计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20170809
56(3):611-622,2019

# 基于智能手机感知数据的心理压力评估方法

王 丰[1,2,5] 王亚沙[1,3] 王江涛[1,2] 熊昊一[4] 赵俊峰[1,2] 张大庆[1,2]
[1](高可信软件技术教育部重点实验室(北京大学) 北京 100871)
[2](北京大学信息科学技术学院 北京 100871)
[3](北京大学软件工程国家工程研究中心 北京 100871)
[4](密苏里科技大学计算机科学系 美国密苏里州罗拉 65409)
[5](计算机网络和信息集成教育部重点实验室(东南大学) 南京 210018)
(wangfeng2013@pku.edu.cn)

# Mental Stress Assessment Approach Based on Smartphone Sensing Data

Wang Feng[1,2,5], Wang Yasha[1,3], Wang Jiangtao[1,2], Xiong Haoyi[4], Zhao Junfeng[1,2], and Zhang Daqing[1,2]
[1](*Key Laboratory of High Confidence Software Technologies* (*Peking University*), *Ministry of Education*, *Beijing* 100871)
[2](*School of Electronics Engineering and Computer Science*, *Peking University*, *Beijing* 100871)
[3](*National Research Center of Software Engineering*, *Peking University*, *Beijing* 100871)
[4](*Department of Computer Science*, *Missouri University of Science and Technology*, *Rolla*, *MO*, *USA* 65409)
[5](*Key Laboratory of Computer Network and Information Integration* (*Southeast University*), *Ministry of Education*, *Nanjing* 210018)

**Abstract** Mental stress is harmful on individuals' physical and mental well-being. It is often easy to be overlooked in the early stage, leading to serious problems. 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 with various embedded sensors, which can capture users' context data contains movement, sound, location and so on, it is an alternative way to access users' behavior by smartphones, which is less intrusive. This paper proposes an automatic and non-intrusive stress detection approach based on mobile sensing data captured by smartphones. By extracting reasonable features from the perceived data, a more efficient psychological stress assessment method is proposed. First, we generate lots of features represent users' behavior and explore the correlation between mobile sensing data and stress, then identify discriminative features. Second, we further develop a semi-supervised learning based stress detection model. Specifically, we use techniques such as co-training and random forest to deal with insufficient data. Finally, we evaluate our model based on the StudentLife dataset, and the experimental results verify the advantages of our approach over other baselines.

**Key words** mental stress; context awareness; feature engineering; automatic detection; machine learning

**收稿日期**:2017-10-25;**修回日期**:2018-12-12
**基金项目**:国家自然科学基金项目(61772045)
This work was supported by the National Natural Science Foundation of China (61772045).
**通信作者**:王亚沙(wangyasha@pku.edu.cn)

**摘 要** 较大的心理压力对大学生的心理和生理均会产生危害.心理压力往往在前期容易被人忽视,从而导致严重的问题.因此,如果能较早发现心理压力,并进行合理干预,有益于人的身心健康.传统心理压力检测方法以问卷调查和借助专业设备的评估为主,但都存在成本较高,且对被评估对象侵扰较大等不足.另一方面,随着智能手机的快速普及,通过手机中内置的位置、声音、加速度等多种传感器感知用户的行为习惯,并基于感知数据评估用户心理压力成为一种低成本、低侵扰的心理压力评估手段.在此背景下,针对基于智能手机感知数据分析,对评估大学生心理压力的方法展开了研究,从感知数据中提取合理的特征,提出了一种更高效的心理压力评估方法.首先,讨论了如何从原始的手机感知数据提取出合理的特征;其次,介绍将心理压力评估转化为分类问题,并使用半监督学习方法构造分类模型;最后,在开放数据集 StudentLife 上对上述模型进行实验验证.实现结果表明:该方法在心理压力检测精确度和召回率等方面均优于基线方法.

**关键词** 心理压力;情境感知;特征工程;自动评估;机器学习

**中图法分类号** TP391.4

在快节奏的生活中,越来越多的人会受到心理压力的影响.美国大学健康协会(America College Health Association)在 2015 年秋季出具的心理学报告[1]中表明有 57.7%的学生在过去的 12 个月中,至少一次感受到"非常焦虑".同时,有研究表明人感受到的压力会显著影响心理和行为习惯,当人们感受到巨大压力时,往往会显得焦虑不安、失眠,严重的可能会导致心理甚至生理疾病[3].更有调查指出,在中途辍学的大学学生中,有 64%是受到了精神方面的疾病的影响[2].这种由压力带来的心理疾病,在初期难以被重视,可能会发展成严重的问题,进而对一个人造成巨大影响[3].因此,在心理压力转化成严重的心理问题之前及时检测心理压力对大学生心理健康方面有很重要的意义.

近些年来,心理压力的检测越来越受到重视.就如何合理检测人的心理压力,心理学领域已经做出了很多的研究.最为传统的方法就是利用依据心理学理论制定的调查问卷,由于其背后的理论支撑,这种方法现在依然是使用最为广泛的调查方法.其次,人的心理压力也可以通过专业的仪器进行监控,比如人的皮肤电阻可以和某些心理指标建立联系[3],监控皮肤电阻即可完成对心理指标的检测,且结果可信度较高.但是,这些方法并不适用于日常的心理压力监控.无论是基于问卷还是专业仪器,由于都需要用户提供额外的时间成本参与测试,对用户造成了较大的侵扰,导致参与积极性不高;因此,我们希望找到一种自动的、低成本、低侵扰的方法来实现对用户的实时心理压力监控.与此同时,智能手机已经成为了人们生活的必需品.为了满足人们生活中更多的需求,手机中也加入了越来越多的感知设备(比如加速度传感器、声传感器、光传感器等).在日常生活中,手机可以持续记录大量和人的日常生活相关的的感知数据,包括运动信息、位置信息、手机使用信息等数据.

