BABEŞ BOLYAI UNIVERSITY, CLUJ NAPOCA, ROMÂNIA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Individual self-assessment using biometric data

– MIRPR report –

Team members
Vintila Dragos, Computer Science - English, 937, dragos.vintila@stud.ubbcluj.ro
Hatisi Bogdan, Computer Science - Romanian, 233, bogdan.hatisi@stud.ubbcluj.ro

2023-2024

Contents

1 Introduction **1**
1.1 What? Why? How? 1

2 Scientific Problem **2**
2.1 Problem definition 2

3 Investigated approach **3**
3.1 Dataset 3
3.2 Data preprocessing 3
3.3 Model training and evaluation 4

Chapter 1

Introduction

1.1 What? Why? How?

The application collects biometric data, such as heart rate, skin conductance, and temperature, via a wristband worn by the user. It starts by creating a baseline of the individual's physiological parameters before exposing them to various stressors and stimuli while measuring their biometric responses. It then logs and organizes this information, associating it with specific stressors or events. The acquired data is evaluated using an unsupervised learning algorithm and various other methods to showcase different patterns that may be reviewed and analyzed for performance barriers.

Chapter 2

Scientific Problem

2.1 Problem definition

The associative relations between different biometric indicators and stressors can be difficult to correlate. Unsupervised learning algorithms can be used to identify patterns in the data that may be used to improve the user's performance. These help us navigate the ambiguous and inconsistent real-world biometric data, by providing:

- **Pattern Recognition**: Intelligent algorithms, such as unsupervised learning, are capable of detecting complicated patterns and relationships in biometric data that analysts may miss. They can detect tiny differences and correlations that can aid in understanding the user's physiological state.
- **Scalability**: These algorithms are capable of processing and analyzing massive amounts of data, which is frequently the case with biometric data acquired over time. The algorithms can adjust and continue to reveal useful insights as the information develops.
- **Automation** of the data analysis process, making it more efficient and less susceptible to human bias. This automation enables continuous monitoring without the need for human intervention.
- **Anomaly Detection**: Anomalies in data, such as unexpected or potentially harmful physiological responses, can be detected by intelligent algorithms, allowing for appropriate intervention.

By wearable biometric sensors, we acquire different physiological and behavioral signals from the user(heart rate, accelerometer, electro-dermal activity, etc.) during exposure to certain stressors and stimuli. Between these exposures to stress we highlight the typology in which they occur.

Chapter 3

Investigated approach

3.1 Dataset

With the help of the Non-EEG Dataset for Assessment of Neurological Status the algorithm attempts to predict different reaction types experienced by the subject based on previously classified stressor conditions exposed to it.

The dataset consists of 20 subjects, each of which were exposed to the same stressors. Each subject was exposed to 7 successive time intervals of different stressors. These were in the following order: relaxation, physical, recovery, emotional, cognitive, recovery, emotional (horror), recovery.

The data collected consists of electrodermal activity (EDA), temperature, 3-axis acceleration, heart rate (HR), and arterial oxygen levels (SpO2).

3.2 Data preprocessing

The recorded data is stored using the wfdb library. Data from each subject is split into two records, the first containing the SpO2 and HR data and the second containing the acceleration and EDA activity. Since the data in the second record is recorded at 8Hz and the first record at 1Hz, the second record is downsampled to match the first.

The acceleration data is recorded along three axes, with the primary interest lying in the magnitude of the acceleration. This can be expressed using the representative formula: $\sqrt{x^2 + y^2 + z^2}$. Temperature measurements were excluded from the analysis due to their perceived unreliability: specifically, temperature exhibited an initial rise during the early stages of the test and demonstrated minimal variation throughout the entire testing period. Similarly, SpO2 measurements were not utilized.

Subsequently, baseline data was defined as the initial interval/phase of data collection, known as the relaxation phase. The median value during this phase was subtracted from each corresponding data point for electrodermal activity (EDA), acceleration (acc), and heart rate (hr).

Electrodermal activity (EDA) is then represented for each data point as the difference between the current point and the point measured 10 data points back (`.diff(period=10)`). Negative values, indicative of non-positive reactions in electrodermal activity, are excluded from consideration, as the focus is solely on instances when sweat production occurs rather than during evaporation.

For acceleration data, the difference between each corresponding point and its immediate predecessor (one measurement back) is computed (`.diff()`). Given the exclusive interest in positive magnitudes of acceleration, arising from both the differences in magnitudes and the frequency of measurements (1Hz), negative values are disregarded and set to zero. This is justified by the fact that negative values denote a loss of acceleration, a factor unsuitable for the research objectives.

The next phase involves encapsulating the information about each stressor into a singular vector. For Electrodermal Activity (EDA), we start by setting a threshold value, with which we calculate a percentage that signifies the portion of observed values that lie above the threshold. This percentage serves as one dimension of the vector. Regarding heart rate, we decompose it into two dimensions: the first dimension consists of the maximum of the observed values within a specific stressor, while the second dimension encompasses the arithmetic mean. This ensures that any spike in heart rate will be represented in a good capacity. The acceleration associated with a particular stressor is also quantified by its average across all observations.

Following the quantification, the subsequent task is its practical application to the data. This involves the creation of a Pandas DataFrame comprising five columns, with each row containing the fields `stressor_name`, `eda`, `hr_max`, `hr_avg`, and `acc`. The label is denoted by `stressor_name`.

3.3 Model training and evaluation

Because the subsequent real-world IoT testing device does not measure EDA, the model will not be trained with it. Using grid search k-fold cross-validation, a random forest classifier with the most optimal hyperparameters will be trained. The library used is scikit-learn. On the aforementioned dataset, using only heart rate and accelerometer data, we have 160 datapoints. The training and testing were split into an 80/20. proportion.

Table 3.1: Classification Report

Class	Precision	Recall	F-score	Support
cognitive	0.5	0.5	0.5	2
emotional	1.0	0.6	0.75	5
emotional(horror)	0.5	1.0	0.67	4
physical	1.0	1.0	1.0	5
relaxation	0.93	0.81	0.87	16
accuracy			0.81	32
macro avg	0.79	0.78	0.76	32
avg	0.87	0.81	0.82	32