Automatic detection of perceived stress in campus students using smartphones

Martin Gjoreski, Hristijan Gjoreski, Mitja Luštrek, Matjaž Gams
Department of Intelligent Systems, Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
E-mail: {martin.gjoreski, hristijan.gjoreski, mitja.lustrek, matjaz.gams}@ijs.si

***Abstract*—This paper presents an approach to detecting perceived stress in students using data collected with smartphones. The goal is to develop a machine-learning model that can unobtrusively detect the stress level in students using data from several smartphone sources: accelerometers, audio recorder, GPS, Wi-Fi, call log and light sensor. From these, features were constructed describing the students' deviation from usual behavior. As ground truth, we used the data obtained from stress level questionnaires with three possible stress levels: "Not stressed", "Slightly stressed" and "Stressed". Several machine learning approaches were tested: a general models for all the students, models for cluster of similar students, and student-specific models. Our findings show that the perceived stress is highly subjective and that only person-specific models are substantially better than the baseline.**

Keywords—stress detection; smartphone; classification;

I. INTRODUCTION AND RELATED WORK

Work-related stress is defined as harmful physical and psychological responses that occur when the requirements of a job do not match the capabilities, resources, or needs of a worker, which can lead to poor health and injury [1]. In a similar way, "study-related" stress can be defined, where instead of workers we observe students. By detecting students' stress and reacting on time, not only their health is protected, but also their future can be improved by helping them to cope with stress, thus allowing them to achieve better performance in the studies and everyday life. The existing studies on stress detection can be roughly separated in two groups. In the first group are the studies [2][3][4] performed in a controlled laboratory environment where the stress is invoked intentionally by using some kind of stress test [5]. In these studies the researchers have complete control over the induced level of stress, and usually high stress detection accuracy is reported (80%–97%). Another approach to stress detection, besides physiological sensors, includes subject's voice analysis [6][7]. In the second group are studies that are analyzing stress in real-life situations where either lower accuracy is achieved [8], or the presented system is quite obtrusive and not well-suited for real-time stress detection, since large number physiological sensors are used [9]. All of the previous mentioned studies have one thing in common: the use of physiological sensors, which can be additional burden to the user regarding price and comfort. However, Zhai et al. [10] reported that by removing the data from one physiological sensor, the accuracy dropped from 93% to 62%. Regarding the research done on stress detection using smartphones, few studies exist. Bauer et al. [11] focused on detecting stress-related changes in the subject's behavior. Sano et al. [12] used smartphone data in combination with a physiological sensor for stress detection. The reported accuracy for a 2 class problem is over 73% by using 10-fold cross-validation. Muaremi et al. [13] reported that the accuracy for a three-class problem dropped from 61% to 55% for user-specific, and from 53% to 45% for the LOSO cross-validation, if the data from the physiological sensor is removed and only smartphone data is used. These results are a good indicator of how hard the problem of detecting stress by using only smartphone data is.

In this study we focus on the detection of perceived stress in campus students using only automatic sensing data extracted from their smartphones. The data used for this study is freely available on web [14]. It is collected as part of a larger study (StudentLife) [15] where the authors used automatic sensing data from smartphones to assess students' mental health, activity level, sociability, academic performance, and behavioral trends. Their approach is based on correlation analysis between different aspects of the students' life. One of the main conclusions of that study is: "Results from the StudentLife study show a number of significant correlations between the automatic objective sensor data from smartphones and mental health and educational outcomes of the student body". In our study we try to take their findings one step further by implementing machine-learning method that will be able to detect the students' stress level. The stress detection is presented as a classification problem with three possible labels: "Not stressed", "Slightly stressed" and "Stressed". The ground truth is obtained from stress questionnaires, which the students answered throughout the StudentLife study in order to assess their perceived level of stress.

The main contribution of our study is the proposed advanced machine learning approaches, including feature extraction (section III), combination of clustering and classification algorithms (section IV.B) and person-adaptive classification approach (section IV.C), applied on data gathered completely in the wild. We present thorough analysis of each approach and insight into their pros and cons.

