2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation

Assessing Mental Stress Based on Smartphone Sensing Data: An Empirical Study

Feng Wang[1,2,4], Yasha Wang[1,3,4*], Jiangtao Wang[1,2,4], Haoyi Xiong[5], Junfeng Zhao[1,2,4], Daqing Zhang[1,2,4]
[1]Key Laboratory of High Confdence Software Technologies, Ministry of Education, Beijing 100871,China
[2]School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
[3]National Engineering Research Center of Software Engineering, Beijing 100871, China
[4]Peking University Information Technology Institute, Tianjin Binhai 300450, China
[5]Big Data Lab, Baidu Research, Beijing, 100085
{wangfeng2013, wangyasha, jiangtaowang, zhaojf, dqzsei} @pku.edu.cn
xhyccc@gmail.com

***Abstract*—Mental stress is a critical factor affecting one's physical and mental well-being. At the early stage, the effect of stress is often underestimated, while it usually leads to serious issue Lateran. Therefore, it is crucial to detect stress before it evolves into severe problems. Traditional stress detection methods are based on either questionnaires or professional devices, which are time-consuming, costly and intrusive. With the popularity of smartphones embedded with a rich set of sensors, which can capture people's context, such as movement, sound, location and so on, it is an alternative way to access people's behavior by smartphones. Through an empirical study, this paper proposes an automatic and non-intrusive stress detection framework based on smartphone sensing data. First, we construct various discriminative features from multi-modality phone sensing data, in which both absolute and relative features are considered to make the model more personalized. Then, to tackle the challenge of label insufficiency, we further develop a co-training based method for stress level classification. Finally, we evaluate our model based on an open dataset, and the experimental results verify its advantages over other baselines.**

***Index Terms*—mental health, mobile crowd sensing, automatic detection, machine learning**

I. Introduction

In a fast-paced life, the health status of massive crowds is frequently invaded by pressure. The American College Health Association's report [1] in 2015 showed that 57.7% of students felt very anxious at least once in the past 12 months. At the same time, studies show that pressure would significantly affect people's psychology and behavioral habits. People tend to have anxiety, insomnia when they feel a lot of pressure, and in some serious cases, leading to mental or physical diseases [2]. This kind of mental illness brought about by stress are often neglected at the initial stage, but it may develop into a serious problem [2]. Therefore, the timely detection of psychological pressure before it transforms into a serious psychological problem has important implications for the wellbeing of college students.

In recent years, the detection of psychological stress has attracted more and more attention. The psychology field has made a lot of research on how to effectively detect people's psychological pressure. Traditional methods focus on using psychological theory-based questionnaires [3]. Because of the theoretical support behind it, this method is still the most widely used one. People's psychological pressure can also be monitored by professional instruments [4], [8]–[10]. For instance, the electric resistance of human skin relates to certain psychological indicators [4]. Monitoring the skin' electric resistance can achieve the detection of stress, and the results are reliable. However, such methods are intrusive and costly. Therefore, we hope to find an automatic, low cost, less intrusive method to realize psychological pressure monitoring.

Meanwhile, smartphones have become a necessity in people's lives. In order to enrich user experience, more and more sensors (such as acceleration sensors, acoustic sensors, etc.) have been integrated into mobile phones [5], [6]. In daily life, mobile phones can continuously record a lot of sensing data related to people's daily-life behavior, including activities, location information, and mobile phone usage information. Studies have shown that people's behavior can reflect people's psychological stress. For example, people under stress tend to exhibit reduced activity positivity, frequent use of mobile phones, and low sleeping quality [2]. The sensing data provided by the mobile phone can reflect behavioral habits, which may be related to pressure. Therefore, there is an opportunity to explore the correlation between the smartphone's sensing data and the people's psychological state and utilize it to develop a learning model for mental stress detection.

The use of mobile phone sensing data to detect people's mental stress has the following technical challenges.

(1) How to extract features from multimodality sensor data that can discriminate two populations (subjects with/without stress). Each data source represents a certain aspect of human behavior. A certain feature may be extracted from multi-source sensing data, resulting in a number of possible combinations during the feature extraction phase. Thus, it is complicated to identify discriminative features from such a large space of candidates.

(2) How to make the generalizable model more personalized. The correlation between sensing data and stress level

∗ Corresponding author

978-1-7281-4034-6/19/$31.00 ©2019 IEEE
DOI 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00200

is quite personalized. Ideally, we should collect a sufficient volume of training data with stress level labels for each person, so that we can develop multiple personalized classification models. However, this is not realistic as continuous labeling is very intrusive. For such reason, we need to develop a universal model shared by all persons with personalized components integrated.

