target variable Y when a variable $Xj \in \mathbf{X} \backslash X_i$ is known. Specifically, CMI is computed as $I(Y; X_i|\mathbf{X}_j) = I(Y; \mathbf{X}_j, X_i) - I(Y; X_i)$. To perform CMI-based FS, we propose the use of an iterative forward-selection algorithm which starts by calculating the MI as $I(Y; X_i)$ and gradually increases the number of selected features until the stopping criterion is reached [2]. The significance of the MI is evaluated through surrogate data generated by randomly permuting each observed feature, with a 95% statistical threshold. The feature with the highest MI value above this threshold is considered the most informative and is added to the selected feature set. Subsequently, the CMI values of each of the remaining features conditioned to the previously selected feature are calculated, according to the same thresholding and selection procedure. This selection continues until the computed CMI values are not significant, ensuring that each selected feature contributes significantly to the information about the target variable.

The Linear Discriminant Analysis (LDA) classifier was used to validate the results obtained with the FS phase to discriminate between genders. The LDA is a supervised linear model where decision boundaries are defined by (D-1)-dimensional hyperplanes within the D-dimensional input space, indicating linear separability when these boundaries can precisely separate classes [4]. The objective was to focus on the performances of the proposed FS approach rather than on the characteristics of the classifier itself. For each class (i.e. the gender), the performances of the classifiers were evaluated through accuracy, recall, and F1-score metrics.

B. Application to gender classification

Data used for this work (Healthy Young POLes Database) were collected from 276 young, healthy volunteers (147 F, 23.8±2.6 years old); we refer to [5] for further details. The analysed beat-to-beat cardiovascular time series, recorded during a rest phase in the supine position, consisted of RR intervals (RR), systolic blood pressure (SBP), and diastolic blood pressure (DBP). To adhere to the guidelines of short-term cardiovascular analysis ($\sim$5 min), 300-point time series

Research supported by PRIN 2022 project "HONEST" (funded by MUR, code 2022YMHNPY, CUP B53D23003020006). I.L. supported by the Faculty of Technical Sciences, University of Novi Sad, project No. 01-3394/1. R.P. supported by the European Social Fund POC 2014/2020 of Sicily Region.



	RECALL	F1 score
F	71.4%	63.4%
M	52.4%	59.0%

Fig. 1. (A) Schematic representation of the FS algorithm procedure and results for each iteration. Grey bullets: features with significant MI or CMI values. Bold X: specific feature selected in a given iteration. Orange bullets: all the features selected for each iteration. In orange: features selected after all the steps. The size of the bullets is proportional to the value of MI or CMI. (B) Confusion matrix and tables of the classification performance results (Recall and F1-score) using only the selected features.

were considered, discarding the first 60 beats. Forty-one features were computed for each cardiovascular time series, divided into three domains (time, frequency, and information). The following time domain indices were calculated: mean (MEAN), standard deviation (STD), and variance (VAR). In the frequency domain, the absolute spectral power values were calculated in Low Frequency (LF, 0.04-0.15 Hz), High Frequency (HF, 0.15-0.4 Hz), and Total Power (TP, 0-0.4 Hz) bands [6]. The normalised power values within the LF and HF bands ($\mathrm{LF_n}$ and $\mathrm{HF_n}$) were also computed, alongside the respiratory peak frequency (as the peak frequency in the HF band, $\mathrm{f_{HF}}$) [6]. In the information domain, the entropy (H), the conditional entropy (CE), and the self-entropy (SE) were calculated [7]. For the RR time series, the root mean square of successive differences (RMSSD), the percentage of successive RR intervals (pNN50), the coefficient of variation (CV) computed as the STD to MEAN ratio, the absolute spectral power in the Very Low Frequency (VLF, 0-0.04 Hz) band and the sympathovagal balance index (SVB), were calculated in addition to the previous features [6].

Herein, the CMI was estimated using the k-Nearest Neighbours approach with $k = 10$ neighbours [7], and computed between the above-indicated features, representing the continuous variables $\boldsymbol{X}$ (as mentioned in Section II.A), and the gender (F, M) being the discrete target variable Y. The LDA classifier was trained using only the selected features, and evaluated over a 10-fold cross-validation for gender discrimination.

III. Results and Discussion

Results shown in Fig. 1(A) suggest that the feature selection algorithm can identify an initial set of relevant features reaching the stopping criterion after six iterations. In the first iteration, only 8 features (out of the 41 total) achieved significant MI values (grey bullets in step 1 of Fig. 1(A)), of which the $\mathrm{LF_n}$-RR feature exhibited the maximum value (orange bullet and bold X). At the end of all the iterations, a total of 6 significant features were selected: $\mathrm{LF_n}$, RR, MEAN RR, H DBP, MEAN SBP, SE SBP, and STD DBP, all with a widely known physiological meaning [6]. Some features remained statistically significant over multiple iterations, indicating their robustness in contributing to the feature set. Recall and F1-score results, reported in Fig. 1(B), proved an overall good ability of the classifier to discriminate between genders, with relatively high values for females (∼71%, ∼63% respectively), despite the sub-optimal overall classification accuracy (∼62%) that is comparable to value achieved obtained using all the 41 features (∼61%). Overall, our findings suggest the capability of the selected features to distinguish between genders.

IV. Conclusion

The study presented an approach employing Conditional Mutual Information for feature selection, suggesting promising results for gender classification. While preliminary, our results highlight the discriminatory power of the selected features and their ability to achieve overall good accuracy comparable to the entire set. Future steps will involve the use of further features (e.g. bivariate [7]), exploring different information-theoretic approaches, such as the Partial Information Decomposition distinguishing among unique, redundant and synergistic information [3], and a more in-depth investigation of the physiological meaning behind the selected features.

References

[1] J. Koenig *et al.*, "Sex differences in healthy human heart rate variability: A meta-analysis," *Neuroscience & Biobehavioral Reviews*, vol. 64, pp. 288–310, 2016.

[2] G. Chandrashekar *et al.*, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[3] P. Wollstadt *et al.*, "A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition," *Journal of Machine Learning Research*, vol. 24, no. 131, pp. 1–44, 2023.

[4] C. M. Bishop *et al.*, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.

[5] P. Guzik *et al.*, "Healthy young poles–hypol database with synchronised beat-to-beat heart rate and blood pressure signals: Hypol–cardiovascular time series database," *Journal of Medical Science*, vol. 92, no. 4, pp. e941–e941, 2023.

[6] F. Shaffer *et al.*, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, p. 258, 2017.

[7] H. Pinto *et al.*, "Testing dynamic correlations and nonlinearity in bivariate time series through information measures and surrogate data analysis," *Frontiers in Network Physiology*, vol. 4, p. 1385421, 2024.