978-1-4673-6654-0/15 $31.00 © 2015 IEEE
DOI 10.1109/IE.2015.27

II. UNDERSTANDING THE DATA

In the following section, the data used in this study is explained. For a more detailed information about the process of collecting the data, one can read the StudentLife study for which this data is originally collected.

We used data extracted from the following sources:

- Activity data, provided by an activity classifier which uses the smartphone's accelerometer to detect the student's activity (stationary, walking or running).
- Audio data, provided by an audio classifier which uses the smartphone's microphone to detect the student's audio surrounding (silence, voice or noise).
- Conversation data, provided by a conversation classifier detecting if the student is near a conversation.
- GPS data, which is logged every 10 minutes.
- Wi-Fi data, which is logged with a varying frequency.
- Call log data, which provides information about the time and duration of the calls.
- Light sensor data, battery charging data and locking data, which are logged if the smartphone had been in dark, charging or locked for a longer period of time.
- Stress questionnaires data. The students had to rate their current stress level on a scale: feeling great, feeling good, little stressed, definitely stressed, stressed out. The stress level that we are trying to detect is a class value derived from the answers of these questionnaires. An instance is labeled "Not stressed", if the student had answered feeling good or feeling great. An instance is labeled "Slightly stressed", if the student had answered little stressed. An instance is labeled "Stressed" if the student had answered definitely stressed or stressed out.

III. FEATURE EXTRACTION

The features used in our study can be separated in three groups: short-term features, date-time features and relative epoch features.

The short-term features are calculated using data only from the last two hours, regarding the answering time of the stress questionnaire. These features are calculated in order to provide information about students' behavior in the last two hours. The number of hours (two) was empirically chosen. In total 5 short-term features were calculated: *stationary ratio* (number of stationary inferences divided by the number of non-stationary inferences), *silence ratio* (number of silence inferences divided by the number of non-silence inferences), *voice ratio*, *noise ratio*, *conversation duration* (sum of the duration of all conversations recognized by the conversation classifier).

In total 3 date-time features were extracted: *number of days until midterm, nominal feature related to the midterm* (before, in and after midterm) and *answering epoch* (at which part of the day, the questionnaire is answered: morning, midday or night).

The relative epoch features have two characteristics. The first one is the epoch (period) of the day for which are calculated, and the second is relativity to the student for which are calculated. Regarding the epoch characteristic, these features are calculated for three different epochs. The first epoch is from 07:30 am until 18:00 pm (roughly from waking up until the end of the classes). The second epoch is from 18:00 pm until 00:00 am (period of the day when the students are studying, exercising, visiting friends, partying etc.). The third epoch is from 00:00 am until 07:30 am (period of the day when the students are probably sleeping). This granularity is introduced in order to distinguish the students' behavior for the three different epochs of the day. For example high activity in the night epoch (00:00 am-07:30 am) might mean that the student did not sleep that night, whereas high activity in the morning epoch might have completely different meaning.

Regarding the relativity characteristic of the relative epoch features, these features are calculated relative to the past behavior of the student for which are calculated. We introduced this relativity since features with absolute values (which we also tested, and performed worse than the relativity features) might have meaning only for user-specific stress detection. For example, a value for the activity level of one student can be obtained from the activity data (number of non-stationary inferences divided by the number of stationary inferences). However, activity level with value *X* for some students can be high activity level, but for other students it can be low or average activity level. By introducing the relative values of the features, we obtained an information about how the activity level changed compared to the past average activity level, regarding the activity data of the student for which it is calculated. For example, the feature *"activity deviation for epoch 1"* was calculated by subtracting the average activity level for epoch 1 of the past two days, from the average activity level for epoch 1 of all the other days, excluding the past two days. In a similar manner, several sources of data are used for calculating relative epoch features. In total 44 relative epoch features were calculated. Each of the following features is calculated for each of the three epochs, except the last two which are for the whole day: *activity deviation*, *silence deviation*, *voice deviation*, *noise deviation* (time spent in silent, "voicy" or noisy audio surrounding relative to the past), *number of calls deviation* (number of smartphone calls relative to the past), *duration of calls deviation* (duration of all smartphone calls relative to the past), *number of conversations deviation* (number of conversation inferences recognized by the conversation classifier relative to the past), *duration of conversations deviation*, *GPS distance deviation* (distance traveled in the last 2 days calculated by using GPS coordinates relative to the past), *stationary&silence deviation* (time spent stationary and in silent audio surrounding relative to the past), *stationary&voice, stationary&noise, stationary&conversation, nonstationary&conversation, maximum-call duration deviation* (duration of the maximum call in the last 2 days relative to the average duration of calls of the past) and *maximum-conversation deviation*.