(3) How to learn an accurate stress detection model using insufficient labeled training data. In the data collection stage, smartphones can continuously collect various data at a low cost and generate eigenvectors. However, the labeled mental stress level corresponding to each feature vector needs to be labeled by the user and cannot be obtained in large quantities, resulting in scarce training data with labels. Therefore, how to carry out accurate model training with insufficient labeled data is another technical challenge.

The contribution of this paper lies in the following four aspects:

First, we extract features from heterogeneous smartphone sensing data (such as location, POI and activity) which can be used to discriminant to two populations. i.e., human subjects with/without stress. In the process of feature extraction, we combine multiple features into views for the analysis. For example, by combining activity data and POI data, we can view people's activity distribution among different POIs as a set of new sematic-rich features incorporating low-level raw data/features.

Second, to integrate personalized components into the generalizable detection model, we extract both relative and absolute features. Absolute features are implemented as the statistics of certain sensing data in a specific time window, while the relative features are estimated by comparing the absolute features with people's historical data, which can depict the personalized change of behavior.

Third, to handle the insufficient labeling problem, we propose a co-trainer-based semi-supervised learning approach for stress detection, which leverages unlabeled data to improve the classification accuracy. This method makes our approach more practical in real-world settings since people do not need to consistently report their stress level as training data labels.

Finally, we evaluate our approach based on an open dataset containing both heterogeneous smartphone sensing data and mental stress levels of 49 college students during a 10-week period. The experimental results justify the advantages of our approach over several baselines.

II. Related Work

A. Emotional Assessment Based on Professional Sensing Devices

As people's psychological changes will inevitably lead to changes in certain physiological indicators, many studies are devoted to using wearable devices to monitor people's daily psychological pressure [8]–[10]. Typically, these devices integrate specialized sensors that can sense changes in people's physiological indicators such as skin electric resistance, body temperature, heart rate, and blood pressure. Because such data can be obtained directly, the wearable device based sensing data is often reliable. However, people have to wear these professional devices, which brings about the problem of high cost and intrusion. In contrast, the psychological pressure sensing based on mobile data can spare people such problems.

B. Emotional Assessment Based on Social Networks

Another type of work can assess people's emotions by their behavior on the social network. Lin et al. [11] use a deep sparse neural network to assess people's psychological pressure based on people's Weibo data; Lin et al. [12] use a convolutional neural network to detect people's pressure. In terms of adolescents, Xue et al. [14] extract a series of features based on their tweets and use classifiers to understand the stress levels of teen's pressure. Jin et al. [15] propose an approach based on co-training, which combines Weibo and trajectory information to perform pressure testing on teenagers.

By using behavior data and posts on social networks, social network-based works deploy natural language processing and deep learning to assess people's psychological pressure with minimum intrusion. However, these methods can only focus on people who frequently use social networks. For people who do not use social networks frequently, it is difficult to predict their psychological pressure. At the same time, since people do not constantly use social networks, it is impossible to continuously monitor people through social network data. Whereas this paper uses mobile phone data to continuously monitor people and thus has better pervasiveness.

C. Emotion Assessment Based on Smartphone Data

In recent years, several research works have been done to focus on the emotion-related analysis based on smartphone data, which can be roughly divided into two categories.

In the first category of work, Mehrotra et al. [16] extract a series of features using daily mobile phone communication data and app usage habits of users and analyze the correlation between multidimensional features and the degree of user frustration through linear regression methods. Bogomolov et al. [20] use data collected by mobile phones (including cell phone call data and SMS data), they build a classification model to recognize people's daily stress. Xiong et al. [17] use linear regression to analyze the correlation between different behavioral habits and people social anxiety through college students' GPS and POI data. Canzian et al. [18] use GPS data to predict people's depression degree. They extract multi-dimensional features from people's GPS data and obtained the people's frustration degree from the PHQ-9 questionnaire. Lu et al. [19] predict people's real-time nervousness on different occasions through the sound data sensed by mobile phones. In their work, users need to wear two mobile phones placed on the neck and waist respectively to collect sound information and use acoustics related methods to extract features. A professional wristband is used to obtain the user's nervousness status. The above works of literature demonstrate that it is feasible to exploit smartphone sensing data to infer a person's emotional status, which inspires the work in this article. However, in

TABLE I
STUDENTLIFE SENSING DATA

No.	Type	Description
1	Activity	User activity status
2	Audio	Audio status around user
3	Conversation	Conversation info of user
4	Bluetooth	Bluetooth scan log
5	Dark	Duration of phone in dark
6	GPS	GPS log
7	Phonecharge	Duration of phone charge
8	Phonelock	Duration of phone lock
9	Wi-Fi	Wi-Fi log
10	Wi-Fi location	Wi-Fi AP location

these studies, only one type of phone sensing data is used. In contrast, we utilize multiple types of data to study the stress detection problem.