与此同时,有研究表明:人的心理压力状态会在人的行为习惯上得到反映,比如在压力较大的状态下,人们往往呈现出活动积极性降低、频繁使用手机、睡眠质量较低等状态[4].手机提供的感知数据可以反映用户的行为习惯特征,而用户的行为习惯可能和心理压力有某种联系,因此可以尝试利用手机感知数据,通过机器学习的方法探究手机感知数据和用户的心理状态之间的联系.

建立两者之间的联系存在着 2 项技术挑战.

1) 如何将基础的手机感知数据转变为有意义的分类特征.原始的手机感知数据以日志形式存在,每一个时刻会生成相应的数据,而对心理压力的评估需要综合某一段时间内的用户行为来进行判断,因此需要将日志数据进行整合,并提取特征.

2) 如何解决带标记的训练数据不足的问题.在数据采集阶段,智能手机可以低成本、持续采集各类数据并生成特征向量.然而,每个特征向量所对应的标记数据需要用户主动标注,无法大量获取,导致带标记的训练数据稀少.因此,如何在带标记的数据不足的情况下进行准确的模型训练是本文要解决的另一个技术挑战.

本文针对上述挑战,利用机器学习的手段,提出了基于智能手机感知数据的心理压力评估方法.主要贡献包括 3 个方面:

1) 对原始的手机感知数据进行分析,提出了特征提取与筛选的方法,基于这些特征生成用于训练

分类模型的样本.通过特征抽取制定出一系列的方法,将抽象的日志数据转化为了带有标记的样本数据,通过对特征进行筛选得到真正对分类有用的特征,减少了数据维度的冗余.

2) 使用半监督学习模型应对训练数据不足的问题,本文充分利用大量的没有标注的数据,使用协同训练(co-training)对这些数据进行标注,并迭代训练,提高模型分类精度.

3) 使用了现有的开放数据集(Dartmouth StudentLife)[5]进行验证,结果证实了本文提出的方法可以对人的心理压力进行有效地监控,且效果优于其他基线方法.

## 1　相关工作

### 1.1　基于专业感知设备感知情绪

由于人的心理变化必然会导致某些生理指标的变化,因此很多研究致力于利用可穿戴设备对人的日常心理压力进行监控[6-9].通常,这些设备上集成了专门的传感器,可以感知人的生理指标的变化,例如皮肤的电阻、体温、心率、血压等.由于可以直接获取这些生理数据,所以基于可穿戴设备进行的感知往往很有说服力,但是代价是人们需要佩戴这些专业设备,这就带来了成本较高,对人打扰程度比较大的问题.而基于手机数据的心理压力感知可以将人从这些专业设备中解脱出来.

### 1.2　基于社交网络感知情绪

随着互联网技术的不断发展,社交网络也在迅猛发展.基于社交网络的和压力相关的研究也越来越多.Lin 等人[10]基于微博用户的数据,利用深度稀疏神经网络对用户的心理压力程度作出判断;之后,Lin 等人[11]基于微博数据,使用卷积神经网络检测用户压力;在青少年方面,Xue 等人[12]从青少年的推特消息出发,提取了一系列特征,利用分类器来了解青少年的潜在压力类别和压力水平;Jin 等人[13]提出了一种基于协同训练的方法,结合微博和轨迹信息来完成对青少年的压力检测.

这些基于社交网络的工作通过对人们在社交网络上的行为和发布的内容,利用自然语言处理以及深度学习可以自动并且在对用户低侵扰的情况下进行心理压力的评估.然而,这些工作也有一定的局限性,那就是这些方法只能聚焦于那些频繁使用社交网络的用户,对于不常使用社交网络的用户,由于缺乏训练数据较难对其进行心理压力的预测.同时,由于人们并不会一直使用社交网络,因此无法通过社交网络数据对一个用户进行不间断的监控,这些问题也是基于手机感知数据的工作所重视的地方.基于社交网络对用户的心理压力进行评估的方法可以和本文工作形成互补.

### 1.3　基于智能手机数据感知情绪

近些年来智能手机不断进步,为了适应各种使用场景,提供了更强大的功能,传感器种类越来越多,精度也越来越高,它们记录了用户使用手机的习惯,提供了大量有价值的数据,陈龙彪等人[14]对智能手机在普适计算领域的应用展开过深入探究.因此也有了越来越多的基于智能手机的感知数据开展的研究,尤其在情感分析领域.这些工作大致上可以分为 2 种类型:1)探究智能手机数据和用户情绪的相关性;2)训练模型利用智能手机感知数据对用户情绪进行预测.

在第 1 类的工作中,Wang 等人[15]利用 StudentLife 数据集,从中提取了多维特征,利用线性回归的方法,分析了用户在学期中的行为习惯和用户在学期中的心理压力,沮丧程度等多种心理指标的关系;Mehrotra 等人[15]利用用户日常生活中的手机通信数据,应用使用习惯等提取了一系列特征,通过线性回归的方法,分析了多维特征和用户情绪沮丧程度的相关性;Xiong 等人[16]利用大学生的 GPS 和 POI 数据,利用线性回归的方法,分析了不同的行为习惯和用户社交焦虑的相关性.这些工作对本文的特征提取工作有着很大的指导意义,缺陷是这些工作并没有完成预测模型,本文在这些工作的启发下完成,并构建了模型对用户心理压力进行预测.