The data from the light sensor and if the phone was charging or locked for longer period of time, was used to extract information about the students' sleep time. From each

of the three sources, duration was calculated (e.g. duration of phone being in dark) using only data from the previous night (22:00 pm until 10:00 am). From the three durations, the maximum was taken. This simple approach might not predict the exact time of sleep duration, but we hypothesize that it has positive correlation with the sleep duration. In this way, two features were calculated, on for the previous night and one for the night before the previous night, regarding the answering time of the stress questionnaire. Also, the data from the Wi-Fi scans was used to obtain information about student's current location, time spent there, and the location before the current location.

Since some of the features were extracted using the same source of data, it was expected that there is a high correlation between some of them. For correlation analysis the Pearson's correlation coefficient was calculated, which is a measure of linear correlation (dependence) between two variables. From each pair of features having the correlation coefficient higher than 0.65, the one with less information was excluded. For example the features "silence deviation" and the features "stationary&silence deviation" for all three epochs, were highly correlated, so the features "silence deviation" were excluded. After the feature filtering using correlation analysis, 47 features were left for further experiments, including the student id as a feature.

IV. Experiments

After the feature extraction, several machine learning approaches were tested in order to provide classification models that can detect the student's perceived level of stress. The classification algorithms were used as implemented in the machine learning toolkit, Weka. In the following subsections each machine learning approach is explained, and discussion about the evaluation results is presented.

A. *Leave one student out (LOSO)*

The LOSO cross-validation was performed by splitting the data into training and test sets, where the test set consisted only of the instances from one student and the training set of the rest of the data. This was done for each student and the results were averaged. SVM, j48, Bagging and Random Forest (RF) classifiers were compared. Random Forest, which proved slightly better than the rest, was also used as the base learner for Weka's Ordinal classifier, which means the order of the stress levels was taken into account (Stressed > Sligtlly stressed > Not stressed). With the LOSO technique, none of the classifiers achieved better accuracy than the majority class classifier (ZeroR), which was 43%. SVM and RF achieved accuracy of 42%. This results are confirmation that building general machine learning model for detection of perceived stress is very challenging task. We believe that the low accuracy is consequence from one of the main characteristics of the perceived stress, which is subjectivity: "Individuals may suffer similar negative life events but appraise the impact or severity of these to different extents as a result of factors such as personality, coping resources, and support." [16]. This characteristic makes it almost impossible to produce a general classification model by using only smartphone data. Our findings are in line with similar stress detection studies. For example Hernandez et al. [8] stated: "Although everyone had the same job profile, we found large differences in how individuals reported stress levels, with similarity from day to day within the same participant, but large differences across the participants." A proof for the high subjectivity of the perceived stress in our data, we can see on Fig. 1, where the class distribution for 10 students is presented. On the y-axis is the number of instances. For example, students S4 and S5 reported almost all the time that they were stressed, while students S6 and S9 reported stress only 10%–20% of the time.