Wang et al. [7] extract multidimensional features from the StudentLife dataset, use linear regression to analyze the relationship between a variety of psychological indicators, such as the user's behavior and psychological pressure during the semester. This work analyzes the correlation between the sensing data and the long-term psychological indicators rather than the short-term indicators. As we mentioned before, the short-term mental health (such as stress level) assessment is very important. Zhang et al. [27] used a factor graph model to predict people's mood by using smartphone sensing data and app usage. They proposed a method for assessing short-term emotions and a universal model was built. In this paper, we use absolute features and relative features to make the universal model more personalized. In addition, we use co-training to tackle the label insufficiency issues which is not considered by Wang's and Zhang's works [7], [27]. This makes our approach more practical in short-term mental status inference, as people do not need to continuously report their mental status as labels.

III. DATASET AND PROBLEM DEFINITION

We use open source dataset StudentLife [7] to train the model. StudentLife dataset is a dataset from Dartmouth, which contains 49 students' data in 10-weeks. The dataset has four parts: sensor data, EMA (Ecological Momentary Assessment) data, survey responses, and educational data. In this paper, we only use sensor data and EMA data. Sensor data contains the data captured by smartphone sensors, and the used sensors are listed in table 1, which describes the activity information, context data, and user behavior. The data is automatically collected by smartphones in users' daily lives. EMA data is ecological momentary assessment data, which is always used to reflect human behavior [26]. During the collection phase, lots of EMA questions were sent to students. The EMA reports reflect students' daily emotions. In this paper, we use stress EMAs. By using this data, we can obtain the user's mental stress status at a specific moment. The stress level is marked from 1 to 5 in StudentLife (1. A little stressed, 2. Definitely stressed, 3. Stressed out, 4. Feeling good, 5. Feeling great). In this paper, we combined levels 1 and 2, marked them as stressful and others are stressless.

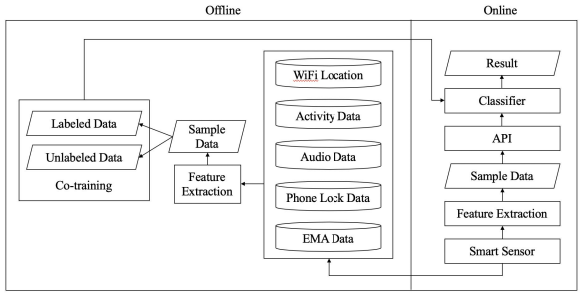

Fig. 1. Mental stress assessment framework

Problem statement: Given a sensor dataset D, we extract feature set $F = \{f_1, f_2, \ldots, f_n\}$, using F to generate sample matrix X_{mn} and its label y_m (where y_i is the label of the ith row in X_{mn}). We aim to train a binary-class classifier C that enables the completion of two classification tasks after a given input.

IV. STRESS INFERENCE FRAMEWORK

A. Overview

The framework of our method is shown in Figure 1, and it consists of offline training and online prediction:

1) *Offline training:* To get the training sample, feature extraction and feature selection are performed on the raw data. The training data includes labeled data and unlabeled data. The classification model is trained by co-training.

2) *Online prediction:* We can get sensing data from mobile phone smart sensors. By using feature extraction, the raw data can be accurately classified by the model.

B. Feature Extraction

Feature extraction works to transform log files into sample data that can be used for classification. To get sample data, we need to introduce the time window. The log data and sensing data in each time window can be transformed into one sample, which is then marked accordingly.

In this paper, we choose 24 hours as one time window. For each EMA sample, the 24-hour sensing data before the user's feedback of the result is selected to generate a sample corresponding to this EMA result flag. We explain why we choose 24 hours as a time window in the following. First of all, people's daily behavior is regular, and many indicators do not change drastically. Secondly, more meaningful information can be sampled when the length of the time window is 24 hours, such as the user's sleep information. The user's sleep and the user's psychological state is closely related to the 24-hour window, so they can be effectively sampled; Finally, with EMA data as the sample annotation, a sample is generated based on the EMA data fed back by each user. In the data set, the frequency of the EMA data is close to one time per user per day. Then samples that are generated according to days are more in line with the physical meaning in actual operation.

In the feature extraction process, the features need to be refined, such as the number of people around the user. The

number of people around the user is relatively large during the daytime and during the nighttime have different meanings. This is because users have more activities in the daytime and more people come into contact with them. At night, they usually go back to the dorms. People around them are relatively fixed and are fewer in number. If a user has more people to come into contact with in the evening, it may indicate that the user is participating in some kind of activity. This discovery is meaningful, but if you measure this feature in days, you cannot get this conclusion. This shows that the refinement of the features lead to more information. In the process of feature extraction, this paper uses two kinds of feature refinement methods, which are refined according to time and point of interest (POI).

1) Refinement according to time: The time is divided into daytime (8:00-18:00) and nighttime (18:00-8:00). In the subsequent feature extraction, each dimension should be considered in conjunction with time.