在第 2 类工作中,Canzian 等人[17]通过 GPS 数据实现了对用户的沮丧程度(depression)进行预测,(他们)从用户的 GPS 数据中提取多维度的特征,从 PHQ-9 问卷中得到用户的沮丧程度,利用线性回归的方法,建立了特征和用户的沮丧情绪之间的联系,并使用 SVM 构建预测模型,达到了 80%的准确率.该工作的主要问题在于,其致力于研究用户一段时间内的沮丧程度,想要达到较好的预测效果,则需要较长时间的 GPS 数据(通常是 2 周左右),无法对用户短时期的心理状态(以 1 天为时间窗口)进行实时评估.同时,该工作的事实依据选用的是问卷调查结果,问卷只在实验结束时进行了一次,所以其无法对用户的实时心理状态进行刻画.Lu 等人[18]通过手机感知的声音数据,可以对用户在不同场合的紧张程度进行实时预测,该方法中用户需要佩戴 2 部手

机采集声音信息,并利用声学的相关方法提取了一系列特征.使用专业的手环获取用户的真实紧张状态,最后利用高斯混合模型(GMM)对这两者得到的数据建立联系,实时预测用户紧张程度,并讨论了如何利用通用模型得到可以更好适配单独用户的个性化模型,最后个性化模型达到了约 80% 的准确率.这个工作虽然能够实时且精确地预测用户的紧张程度,可是其数据采集设备对用户的打扰程度较高,不适用于日常生活中对用户进行心理压力的检测.Bogomolov 等人[19]在使用手机收集到的通信数据(包括手机通话数据和短信数据)之外,还使用了天气数据和用户的心理学问卷信息,利用决策树构建模型,在预测用户是否有压力的二分类问题中,取得了 72% 的准确率.但是,和本文工作相比,有 2 个不同:1)这个工作中也要求用户完成心理学问卷,并且结果反映心理学问卷对预测准确度有很大的贡献,而当只使用收集数据的时候,模型很难达到较高的精度,但心理学问卷需要耗费用户的大量时间,且对用户的打扰程度较大,本文工作不需要用户去额外填写这些问卷,同时也达到了可以接受的模型精确度;2)这个工作中没有考虑如何利用未标记数据,而在本论文中,本文通过协同训练使用了大量的没有标记的数据用于训练,可以有效提高模型的预测精度.

## 2 数据集介绍与问题定义

### 2.1 数据集介绍

为了训练模型,本文使用了开放数据集 StudentLife[5,20].StudentLife 数据集是 Dartmouth 学院的研究团队于 2013 年在 StudentLife Study 中获取的数据集.在研究中,学生被要求使用安装有 StudentLife 程序的手机,手机在后台记录了一系列信息.这次研究一共收集了 49 个学生连续 10 周的感知数据.这些数据大致分为 4 个类别:传感器数据、EMA (ecological momentary assessment)数据、问卷调查数据、学业数据.

本文使用了 StudentLife 数据集中的传感器数据和 EMA 数据,其中,传感器数据包含了所有利用手机传感器得到的数据,描述了人在使用手机的过程中活动信息、环境信息、手机使用习惯信息等.EMA 数据是即时的生理状态评估,用户在使用软件时会不定期收到简单的问题,用户对自己的心理状态做出评估后,实时反馈给服务器[21].

### 2.2 问题定义

本文目标是利用手机的感知数据建立模型预测用户的心理压力,即通过机器学习算法训练一个分类器用来完成分类问题.因此需要明确如何形成用以训练的样本数据.

训练数据使用 StudentLife 数据集中的传感器数据,原始的文件一共分为 10 个类别,如表 1 所示.在试验中,利用这些数据一共生成了 49 个学生的 2 167 个有标记样本数据,样本生成的方法在第 4 节会进行详细介绍.

**Table 1 StudentLife Sensing Data**

**表 1 StudentLife 感知数据**

| Type Number | 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 | WiFi | WiFi log |
| 10 | WiFi location | WiFi AP location |

不同于以往的工作使用心理学问卷得到的结论作为样本数据的标注,为了追求预测的实时性,需要使用用户实时反馈的心理状态评估,这里利用了 EMA 数据.

在 EMA 数据中,真正需要关心的是用户反馈的心理压力数据,EMA 问卷中关于心理压力的问题是(figure),用户有 5 个选项可供选择,分别是:

1) 有一点压力(a little stressed);

2) 确定有压力(definitely stressed);

3) 压力很大(stressed out);

4) 感觉较好(felling good);

5) 感觉好极了(felling great).

为了使类目之间的显著性更大,综合以往的工作[19],本文将这 5 类进行融合,将 1)~3)合并为有压力,将 4)5)合并为无压力.从而这个分类问题变为二分类问题.2 类的物理意义也更加明确.

问题定义:通过感知数据集 $D$,从 $D$ 中提取出特征集合 $F=\{f_1,f_2,\cdots,f_n\}$,从数据集 $D$ 中,按照特征抽取规则得到样本矩阵 $\boldsymbol{X}_{m\times n}$($m$ 行,每一行是一个样本),从 EMA 数据中获取标注集 $y_m$($y_i$ 对应

样本矩阵第 $i$ 行的标注),利用 $\boldsymbol{X}_{m\times n}$ 和 $y_m$ 训练得到二分类分类器 $C$,使得 $C$ 能够在给定输入后完成二分类任务.

## 3　方法概览

本文所设计的方法框架图如图 1 所示,方法共包括在线预测部分和离线训练部分:

1) 离线训练部分. 将原始数据进行特征提取、特征筛选,得到数据样本,包括有标记样本和未标记样本,然后利用协同训练得到分类模型.

2) 在线预测部分. 从数据源获取传感器数据,对数据进行特征提取,通过接口调用预测模型,可以实时完成对用户心理压力的评估.