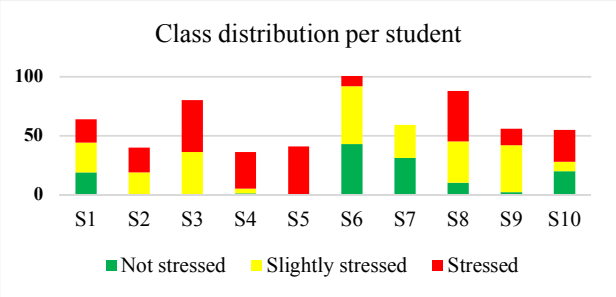

Figure 1. Class distribution for 10 students

B. *Cluster-specific classification*

With the previous approach we realized that building a general stress detection model for all the students is not feasible with our data. Therefore we wanted to try if clusters of students can be created for which accurate cluster-specific classification model can be built. First, the data of one student was removed to be used as test data. On the remaining data, a clustering algorithm (Weka's expectation maximization) was used to create clusters of students. For each student, all instances were assigned to the cluster which contained most of his/hers instances. Next, a classification algorithm (SVM, j48, Bagging, Random Forest, or Ordinal classifier) was used to train a cluster-specific model using only the data of the specific cluster. In the testing phase, the test student was assigned to the cluster in which most of his\hers instances were clustered by the clustering algorithm, and the cluster-specific classification model was used to classify the instances of the test student. In the clustering phase (training and testing), only the relative epoch features and the Wi-Fi features were used, since these features contained information about students' average behavior, and we wanted to cluster students with similar behavior. In the classification phase all 47 features were used. This procedure was repeated for each student and the results averaged. With this technique, no significant improvement in the overall accuracy was achieved. The highest overall accuracy of 43% was achieved by J48 which was equal to the accuracy of the majority classifier.

C. *Learning with a calibration phase*

The idea behind the learning with a calibration phase is that at the beginning there exists a general classification model for stress detection which needs a calibration phase during which it is adapted to a specific student. Like previously, the data of one student was first removed to be used as test data. From that test data, *X* random instances were moved to the training data as calibration data. This means that the training data consisted of

all the data from the other students and *X* instances belonging to the test student. A general model was then trained on the training data, and evaluated on the test data. This procedure was repeated for each student and the results averaged. With this technique Random Forest was tested and compared to classifiers trained only on the calibration data of the test student. In Fig. 2 it is presented how the accuracy changes as the X – the amount of calibration data – increases. The "General RF" classifier was trained on the calibration data and the data of other students, while the "Specific" classifiers were trained on the calibration data only. In order to avoid overfitting (since the "Specific RF" model had small amounts of data to train), the number of features is reduced to 20. The reduction is done by averaging the features for different epochs of the day. For example, the three features *average activity for epoch 1*, *average activity for epoch 2* and *average activity for epoch 3*, are replaced by one feature, which is average of the three. Also, the short-term features were excluded. From the results in Fig. 2, we can see that starting from three calibration instances, the "Specific RF" model performs better than the "Specific ZeroR" and the "General RF" (even though the "General RF" had more training data), which is another indicator of how superior are the specific models for this data. In general, the accuracy of the "Specific RF" increases as the amount of calibration instances increases. The first significantly high accuracy of 60% is achieved for 23 calibration instances which correlates with the number of features used in this approach (20).

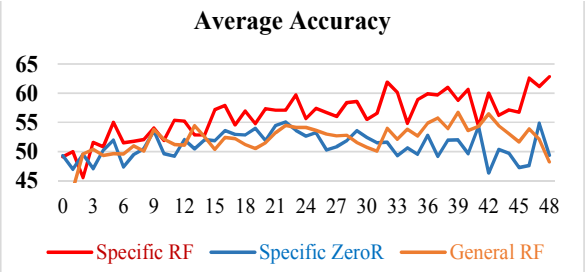

Figure 2. Average accuracy for varying number of calibration instances.