2) Refinement according to POI: POI describes the user's location, and similar to time, we can also get more information when the user's performance is divided according to the POI.

Absolute Features F_a

Absolute features are the features that are extracted from data sources directly. Each dimension clearly reflects the absolute value of a certain statistical value. All the absolute features are listed in Table 2.

POI Based Features

Inspired by Jin's work, the frequency of user access to different POIs has a strong relationship with the degree of user's psychological anxiety. In this paper, we obtain the user's POI information from the Wi-Fi location data of the StudentLife data set.

Wi-Fi location data is the Wi-Fi scan logs. They tell the POI of the access point. We divide the POIs into three types: 1) Teaching area. Teaching building, laboratory, library. 2) Dormitory area. Student apartments, hotels. 3) Diet and wellness area. Canteen, gym, art gallery. This helps with the refinement of the POI information. Just as time is refined, the time is divided into daytime and nighttime, and each property is calculated according to these two time periods. By using POI refinement, the position is divided into teaching areas. Dormitory area, eating and drinking area.

At the same time, we hope to use an indicator to reflect how much time users spend on diffirent POIs each day. This kind of information can reflect a user's daily behavioral habits. For example, users who love to study would spend more time in the teaching area every day, while users in homes have more time in the dormitory area, and those who love fitness may spend more time in the catering and health area. Based on this idea, the POI data of each user in the time window is counted. Since the sampling frequency is basically constant, the number of entries of each type of POI is proportional to the time the user stays in each type of POI.

In addition to considering the length of time users spend in a certain type of POI, these three types of data constitute a distribution. This paper introduces entropy to express the characteristics of this distribution. For a multi-class distribution X, we define the entropy as:

$$H(X) = -\Sigma_i P(x_i) lg(P(x_i)) \tag{1}$$

Activity-relevant Features

POI data can reflect the user's activity. The StudentLife dataset uses the physical motion classifier to classify the raw data by using the data collected by the accelerometer sensor to obtain the user's movement state at a certain moment: stationary, walking or running (Others are unknown). The frequency of sensor sampling is constant, so within a time window the number of labels in each category corresponds to the length of time the user is in motion state. This paper counts the number of labels for each category in the time window, and calculates the entropy. At the same time considering the refinement according to time.

GPS data can also reflect the users' activity patterns. The GPS sensor samples the user's latitude and longitude each 10 minutes. In Jin's work[13], researchers explored the impact of the user's moving distance on the user's psychological anxiety level. We use GPS data to calculate the user's moving distance during the day in the time window and the distance in the night. The distance is the cumulative distance between adjacent sampling points. The distance between the two latitude and longitude points is calculated using the more accurate Flat earth distance. The formula is as follows: (x, y) is the latitude and longitude of a data point:

$$D((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 x_2)^2 + cos^2(\frac{x_1 + x_2}{2})(y_1 y_2)^2} \tag{2}$$

Conversational Features

From Lu's work [17], it has been known that the conversational information of the user's environment can be used to predict the user's stress level, which shows that the sound information is closely related to the user's psychological indicators. By using mobile phone sensing data, we can obtain the conversational information of the user's environment. The StudentLife data set uses sound classifiers and conversational classifiers to obtain a series of sound-related information: one to sense the presence of sound in the phone's environment, and the other to sense whether the sound is vocal. This information includes the type of sound the user is in at each moment: silence, noise, and voice. The data process method is similar to the method of processing user activity information, combined with the time refinement, the number of 3 types of tags of the user during the day or night is counted, that is, the user's time in the corresponding space. At the same time, the entropy is also calculated to describe this distribution.

The dataset also contains the dialogue information of the user's environment. Combined with the psychological point of view, it is found that the user's psychological pressure would affect whether they are willing to communicate.

Bluetooth Features

TABLE II
ABSOLUTE FEATURES

Dimension	Data source	Number of features	Feature description
POI-based Features	GPS, Wi-Fi	8	POI count in different areas and different periods, entropy of POI count distribution
Activity-relevant Features	Activity	8	Number of different labels in different time periods, entropy of label distribution
Conversational Features	Conversation	18	Number of different labels in different periods, entropy of label distribution, dialog count in different areas and different periods
Bluetooth-based Features	Bluetooth	5	Number of scanned devices in different areas and different periods
Sleeping-relevant Features	Phone Lock	2	Sleeping time, length of sleeping duration

Studies have shown that when users are in a depressing state, they tend to be more autistic and do not want to communicate with people. Therefore, we can extract features to describe the user's social information. Different from the previous work, the researchers can obtain the user's social-related information through the user's phone text message data. Mobile phone sensing data does not contain related information, but it can use bluetooth scan data to approximate social information. The phone would periodically scan and record the total count of the scanned bluetooth device. Based on our knowledge, most of the bluetooth devices that can be scanned are smartphones, computers, and other devices. For unsolicited users, they usually do not go to densely populated places. On the contrary, for users who are active and outgoing, they are keen on participating in various activities. The scanned count would also be more. Based on this, the article counts the number of devices scanned by different POIs and bluetooth devices at different time periods in order to describe the user's social habits.