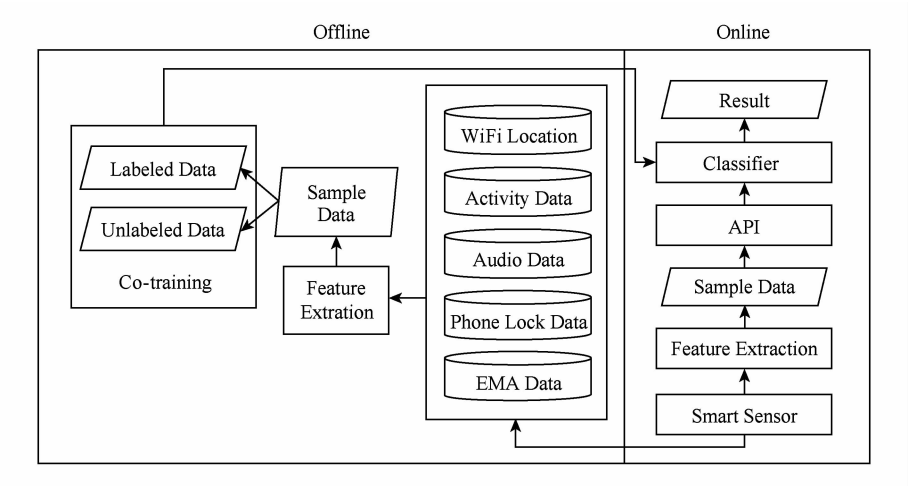


Fig. 1　Mental stress assessment framework

图 1　基于移动感知数据的心理压力检测框架

整体的工作流程如下. 首先,在用户日常使用过程中,手机会不断收集感知数据,然后这些感知数据会被传输到服务器上,通过特征抽取的方法,从中提取出一系列有意义的特征,继而通过特征筛选的方法保留有利于分类的特征,将日志数据会被转化为样本数据. 在生成有标记样本的同时,也会利用相似的方法得到大量的未标记样本,这些样本将被用于协同训练,从而提升分类模型的分类效果,本文构建的协同训练分类器基于随机森林分类器. 分类器的训练工作都是离线完成,从而用户可以在线调用已经完成的分类器,实现对用户的心理压力的实时预测的工作.

## 4　核心方法

### 4.1　特征抽取

特征抽取的工作是为了将日志文件转变为可用于分类的样本数据.

首先需要引入时间窗口的概念,目的是将批量的日志数据转化为一个样本. 对于每一个样本,只使用这个样本对应的时间窗口内的传感器数据进行特征提取,生成样本,并打上对应的标记.

本文选择了 24 h 作为时间窗口,对于一个 EMA 数据,选取用户反馈该结果的时刻之前的 24 h 的感知数据用以生成对应这个 EMA 结果标记的样本. 选定时间窗口为 24 h 有 3 个原因:1)人每天的行为是有规律性的,很多指标不会剧烈变化,这样会更有利于控制变量,进行样本间的对比,选取其他的时长作为时间窗口,便失去了这个保障;2)以 24 h 为时间窗口可以采样到更多有意义的信息,比如用户的睡眠信息,用户的睡眠情况和用户的心理状态息息相关,以 24 h 为时间窗口可以很自然地采样到这个数据;3)以 EMA 数据作为样本的标注,根据每一条用户反馈的 EMA 数据生成一个样本,在该数据集中,EMA 数据的频率接近每个用户每天 1 次,那么按照天来生成样本,更符合实际操作中的物理含义.

在特征抽取的过程中,需要对特征进行细化. 例如描述用户周围人数的特征. 用户在白天周围人数较多和在晚上周围人数较多实际上对应着不同的含义. 这是因为用户在白天活动较多,接触到的人也比较多,而晚上一般都会回到宿舍,周围接触到的人比较固定,数量较少. 而如果一个用户晚上接触到的人也比较多,那么可能会说明这个用户在参加某种活动,这种发现是有意义的,而如果以天为单位去度量这一特征,则无法得出这个结论. 这说明对特征进行

细化可以得到更多信息.在特征抽取的过程中,本文使用了2种特征细化方法,分别是按照时间细化和按照POI(point of interest)细化.

1) 按照时间细化.将时间分成白天(8:00—18:00)和夜间(18:00—8:00),在之后的特征提取中,会把每一种维度结合时间进行细化考虑.

2) 按照POI细化.POI描述了用户所处的位置,和时间类似,用户的表现按照POI进行划分也可以得到更多的信息.

同时,将特征分为2个类别,绝对特征和相对特征.绝对特征只描述当前时间窗口内的数据,而相对特征则由当前时间窗口的数据和用户的历史数据对比得到.

本节将详细阐述针对不同数据集提取的特征以及提取这些特征的指导思想.

4.1.1 绝对特征 $F_a$

4.1.1.1 POI相关特征

Jin等人[17]的工作是基于用户的轨迹对用户的心理焦虑程度进行预测,提取了POI相关的特征,并证明了用户访问不同POI的频率和用户的心理焦虑程度是有一定的相关性的.本文从该角度出发,基于StudentLife数据集的WiFi location数据,获取用户POI信息.

WiFi location数据是手机的WiFi模块的扫描结果,每1～2 min会记录下一次扫描结果,出于对隐私的保护,数据只给出了接入点的位置信息(接入点和学校内POI的关系).本文将POI分为3类:1)教学区,包括教学楼、实验室、图书馆.2)宿舍区,包括学生公寓、宾馆.3)饮食康健区,包括食堂、健身房、艺术馆.进而得到每个用户在每个时间所处的POI类别,这为通过POI信息细化特征提供了帮助,正如通过时间细化是把时间分为了白天和夜间,对各个指标按这2个时间段分别计算;通过POI细化则是把位置分成教学区、宿舍区、饮食康健区,对不同的区域进行计算.

与此同时,我们希望利用一个指标反映用户每天在各类POI花费的时间.这类信息可以反映一个用户每天活动的行为习惯,比如热爱学习的用户每天会花费更多时间在教学区,而较宅的用户在宿舍区的时间更多,热爱健身的人可能在饮食康健区停留更久.基于该思想的指导统计了每名用户在时间窗口内的POI数据,由于采样的频率基本恒定,每一类POI的条目数量正比于用户在每一类POI停留的时间.

