1. Introduction

Lots of students are suffering stress and strain of student life nowadays. However, most of time stress and strain of a student is not obvious. For judging students’ mental states as well as further research, it is essential for us to acquire large quantity of information from each student. From the paper ‘StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students using Smartphones’, it shows that smartphones have significant meaning in analyzing mental state of students. In addition, the study also mentioned that some other features will also have some invisible influence on students’ mental states which including positive affect, conversation levels, sleep, activity and so on. We totally collect 10 useful features which are ‘activity’, ’audio’, ‘Bluetooth’, ‘conversation’, ‘dark’, ‘GPS’, ‘phonecharge’, ‘phonelock’, ‘wifi’, ‘wifi\_location’. By selecting features some more important features among collected features as well as analyzing and processing them, we can make some prediction on students’ mental state with a high accuracy.

1. Dataset

Description

In total, there are 10 features we collect from students. However, the method of collecting these features are different.

Although we use various methods to collect features, the category of timestamp is the same. The Timestamp in dataset is the Unix time when the inference was collected. The time zone is Eastern time zone.

In Physical Activity Inferences, we totally collect two items of information. The first is timestamp and the second is activity interface. Student’s Activity is detected every two or three seconds and when an activity is generated, we will note down the timestamp. In addition, activity has four different interface ID from ‘0’ to ‘3’ which represent four various activity state. In detail, ‘0’ is ‘Stationary’, ‘1’ is ‘Walking’, ‘2’ is ‘Running’ and ‘3’ is ‘unknown’ which means we do not know what this student is doing.

About the dataset of Audio, this dataset is mainly about the participants’ audio state. There are two pieces of information which are timestamp and audio inference. In this dataset, audio inference is generated every second and we note down the timestamp at the same time. There are also four audio inference id from ‘0’ to ‘3’ which represents ‘Silence’, ‘Voice’, ‘Noise’ and ‘Unknown’ in the same order.

For conversation, this dataset is used to record the start time and the end time of each conversation. Therefore, it only has two element which are start\_timestamp and end\_timestamp. If the participant does not have a conversation, the time stamp will not be recorded.

We also record the information of GPS Location. Because the tool which is used to gather information from participant is a mobile application, it is easy for us to collect large quantity of information about the GPS Location of participants. In this dataset, besides the timestamp, the source of GPS coordinate(provider), network\_type, accuracy, latitude, longitude, altitude, bearing, speed and travel state are also included. Moreover, the GPS coordinates were collected every 10 minutes.

In the dataset of Bluetooth, we collected some useful information about the connection between participants’ mobile phone and other equipment. The participants’ Bluetooth scan log file includes timestamp, the MAC address of surrounding Bluetooth device(MAC), general characteristics and capability of a device(class\_id) and signal strength(level). Time is recorded every 10 minutes.

The information of WIFI state is vital to the research. So, we gather some WIFI AP scan log file which may useful. In this dataset, the information of time, BSSID, freq and level is collected. In detail, BSSID is AP’s MAC address and freq is AP's working channel frequency. Level means signal strength.

We believe that the participants’ locations in campus will also have influence on participants’ mental state. With the help of WIFI, we can roughly get participants location in the campus and this is recorded is the location information in the WIFI Location dataset.

The environment of mobile phone can have a side reflect of a participant’s mental state. Therefore, in the dark dataset, we note down start time and end time when participant’s mobile phone has already been in a dark environment for more than one hour.

The Phone Lock dataset describes the start time and the end time that participant’s phone has locked for more than one hour.

Finally, in the dataset of Phone Charge, we only record the start time and end time of charging mobile phone. Only the charging duration which is over one hour can be recorded in this dataset.

For all these datasets, if there are some data loss of a certain participant, we ignore the lost data. For all the exists data, we use the method of min-max normalization to modify the real data which can simplify data and easy to analyze.