Sleeping-relevant Features

Psychological pressure can cause anxiety, poor sleep quality and other physiological responses, therefore, in order to predict the user's psychological pressure, the quality of sleep is a very important feature. The duration of the user's sleep and the time to fall asleep can help classification.

This paper uses the mobile phone lock screen records to get the user's sleep habits. Each record records the turned-on and turned-off time of the mobile phone's lock screen (time span exceeding 1h would be recorded). During the study, it is found that a long record appears every 24h, and the record is in the middle of the night, that is, the record corresponds to the user's nighttime rest (the user does not use the mobile phone during sleeping time, so he would leave a record of length equal to the length of sleep). This paper uses the start time of the user's record as the user's sleep time, and constructs the user's sleep duration and sleep time features.

Relative Features F_b

Absolute features characterize the absolute amount of human behavior in a given time window. But we think that the change of user's behavior can reflect the change of user's mental state. So we propose the relative feature to depict the change of user's behavior. The relative feature ignores the difference between absolute values, and only concern about the change of user's behavior, so it can also make the universal model more personalized.

The benchmark used for comparison is the average value. Since the time window is 24h, the data includes a total of 49 students' perception data over 70 days. Therefore, the 49-day data of these 49 students are sampled and generated according to the features mentioned in the previous section. After averaging the samples, an average sample can be obtained. (In the dimensions mentioned in the previous feature extraction part, not all dimensions there are contrasting average value. For example, the entropy contrast has no physical meaning. This kind of value would not be compared.) After that, for all the comparable dimension values, calculate the following formula, for dimension i:

$$r_i = (v_i - avg_i)/avg_i \tag{3}$$

The obtained feature r_i is added to the original feature vector as a new feature dimension, and each row of the previously obtained sample matrix is processed to obtain a sample matrix with the added relative features.

C. Feature Selection

According to the above method, 83 features are extracted. However, because the features are manually extracted, there must be some features that are not significantly correlated to the user's psychological stress. Therefore, it is necessary to reduce the dimension of the features.

For the interpretability of features, this paper does not use PCA-like methods for dimension reduction, so the feature screening problem is equivalent to selecting the optimal subset, and the optimal subset problem is NP-hard, so we use approximate methods to solve this problem. Selecting a subset can make the classifier achieve the best possible results and can be solved in polynomial time. Combined with the random forest classifier used in the verification process, this paper proposes a dimension reduction method based on Gini impurity.

Gini impurity is a statistic used to measure the purity of a data set and is widely used in decision trees. And the Gini index describes the change in the Gini impurity of the data set during the partitioning process.

For a specific dataset D, assuming the feature set $F = \{f_1, f_2, \ldots, f_n\}$, the Gini impurity is defined as $I_G(D)$:

$$I_G(D) = 1 - \Sigma_{k=1}^{|K|} p_k^2, (p_k = \frac{n_k}{|D|}) \tag{4}$$

when the data is divided by feature f_i, the Gini index is:

$$\Delta I(D, f_i) = I_G(D) * (p_l I_G(D_l) + p_r I_G(D_r)) \tag{5}$$

The greater the change in Gini impurity, the higher the purity of the data obtained by dividing by this dimension. When feature selection is performed, these kinds of features should be retained as much as possible. Based on this idea, all features are used to calculate the Gini index, and the features are sorted by the Gini index, and a larger Gini index would be preferred. Then the classifier evaluation is performed. The selection results are shown in section 6.

D. Semi-supervised Learning Based Stress Inference

1) Co-training: Co-training is a semi-supervised model that can use a large amount of unlabeled data to train model, which can help to improve the accuracy of the classifier when there are few labeled data [25]. Co-training needs to analyze data from two different "perspectives". It requires that the data set has two relatively independent feature sets, and the two are independent of each other. Collaborative training combines two classifiers from different perspectives to build a more accurate classification model.

In this paper, two kinds of feature extraction methods are used to obtain the absolute features directly extracted from the original data and the relative features based on personal historical data. These two types of features are also mathematically independent. Classifiers are constructed separately for the two types of features, and two classifiers can be obtained by using co-training. The algorithm is shown in Algorithm 1.