除了考虑用户在某种类型POI所处的时间长短,这3种类型的数据构成了一个分布,本文引入了熵来表达这个分布的特征.对于一个多类分布 $X$,定义熵为

$$H(X)=-\sum_i P(x_i)\lg(P(x_i)).$$

结合上文,形成的基于POI的特征如表2所示:

**Table 2 POI Based Features**

**表2 基于POI的特征**

| Feature Number | Meaning |
|---|---|
| 1 | POI number in teaching area, daytime |
| 2 | POI number in accommodation area, daytime |
| 3 | POI number in eating and heathy area, daytime |
| 4 | POI number in teaching area, nighttime |
| 5 | POI number in accommodation area, nighttime |
| 6 | POI number in eating and healthy area, nighttime |
| 7 | Entropy of POI number distribution, daytime |
| 8 | Entropy of POI number distribution, nighttime |

1) 活动信息相关特征

POI信息可以在一定程度上反映用户的活动情况,但其粒度较大.StudentLife数据集中通过利用加速度传感器收集到的数据,使用物理运动分类器[22]对原始数据进行分类,得到用户在某个时刻的运动状态:静止(stationary)、走(walking)或跑(running),还有一类标签是未知.为了得到更加精细的数据,本文利用了这类数据.传感器采样的频率是恒定的,所以在一个时间窗口内,用户的某一类标签数目的多少就对应了用户处在这种运动状态下的时间长短.本文对时间窗口内每一类的标签数量进行统计,并计算熵,同时考虑按照时间进行细化,得到特征如表3所示:

**Table 3 Activity Features**

**表3 活动信息相关特征**

| Feature Number | Meaning |
|---|---|
| 1 | Number of Stationary, daytime |
| 2 | Number of Stationary, nighttime |
| 3 | Number of Walking, daytime |
| 4 | Number of Walking, nighttime |
| 5 | Number of Running, daytime |
| 6 | Number of Running, nighttime |
| 7 | Entropy of label distribution, daytime |
| 8 | Entropy of label distribution, nighttime |

除此之外,可以反映用户活动信息的还有 GPS 信息. GPS 传感器采样以 10 min 为间隔,对用户的经纬度信息进行采样,在 Jin 等人[17]的工作中,研究者们探究了用户的移动距离对用户的心理焦虑程度的影响,本文利用 GPS 信息计算出用户在时间窗口内的白天移动距离和夜间移动距离,移动距离为相邻的采样点间的距离的累加,2 经纬度点间的距离计算方式为

$$D((x_1,y_1),(x_2,y_2))=$$

$$R\times\sqrt{\left((x_1\times x_2)^2+\cos^2\left(\frac{x_1+x_2}{2}\right)\times(y_1\times y_2)^2\right)},$$

其中$(x,y)$为一个数据点的经纬度.

2) 声音相关特征

从 Lu 等人[13]的工作中已经得知,用户所处的环境的声音信息可以用来预测用户的紧张程度,这说明声音信息和用户的心理指标是紧密相关的. 利用手机的感知数据可以获取用户所处环境的声音信息. StudentLife 数据集利用声音分类器和对话分类器获取了一系列的声音相关信息:其中一个用以感知手机所处环境是否有声音,另一个用以感知这个声音是否是人声. 这些信息包括用户每一时刻所处环境的声音类型:安静(silence)、噪音(noise)、人声(voice). 对于这一类的数据,采用类似于处理用户活动信息的方法,结合时间细化,统计了用户白天或者夜间 3 类标签的数量,即用户处在相应空间的时长. 同时也计算了熵用以描述此分布.

数据集中还包含了用户所处空间的对话信息,结合心理学角度发现,用户的心理压力程度会影响其是否乐于交流,本文据此统计了用户在不同的时间段以及不同的类别 POI 所进行的对话次数和时长,声音相关的特征维度如表 4 所示.

3) 社交相关特征

有研究表明:当用户处于较为压抑的状态下时,往往表现得更加自闭,不愿与人交流. 因此,可以提取特征来刻画用户的社交信息. 区别于以往的工作,研究者可以通过用户的电话短信数据获取用户的社交相关的信息. 手机感知数据中不包含相关信息,但可以使用蓝牙扫描数据近似刻画社交信息. 手机会定期进行扫描,并记录下扫描到的蓝牙设备. 基于我们的认知,可以被扫描到的蓝牙设备大多数为智能手机、电脑等设备. 对于孤僻的用户而言,不会倾向于去人流密集的地方,蓝牙扫描记录的设备数量也就越少;相反,对积极外向的用户而言,热衷于参加各种活动,那么记录中设备数量也就会越多. 据此,本文统计了用户在不同时间段,在不同 POI,蓝牙扫描到的设备数量,以此来描述用户所处环境的热闹程度,也一定程度上反映了用户的社交习惯. 具体的特征如表 5 所示.

**Table 4　Conversation Features**

**表 4　声音相关特征**

| Feature Number | Meaning |
|---|---|
| 1 | Number of Silence, daytime |
| 2 | Number of Silence, nighttime |
| 3 | Number of Voice, daytime |
| 4 | Number of Voice, nighttime |
| 5 | Number of Noise, daytime |
| 6 | Number of Noise, nighttime |
| 7 | Entropy of label distribution, daytime |
| 8 | Entropy of label distribution, nighttime |
| 9 | Number of dialog, daytime |
| 10 | Total time of dialog, daytime |
| 11 | Number of dialog, nighttime |
| 12 | Total time of dialog, nighttime |
| 13 | Number of dialog, teaching area |
| 14 | Total time of dialog, teaching area |
| 15 | Number of dialog, accommodation area |
| 16 | Total time of dialog, accommodation area |
| 17 | Number of dialog, eating and health area |
| 18 | Total time of dialog, eating and health area |

**Table 5　Bluetooth Features**

**表 5　蓝牙相关特征**

| Feature Number | Meaning |
|---|---|
| 1 | Number of scanned device, daytime |
| 2 | Number of scanned device, nighttime |
| 3 | Number of scanned device, teaching area |
| 4 | Number of scanned device, accommodation area |
| 5 | Number of scanned device, eating and health area |

4) 用户睡眠信息

基于心理学的预知知识,当用户处在一定的心理压力下时,会引起焦虑、睡眠质量差等生理反应,因此,为了预测用户的心理压力,因此为睡眠质量是很重要的一个特征,刻画出用户的睡眠时长以及入睡时间,可以为分类提供帮助.