V. Conclusion

Building general classification models for perceived stress detection using only smartphone data is challenging problem. Even clustering students' with similar behavior using automatic sensing data, and building cluster-specific classification models didn't improve the average accuracy for the stress classification. These findings were confirmed by the technique "Learning with a calibration phase", where the person-specific model yielded best results. Once the classification algorithm had enough data to build a model, it performed better than the majority class classifier and the general classifier. Perceived stress is very subjective and each individual is specific, so smartphone stress detection can be done by building person-specific models, where certain period of time (e.g. 20-25 days) user input is needed. In the future, we plan to add smartphone based voice analysis [17], serious sleep analysis [18] and social media analysis to automatically and unobtrusively detect stress. Also, new smartphone technologies (such as Samsung Galaxy S5's heart rate monitor) can be easily integrated into our approach.

Acknowledgment

The authors would like to thank prof. Andrew Campbell and the team of the original "StudentLife Study" [15].

References

[1] S. M. Jex, "Stress and job performance: theory, research, and implications for managerial practice," Thousand Oaks, CA: Sage, 1998.

[2] J. Wijsman, B. Grundlehner, H. Hermens, "Wearable physiological sensors reflect mental stress State in office-like situations," Humaine Association Conference on Affective Computing and Intelligent Interaction. Geneva, pp. 600–605, 2013.

[3] D. McDuff, S. Gontarek, R. Picard, "Remote measurement of cognitive stress via heart rate variability," Engineering in Medicine and Biology Society, pp. 2957–2960, 2014.

[4] B. Ionut-Vlad, G. Ovidiu, "A study about feature extraction for stress detection," The 8th International Symposium on Advanced Topics in Electrical Engineering, Bucharest. pp. 1–4, 2013.

[5] K. Dedovic et al., "The montreal imaging stress task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain," Journal of Psychiatry and Neuroscience, vol. 30, no. 5, pp. 319, 2005.

[6] S. Scherer et al., "Emotion recognition from speech: Stress experiment," in Proceedings of the Sixth International Language Resources and Evaluation, pp. 1325-1330, Marrakech, Morocco, 2008.

[7] Phil Adams et al., "Towards personal stress informatics: Comparing minimally invasive techniques for measuring daily stress in the wild," in the 8th International Conference on Pervasive Computing Technologies for Healthcare, 2014.

[8] J. Hernandez, R. R. Morris, R. W. Picard, "Call center stress recognition with person-specific models," in Affective Computing and Intelligent Interaction, pp. 125– 134. Springer, 2011.

[9] J. A. Healey, R.W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," Intelligent Transportation Systems, IEEE Transactions on, vol. 6, pp. 156– 166.

[10] J. Zhai, A. Barreto, "Stress detection in computer users based on digital signal processing of noninvasive physiological variables," Engineering in Medicine and Biology Society, 2006. IEEE, 2006, pp. 1355–1358.

[11] G. Bauer, P. Lukowicz, "Can Smartphones Detect Stress-Related Changes in the Behaviour of Individuals?" Pervasive Comp and Comm. Workshops, IEEE, Int.Conference, pp. 423– 426, 2012.

[12] A. Sano, R.W. Picard, "Stress Recognition using Wearable Sensors and Mobile Phones," Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, pp. 671 – 676, 2013.

[13] A. Muaremi, B. Arnrich, G. Troster, "Towards measuring stress with smartphones and wearable devices during workday and sleep". BioNanoSci, pp. 172–183, 2013.

[14] http://studentlife.cs.dartmouth.edu/

[15] R. Wang et al., "StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones," ACM Conf. on Ubiquitous Computing, pp. 3-14, 2014.

[16] A. C. Phillips, "Encyclopedia of Behavioral Medicine", 2013, pp 1453-1454.

[17] Lu H et al., "StressSense: Detecting Stress in Unconstrained Acoustic Environments using Smartphones," ACM UbiComp, pp. 351-360, 2012.

[18] Chen Z et al., "Unobtrusive Sleep Monitoring using Smartphones", 7th International Conference on Pervasive Computing Technologies for Healthcare, Venice, pp. 145 – 152, 2013.