Algorithm 1 Co-training

Input: Labeled data set D_l, unlabeled data set D_u, number of iteration round θ, selected count limit n, feature sets f_1, f_2

Output: classifier h_1, h_2

1: Project D_l by f_1, f_2, get projected data sets D_{f_1}, D_{f_2}
2: Train classifier h_1 by D_{f_1}, h_2 by D_{f_2}
3: For each sample in D_u, calculate it's label by h_1, choose n sample with higher confidence, add them into D_l
4: For each sample in D_u, calculate it's label by h_2, choose n sample with higher confidence, add them into D_l
5: $\theta = \theta - 1$
6: **if** $\theta < 0$ or $D_u = \Phi$ **then**
7: **return** h_1, h_2
8: **end if**
9: go to step 1

In each round of iterations, n samples with higher confidence given by the random forest classifier are added to the training sample set from the unmarked sample set until the unlabeled sample set is empty. Finally, we can get 2 random forest classifiers: h_1, h_2.

TABLE III
NUMBER OF SAMPLE

Data Type	Stressed	Not Stressed	Total
Labeled Sample	1587	580	2167
Unlabeled Sample	-	-	9800

2) Online Prediction: Through co-training, two classifiers h_1, h_2 based on different feature dimensions can be obtained. In the prediction process, the results given by the two classifiers need to be combined to make an assessment.

Both classifiers are based on random forests. Random forest is a compound classification model obtained from a combination of many individual decision trees. The result of random forest is determined by the output of the individual decision tree. For the two-class c_1, c_2 problem, we set up a random forest with m decision trees, the number of trees who vote c_i is n_i, then the probability that the random forest thinks the sample belongs to c_i is

$$P^{c_i} = \frac{n_i}{m_i} \tag{6}$$

In the prediction process, the following formulas are used to obtain the classification results of the two classifiers, and a category with a higher probability is selected as the classification result:

$$c = argmax_{c_i \in C}(p_1^{c_1} * p_2^{c_2}) \tag{7}$$

Feature extraction is performed on the given user data. After the feature selection, the sample can be obtained. The above method can be used to complete the online classification.

V. EXPERIMENT

A. Dataset, Methods, and Baselines

The Dartmouth College StudentLife dataset was used as the experiment data, which includes sensor data and EMA feedback data of 49 students of more than 10 weeks. Using the feature extraction and selection methods proposed in the previous section, labeled data and unlabeled data are generated. The quantity information is shown in Table III.

In the experiment validation section, this paper validates three aspects of the method.

1) Comparison with the Baseline method. This is to prove that the proposed model is superior to other Baseline methods. The methods involved in the comparison include: decision tree, support vector machine (SVM), and K-nearest neighbor (KNN) and logistic regression;

2) Evaluation of the effectiveness of feature selection. This is to prove that the application of feature selection can significantly reduce the number of features participating in training while ensuring the model's effectiveness;

3) Evaluation of the performance of co-training. This is to prove that the co-training can make use of unlabeled data, and help to increase the accuracy of the classification model.

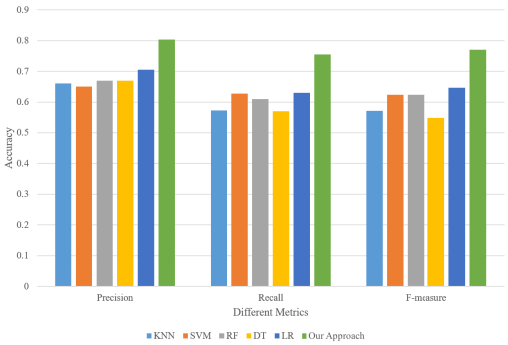

Fig. 2. Comparison of different classifiers

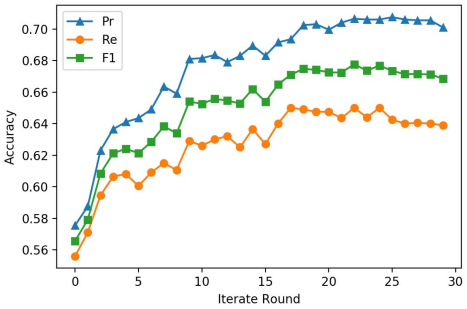

Fig. 3. Performance with number of selected feature

TABLE IV
TOP 10 FEATURES SELECTED BY GINI INDEX

No.	Feature
1	Length of sleeping duration, relative
2	Number of scanned device, daytime, relative
3	POI number in teaching area, daytime, absolute
4	Entropy of label distribution, daytime
5	Total time of dialog, teaching area, absolute
6	Number of scanned device, nighttime, absolute
7	Number of Walking, nighttime, absolute
8	Length of sleeping duration, absolute
9	Number of Silence, daytime, relative
10	POI number in accommodation area, nighttime, absolute

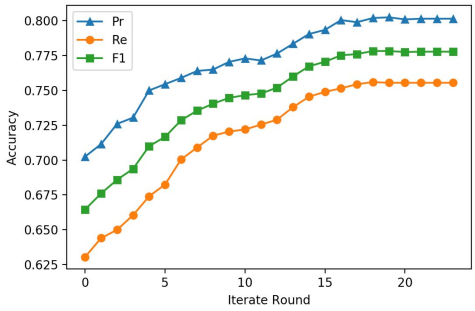

Fig. 4. Performance with number of iteration round

B. Experiment Result

The experiment results compared with the baselines are shown in Figure 2. It can be seen that the performance of our method on Precision (Pr), Recall (Re), and F-measure (F1) is better than using these classifiers directly. This should give credit to the co-training which utilizes a large amount of unlabeled data, and random forest classifier helps overcome over-fitting caused by a small number of data samples.