本文基于手机锁屏的记录,获取用户的睡眠习惯. 每一条记录了一次手机锁屏到开启的起止时间(超过 1 h 才会被记录),在探究过程中发现,每 24 h 都会出现一个较长的记录,且该记录都处在深夜,即

该记录对应了用户夜间的休息(用户在休息的时候不会使用手机,因此会留下一段长度相当于睡眠时长的记录).本文使用用户这条记录的开始时间作为用户入睡时间,构造了用户的睡眠时长以及入睡时间的特征,如表 6 所示:

**Table 6 Sleeping Features**

**表 6 用户睡眠信息**

| Feature Number | Meaning |
|---|---|
| 1 | Sleeping time |
| 2 | Length of sleeping duration |

4.1.2 相对特征 $F_b$

绝对特征在描述问题的时候依然有局限性:某些用户性格比较孤僻,数据显示他接触到的人会一直较少,有的用户相对外向,其接触到的人较多.因此,对于某一个指标的特定值,对一些用户来说较高,而对另一些用户是较少的,相同的值对应了相反的变化,因此本文利用将用户的数据和自身的历史数据进行纵向对比的方法来刻画这种现象.

用于对比的基准是均值.由于时间窗口是 24 h,数据一共包括了 49 名学生超过 70 d 的感知数据.因此对这 49 名学生的 70 d 数据分别按照 4.1 节提到的特征抽取生成样本,再对样本求取平均值,可以得到均值样本(在前面特征抽取部分提到的维度中,并非所有维度都有对比意义,比如熵的对比值就没有物理意义,这种值不会进行对比).

之后,对所有可生成相对特征的维度 $i$,计算其相对特征 $r_i$:

$$r_i=(v_i-avg_i)/avg_i,$$

将得到的 $r_i$ 作为新的特征维度加入原本的特征向量,对之前得到的样本矩阵的每一行进行处理,得到增加了相对特征的样本矩阵.

## 4.2 特征选择

按照上面的方法,一共提取了 83 个维度的特征,但是由于特征都是手工提取,一定会存在很多特征与用户心理压力程度在数学上相关性不足.因此,对特征进行降维是有必要的.

为了特征的可解释性,本文没有采用 PCA 一类的方法进行降维,因此特征筛选问题等价于选取最优子集,而最优子集问题是 NP 难的,从而问题变为利用近似方法选取一个子集,能使分类器达到尽量优的效果,且可以在多项式时间内求解.结合验证过程中使用的随机森林分类器,本文提出了一种基于基尼不纯度的降维方式.

基尼不纯度是用来衡量数据集纯净程度的统计量,在决策树中使用广泛.对决策树而言,随着树的节点不断分裂,目标希望叶子节点中的数据点尽可能属于同一类别,即希望节点样本的"纯度"越来越高.决策树选择节点的划分方式的时候可以依据基尼不纯度来选择特征维度[23].这里有 2 个概念:基尼不纯度和基尼指数,其中基尼不纯度描述了一个数据集的纯净度,而基尼指数则描述了数据集在划分过程中基尼不纯度的变化.

对于数据集 $D$ 来说,设维度集为 $F=\{f_1,f_2,\cdots,f_n\}$,那么基尼不纯度可以用 $I_G(D)$ 来表示:

$$I_G(D)=1-\sum_{k=1}^{|K|}p_k^2,$$

其中,$p_k=\frac{n_k}{|D|}$.

当对 $D$ 按照某一个数据维度进行划分的时候,会得到这个基尼不纯度的变化,也就是基尼指数,当按照 $f_i$ 进行划分的时候,基尼指数为

$$\Delta I(D,f_i)=I_G(D)\times(p_lI_G(D_l)+p_rI_G(D_r)).$$

在决策树选取特征划分时,会选择划分后基尼不纯度变化最大的特征维度 $f_i$ 进行划分,即:

$$f_*=\arg\max_{f_i\in A}\Delta I(D,f_i).$$

基尼不纯度变化越大,说明按照这个维度划分后得到的数据的纯度越高,特征筛选时,应尽量保留这些数据维度.基于这个思想:本文所有特征计算基尼指数,将特征按照基尼指数倒排,按照排序增量选取特征,继而训练分类器评估,筛选结果在第 6 节中展示.

## 4.3 模型训练与在线识别

4.3.1 分类器选择

由于数据集中有标记的样本数量较少,在分类器的选取上,为了避免数据较少引起的分类器过拟合问题,采取了可以有效规避过拟合问题的随机森林(random forest)来作为本文使用的分类器.

随机森林是一个包含若干决策树的分类器,其输出的类别由个别树输出的类别的众数而定.而森林中的每一棵决策树只利用一部分特征进行分类,每一棵决策树使用的样本也是从原始样本集合中通过 Bootstrap 自举方法生成.

4.3.2 协同训练

协同训练(co-training)是一种半监督模型,在有标记的数据量较少的情况下,可以使用大量的未标记数据进行训练,以提升分类器的精度[24].协同

训练需要从 2 个不同的“视角”去分析数据,它要求数据集合有 2 个不同的特征集合,且两者是相互独立的,任意一组数据集合可以训练出预测类别的分类器.协同训练通过结合着 2 个从不同视角出发的分类器,构建出更准确的分类模型.

本文采用了 2 类特征抽取方法,分别得到从原始数据中直接抽取到的绝对特征和基于个人历史数据的相对特征,这 2 类特征在数学上也是相互独立的.对 2 类特征分别构建分类器,利用协同训练可以得到 2 个分类器,具体算法见算法 1:

**算法 1.** Co-training.

输入:有标记数据集 $D_{\text{labeled}}$、无标记数据集 $D_{\text{unlabeled}}$、迭代轮数 $\theta$、每轮选择的样本数 $n$;

输出:分类器 $h_1$,$h_2$.