Manual feature extraction has certain redundancy. In this paper, feature selection is based on Gini impurity. For each feature, calculate the change in Gini impurity, invert it accordingly, and then continue to add in features to train the model, the evaluation of which is plotted as a curve as shown below. The x-axis represents the number of features used in the training, and the y-axis represents the evaluation indicators of the model after each iteration. As can be seen from the figure, as the number of features increases, the values of the various assessment indicators first show an upward trend and then tend to be stable (Figure 3). It shows that after adding a certain amount of features, the classifier effect tends to be stable, the new features no longer need to be added. According to this method, the first 26 features are selected for the final classifier training. The top 10 features are listed in Table IV.

This paper uses co-training to label data without labels and use it for iterative training. Two basic random forest classifiers are trained using labeled data, and used to predict unlabeled data. Samples with higher prediction confidence are selected to enter the labeled sample set. Based on the new training dataset two random forest classifiers are trained, and the classification results of the two classifiers are evaluated. Iterations are continued until the classification effect becomes stable. In Figure 4. The x-axis is the number of iterations, and the y-axis is the evaluation index of the classifier. It shows that with the addition of the sample data labeled by the classifier, the effect of the classifier is first gradually increasing, and the curve tends to be stable, indicating that the effect of the classifier is stable. Since then adding more sample can no longer significantly improve the classifier effect.

It can be inferred from the experiment that with the iteration of co-training, the training model is optimized. Therefore, in the absence of a large amount of labeled data, co-training can effectively utilize unlabeled data, thereby improving the effectiveness of the original classifier.

VI. LIMITATION AND FUTURE WORK

Using the methods provided in this paper, data collected from smartphone sensors can be analyzed to assess whether users are under pressure. However, the following problems still exist in this work.

First, this paper uses 49 students' data of 10 weeks to train a model to predict the user's psychological pressure. The relevance between mental stress level and smartphone sensing data is different among different people. In this paper, as the

sample size is limited, it is difficult to generate a classification model for each user. If there is enough personal data, a user-specific classification model can be generated for each user, and correspondingly, better prediction results can be achieved.

Second, this paper extracts absolute features and relative features for different types of data use Gini impurity for feature selection. Although the prediction has achieved certain results, there is still room for improvement in terms of correlation interpretation. More psychology-related knowledge can be applied to extract features, and the correlation between features and user psychological pressure can also be studied and interpreted. At the same time, better dimension reduction methods and model training methods can be employed for more abundant features to achieve better prediction results.

Third, this paper employs the method of dividing the time window to convert the original sensor data into a training unit for feature extraction. The size of the time window may affect the prediction results. However, this paper only set the size through several experimental attempts. With a more appropriate time window size found, the prediction performance could be further improved.

VII. Conclusion

In this paper, we proposed an automatic psychological stress sensing method using mobile phone sensing data. First, we extract various discriminative features from multi-modality phone sensing data, in which both absolute and relative features are considered to make the model more personalized. Then, to tackle the challenge of label insufficiency, we further developed a co-training based method for stress level classification to handle the insufficient labeled data problem. We evaluated our model based on an open dataset, and the experimental results verify its advantages over other baselines.

Acknowledgment

This work is supported by the National Key Research and Development Program of China (No. 2018YFB1004403) and the National Natural Science Foundation of China (No.61772045).

References

[1] America College Health Association. Fall 2015 reference group executive summary. 2015 [2017-01-23].

[2] Kirsten S. Statistics on college student stress. 2015[2017-01-23]. http://stress.lovetoknow.com/Statistics_on_College_Student_Stress

[3] Selye H. Stress in Health and Disease. Oxford: Butterworth-Heinemann, 1974

[4] Kahneman D, Tursky B, Shapiro D, et al. Pupillary, heart rate, and skin resistance changes during a mental task. Journal of Experimental Psychology, 1969, 79(1): 164-167

[5] Jiangtao Wang, Yasha Wang, Daqing Zhang, Feng Wang, Haoyi Xiong,Chao Chen, Qin Lv, Zhaopeng Qiu (2018). Multi-Task Allocation in Mobile Crowd Sensing with Individual Task Quality Assurance. IEEE Transactions on Mobile Computing.

[6] Jiangtao Wang, Yasha Wang, Daqing Zhang, Feng Wang, Yuanduo He, Liantao Ma: PSAllocator: Multi-Task Allocation for Participatory Sensing with Sensing Capability Constraints. The 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2017); 02/2017, Portland, USA.

[7] Wang Rui, Chen Fanglin, Chen Zhenyu, et al. StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. Proc of the 2014 ACM Conf on Ubiquitous Computing. NewYork: ACM, 2014: 3-14

[8] Hetz C, Martinon F, Rodriguez D, et al. The unfolded protein response: Integrating stress signals through the stress sensor IRE1. Physiological Reviews, 2011, 91(4): 1219-1243

[9] Healey J A, Picard R W. Detecting stress during real-world driving tasks using physiological sensors. IEEE Trans on Intelligent Transportation Systems, 2005, 6(2): 156-166

[10] Mozos M, Sandulescu V, Andrews S. Stress detection using wearable physiological and sociometric sensors. Int Journal of Neural Systems, 2017, 27(2): 1-17

[11] Lu Hong, Frauendorfer D, Rabbi M, et al. StressSense: Detecting stress in unconstrained acoustic environments using smartphones. Proc of the 2012 ACM Conf on Ubiquitous Computing. New York: ACM, 2012:351-360

[12] Lin Huijie, Jia Jia, Guo Quan, et al. Psychological stress detection from cross-media microblog data using deep sparse neural network. Proc of the 2014 IEEE Int Conf on Multimedia and Expo. Piscataway, NJ: IEEE, 2014: 1-6

[13] Lin Huijie, Jia Jia, Guo Quan, et al. User-level psychological stress detection from social media using deep neural network. Proc of the 22nd ACM Int Conf on Multimedia. New York: ACM, 2014: 507-516

[14] Xue Yuanyuan, Li Qi, Jin Li, et al. Detecting adolescent psychological pressures from micro-blog. LNCS 8423:Proc of the 3rd Int Conf on Health Information Science. Berlin: Springer, 2014: 83-94

[15] Jin Li, Xue Yuanyuan, Li Qi, et al. Integrating human mobility and social media for adolescent psychological stress detection. LNCS 9643:Proc of the 21st Int Conf on Database Systems for Advanced Applications. Berlin: Springer, 2016: 367-382

[16] Mehrotra A, Pejovic V, Vermeulen J, et al. My phone and me: Understanding people's receptivity to mobile notifications. Proc of the 2016 CHI Conf on Human Factors in Computing Systems. New York: ACM, 2016: 1021-1032

[17] Xiong H, Huang Y, Barnes L E, et al. Sensus: A cross-platform, general-purpose system for mobile crowdsensing in human-subject studies. Proc of the 2016 ACM Int Joint Conf on Pervasive and Ubiquitous Computing. New York: ACM, 2016: 415-426

[18] Canzian L, Musolesi M. Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. Proc of the 2015 ACM Int Joint Conf on Pervasive and Ubiquitous Computing. New York: ACM, 2015: 1293-1304

[19] Lu Hong, Frauendorfer D, Rabbi M, et al. StressSense: Detecting stress in unconstrained acoustic environments using smartphones. Proc of the 2012 ACM Conf on Ubiquitous Computing. New York: ACM, 2012:351-360

[20] Bogomolov A, Lepri B, Ferron M, et al. Daily stress recognition from mobile phone data, weather conditions and individual traits. Proc of the 22nd ACM Int Conf on Multimedia. New York: ACM, 2014: 477-486

[21] Wang Rui, Harari G, Hao Peilin, et al. SmartGPA: How smartphones can assess and predict academic performance of college students. Proc of the 2015 ACM Int Joint Conf on Pervasive and Ubiquitous Computing. New York: ACM, 2015: 295-306

[22] Shiffman S, Stone A, Hufford R. Ecological momentary assessment. Annual Review of Clinical Psychology, 2008, 4: 1-32

[23] Lu Hong, Yang Jun, Liu Zhigang, et al. The Jigsaw continuous sensing engine for mobile phone applications. Proc of the 8th ACM Conf on Embedded Networked Sensor Systems. New York: ACM, 2010: 71-84

[24] Zheng Yu, Liu Furui, Hsieh P. U-air: When urban air quality inference meets big data. Proc of the 19th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2013: 1436-1444

[25] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. Proc of the 11th Annual Conf on Computational Learning Theory. New York: ACM, 1998: 92-100

[26] Chan L, Swain V D, Kelley C, et al. Students' Experiences with Ecological Momentary Assessment Tools to Report on Emotional Well-being[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2018, 2(1): 3.

[27] Zhang X, Li W, Chen X, et al. MoodExplorer: Towards Compound Emotion Detection via Smartphone Sensing[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2018, 1(4): 176.