① 将 $D_{\text{labeled}}$ 在 2 个不同的不相关的特征组合 $f_1$,$f_2$ 上进行投影,分别得到投影后的数据集 $D_{f_1}$,$D_{f_2}$;

② 利用 $D_{f_1}$ 训练出分类器 $h_1$,利用 $D_{f_2}$ 训练出分类器 $h_2$;

③ 对每一个 $D_{\text{unlabeled}}$ 中的未标记数据,利用 $h_1$ 标注,选择置信度最大的 $n$ 个样本,加入 $D_{\text{labeled}}$;

④ 对每一个 $D_{\text{unlabeled}}$ 中的未标记数据,利用 $h_1$ 标注,选择置信度最大的 $n$ 个样本,加入 $D_{\text{labeled}}$;

⑤ $\theta=\theta-1$;

⑥ if $\theta<0 \parallel D_{\text{unlabeled}}=\varnothing$

⑦　Return $h_1$,$h_2$;

⑧ else

⑨　返回算法步骤①继续执行;

⑩ endif

在每一轮迭代中,从未标记样本集中选取随机森林分类器给出的置信度较高的 $n$ 个样本加入训练样本集,直到未标记样本集为空.最终可以得到 2 个随机森林分类器:$h_1$ 和 $h_2$.

4.3.3　在线评估

通过协同训练可以得到 2 个基于不同特征维度的分类器 $h_1$,$h_2$,在预测过程中,需要综合 2 个分类器给出的结果做出评估.

2 个分类器都是基于随机森林得到的.随机森林是由很多单独的决策树组合得到的复合分类模型,随机森林的结果是由单独的决策树输出的类别的众数决定.对二分类 $\{c_1,c_2\}$ 问题来说,设一个包含 $m$ 棵决策树的随机森林,给出分类结果为的决策树分别为 $n_i$ 棵,那么随机森林认为样本属于 $c_i$ 的概率为

$$p^{c_i}=\frac{n_i}{m_i}.$$

在预测过程中,利用如下的公式得到 2 个分类器综合的分类结果,选取概率较大的一个类目作为分类结果:

$$c=\arg\max_{c_i\in C}(p_1^{c_i}\times p_2^{c_i}).$$

模型的训练过程是离线完成的,识别可以在线完成,对给定的用户数据进行特征提取,特征选择之后得到样本,就可以进行在线分类.

## 5　实验验证

### 5.1　实验数据

实验数据采用达特茅斯学院的 StudentLife 数据集,包括 49 名学生超过 10 周的传感器数据和 EMA 反馈数据.对数据集采用前文提出的特征提取和特征筛选方法,生成了有标注数据和未标注数据,数量信息如表 7 所示:

**Table 7　Number of Sample**

**表 7　样本数量**

| Data Type | Stressed | Not Stressed | Total |
|---|---|---|---|
| Labeled Sample | 1 587 | 580 | 2 167 |
| Unlabeled Sample | | | 9 800 |

### 5.2　实验方法

本文在样本集合上使用不同的方法构建模型,并评估模型的效果.在验证过程中,对数据采用 10 折的交叉验证,将数据随机分成 10 份,其中 9 份用于训练得到分类模型,最后 1 份用作验证.本文使用 3 种评价指标,分别是精确率(precision, $Pr$)、召回率(recall, $Re$)、$F$ 值(F-measure, $F_1$).

1) $Pr$:是正确被分到某一类的样本数量占所有被分类器标为该类的样本数量的比例;

2) $Re$:是正确被分到某一类的样本数量占实际这一类的样本数量的比例;

3) $F_1$:是综合精准率和召回率的一个指标,在认为二者权重相等的时候,是取二者的调和平均,也就是 $F_1$ 值:

$$F_1=\frac{2\times Pr\times Re}{Pr+Re}.$$

本文对模型进行了 3 个角度的评估,和基线方法进行对比,对特征筛选的效果进行评估,对协同训练的效果进行评估.

### 5.3 实验结果

#### 5.3.1 和基线方法进行对比

本文提出了基于随机森林的协同训练模型,在本节中,将此方法和基线方法进行对比,作为对比的基线方法包括:

1) 决策树(decision tree);

2) 支持向量机(SVM);

3) $K$ 近邻($K$NN);

4) 逻辑斯蒂回归(logistic regression).

对比的实验结果如图 2 所示:

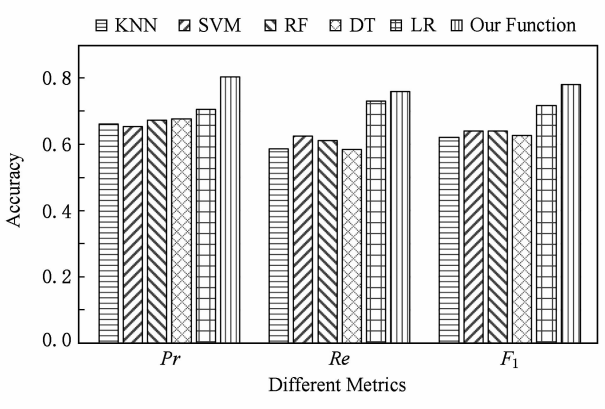


Fig. 2 Compare to other classifier

图 2 对比其他分类方法

可以看出本文方法在 $Pr$,$Re$,$F_1$ 上的表现都要好于直接使用这些分类器. 这得益于协同训练的方法利用了大量的无标记数据,以及随机森林分类器在一定程度上克服了数据样本数量少容易带来的过拟合问题,本文方法在 3 种指标上的值如表 8 所示:

**Table 8 Performance of Our Function**

**表 8 本篇论文方法效果**

| Metrics | Our Function |
|---|---|
| $Pr$ | 0.804 |
| $Re$ | 0.755 |
| $F_1$ | 0.770 |

#### 5.3.2 对特征筛选的评估

基于手工提取的特征具有一定的冗余性,本文采用基于基尼不纯度的特征筛选方法. 对每一个特征,计算基尼不纯度的变化,并依此倒排. 然后从中不断添加特征进行模型训练,并进行评估,得到如图 3 的曲线.

其中 $x$ 轴表示训练使用到的特征数量,$y$ 轴表示每轮迭代后模型的各项评估指标. 从图 3 中可以看出,随着添加的特征数目不断增加,各项评估指标的值首先呈上升趋势,然后趋于平稳. 说明在添加了一定量的特征之后,分类器的效果趋于稳定,那么就不需要继续添加新的特征了. 按照这个方法,最终选取了前 26 个特征用于最后的分类器训练.

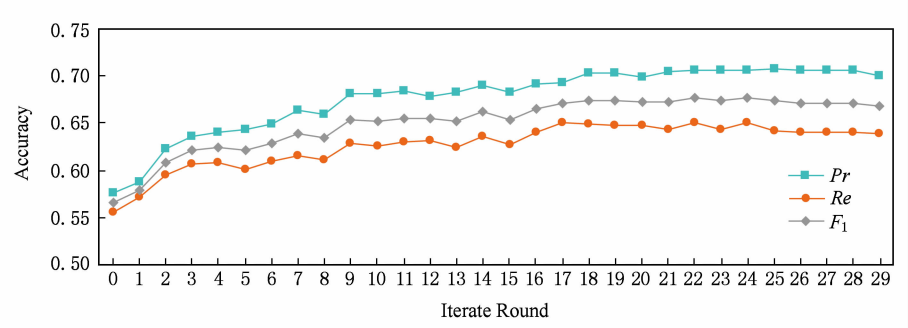


Fig. 3 Performance with number of selected feature

图 3 模型效果随特征数量的变化

#### 5.3.3 对协同训练的评估

本文采用协同训练对未标记的数据进行标记并用于迭代训练. 利用有标记的数据训练出 2 个基本的随机森林分类器,再用 2 个分类器对无标记数据进行预测,选取预测置信度较高的样本进入有标记样本集合,然后基于新的训练数据集合训练出 2 个随机森林分类器,并对 2 个分类器的综合分类结果进行评估,再不断迭代,直到分类效果趋于稳定. 据此得到了图 4 所示.

其中 $x$ 轴为迭代的轮数,$y$ 轴为分类器的评价指标值. 可以看出,随着逐渐加入分类器标注的样本数据,分类器的效果先是逐渐提高,继而曲线趋于平稳,说明分类器效果达到稳定,此后再加入样本不再能使分类器效果有显著提升.

从实验中可以看出随着协同训练的不断迭代,训练模型的效果得到了很好的优化. 因此在缺少大量有标记数据的情况下,使用半监督训练可以有效利用无标记数据,从而改善原本分类器的效果.

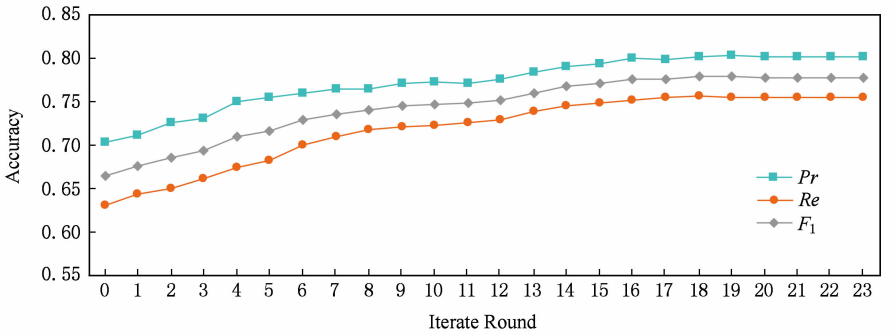


Fig. 4　Performance with number of iteration round

图 4　模型评估指标随协同训练迭代轮数变化

## 6　总　　结

本文分析了心理压力对人的身心的重要性,并针对以往的评估人的心理压力的方法,提出了基于手机感知数据的自动心理压力感知方法,在对用户低侵扰的情况下实现对用户的心理压力的评估.针对手机获取到的日志数据,本文提出了一系列的特征抽取方法,将原始的日志数据转化为可用于分类的样本数据.基于基尼不纯度提出了特征筛选方法,在多项式时间内筛选出对分类有利的特征.然后提出了基于随机森林的协同训练模型,实现了通过手机感知数据对用户的心理压力进行感知的任务($Pr$=80.4%,$Re$=75.5%,$F_1$=77.0%),效果好于基线方法.

## 参　考　文　献

[1] America College Health Association. Fall 2015 reference group executive summary [EB/OL]. 2015 [2017-01-23]. https://www.acha.org/documents/ncha/NCHA-II_FALL_2017_REFERENCE_GROUP_EXECUTIVE_SUMMARY.pdf

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

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

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

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

[6] Hetz C, Martinon F, Rodriguez D, et al. The unfolded protein response: Integrating stress signals through the stress sensor IRE1α [J]. Physiological Reviews, 2011, 91(4): 1219-1243

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

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

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

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

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

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

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

[14] Chen Longbiao, Li Shijian, Pan Gang. Smartphone: Pervasive sensing and applications [J]. Chinese Journal of Computers, 2015, 38(2): 423-438 (in Chinese)
(陈龙彪,李石坚,潘纲. 智能手机:普适感知与应用[J]. 计算机学报, 2015, 38(2): 423-438)

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

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

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

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

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

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

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

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

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

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

**Wang Feng**, born in 1995. Master candidate at the School of Electronic Engineering and Computer Science, Peking University, China. His main research interest is mobile crowd sensing.

**Wang Yasha**, born in 1975. Professor and PhD supervisor at Peking University. His main research interests include urban data analytics, ubiquitous computing, software reuse, and online software development environment.

**Wang Jiangtao**, born in 1987. PhD. Assistant professor at Peking University. His main research interests include collaborative sensing, mobile computing, and ubiquitous computing.

**Xiong Haoyi**, born in 1987. PhD. Assistant professor at the Department of Computer Science, Missouri University of Science and Technology, Rolla, MO, USA. His main research interests include ubiquitous data science, crowdsourcing, and applied optimization & statistics.

**Zhao Junfeng**, born in 1974. PhD. Associate professor at Peking University. Her research interests include software engineering, sofeware reuse, medical data analysis.

**Zhang Daqing**, born in 1965. PhD. Professor at Peking University, China, and Telecom SudParis, France. His main research interests include context-aware computing, urban computing, mobile computing.