计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211100014

大数据驱动的社会经济地位分析研究综述

么晓明[1,2]　丁世昌[3]　赵　涛[4]　黄　宏[5]　罗家德[6]　傅晓明[1]
1 哥廷根大学计算机系　哥廷根 37077 德国
2 中国电信集团云计算分公司大数据事业部　北京 100033
3 信息工程大学网络空间安全学院　郑州 276800
4 国防科技大学前沿交叉学科学院　长沙 410073
5 华中科技大学计算机学院　武汉 430074
6 清华大学社会学系　北京 100084
(yaoxm@chinatelecom.cn)

摘　要　一个人的社会经济地位(Socioeconomic Status,SES)是结合经济学和社会学等因素相对于其他人的经济和社会地位的总体衡量,包含其职业、学历、收入等多维度信息。对这些信息进行综合评估可以帮助政府和相关机构制定各种政策、决策(如政府制定社会政策、企业进行广告个性化服务等),因此该研究得到了研究人员的广泛关注。随着近几年大数据技术和机器学习的发展,以数据驱动的方法来评估社会经济地位时,可以通过融合多维数据和利用各种算法来自动评估人们的社会经济地位,解决传统方法数据采集困难、成本过高的问题。文中旨在概述近年来将大数据技术应用于社会经济地位分析的相关研究进展。首先介绍社会经济地位的基本概念,并讨论大数据方法与传统方法所带来的不同挑战;然后,根据学习过程中的信息,系统性地总结各种相关方法,并详细讨论各类方法的利弊;最后,讨论目前个人社会经济地位分析存在的挑战和问题,并展望未来的相关研究方向。

关键词:社会经济地位;机器学习;深度学习;数据挖掘;社交媒体
中图法分类号　TP391

Big Data-driven Based Socioeconomic Status Analysis:A Survey

YAO Xiao-ming[1,2],DING Shi-chang[3],ZHAO Tao[4],HUANG Hong[5],LUO Jar-der[6] and FU Xiao-ming[1]
1 Institute of Computer Science,University of Goettingen,Goettingen 37077,Germany
2 Cloud Branch Big Data Department,China Telecom Co. Ltd,Beijing 100033,China
3 School of Cyberspace Security,State Key Laboratory of Mathematical Engineering & Advanced Computing,Zhengzhou 276800,China
4 College of Advanced Interdisciplinary Studies,National University of Defense Technology,Changsha 410073,China
5 College of Computer Science and Technology,Huazhong University of Science & Technology,Wuhan 430074,China
6 Department of Sociology,Tsinghua University,Beijing 100084,China

Abstract　Socioeconomic Status (SES),an overall measure of a person's economic and social status relative to others combining factors such as economics and sociology,has received a lot of attention from researchers,as its assessment can help relevant organizations to make various policies and decisions (governmental formulation of social policies,advertising personalized services,etc). In addition,with the development of big data technology and machine learning in recent years,assessing people's socioeconomic attributes (SEAs) and further obtaining the corresponding socioeconomic status with a data-driven approach can address the issue of extremely high cost of traditional methods. Therefore,this paper summarizes the research progresses of applying big data techniques to socioeconomic status analysis in recent years. It first introduces the basic concept of socioeconomic status and discusses the challenges posed by big data methods compared to traditional methods. After that,it systematically summarizes and classifies the state-of-the-art related methods based on the information in the learning process,and present them in detail,discusses the pros and cons of each type of method. Finally,it discusses the challenges and problems of inferring people's socioeconomic status and provides an outlook on future research directions.

Keywords　Socioeconomic status,Machine learning,Deep learning,Data mining,Social media

到稿日期:2021-10-29　返修日期:2022-02-16
基金项目:欧盟水平线 2020 COSAFE 项目(824019);国家重点研发计划(2020YFE0200500)
This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Grant Agreement(824019) and Chinese National Key R & D Program (2020YFE0200500).
通信作者:傅晓明(fu@cs.uni-goettingen.de)

推断人们的社会经济属性（Socioeconomic Attribute，SEAs），如收入水平、教育水平和职业类型，是社会计算的一个重要问题[1]。这些属性在社会分层和社会福利等研究中发挥了至关重要的作用，它们也是计算人们的社会经济地位（Socioeconomic Status，SES）的基本因素。SES 是社会学的一个关键概念，它是结合经济学和社会学，关于某个人的工作经历和个体或家庭基于收入、教育和职业等因素相对于其他人的经济和社会地位的总体衡量，可以帮助政府设计和评估社会政策，特别是福利政策。近年来，SES 也成为在线服务提供商在推荐和广告中提供个性化服务的关键考量要素[2-5]。传统的调查方法普遍是在一个地区人工进行大量的个人或家庭访谈，虽然可以得到准确的信息，但其代价非常昂贵而且十分耗时，不能适用于大规模的调查，这在一定程度上阻碍了该领域的发展。

幸运的是，随着近年来大数据的发展，大量在线用户生成的数据提供了一个很好的条件，使得相关研究人员可以低成本、有效地评估社会经济地位。近年来，研究人员们提出了各种机器学习方法，以便从人们的网络空间行为中自动估计人类的多种社会经济属性[6-12]。例如，文献[10-12]探讨了如何根据推特内容中的语言模式、主题，甚至是情绪来估计人们的收入或职业。文献[6-9]则专注于从人们的手机使用习惯来预测他们的社会地位或家庭收入。最近，研究人员开始尝试从人们的行为表现中推断 SEA。例如，文献[13-14]根据人们在线下零售商购买商品的方式来估计人们的收入和教育水平；文献[15]根据人们在地铁系统中的移动模式来推断人们的收入水平。此外，文献[16]从家庭所在地来预测一个人的个人收入水平、家庭收入水平、职业类型和教育水平。

然而，虽然基于大数据的社会经济属性推断在不同领域的应用已经有了不少的研究，但仍未对该领域方法进行系统、全面的总结，也未深入分析这些方法的优缺点，详细讨论它们的适用性。同时，该领域数据集都是分别采集的，未进行梳理及分析。为了弥补这些空缺，本文将全面回顾社会经济属性推断方面的现有工作，包括代表性方法和技术以及可公开使用的数据集。本文的贡献主要为：

（1）讨论了基于机器学习的社会经济属性推断与传统方法相比所带来的机会和挑战，介绍了现有基于大数据的社会经济属性推断方法，并且将这些方法根据它们在学习过程中使用的信息进行分类。

（2）对各具代表性的基于大数据的社会经济属性推断方法和技术进行了详细介绍，并进一步分析了其优缺点。

（3）介绍了该领域广泛使用的数据集，以促进未来在这一领域的研究和应用。

（4）探讨了基于大数据的社会经济属性推断方法存在的其他问题和挑战，并对未来的研究方向进行了展望。

本文第 1 节介绍了几个相关概念的定义及其意义；第 2 节根据大数据分析中使用的数据源对社会经济地位推断方法进行详细的分类和介绍，并在此基础上分析它们的优缺点，讨论了它们的适用性；第 3 节进一步对现有的社会经济属性推断方法进行了系统的总结，介绍了基于不同任务的分析方法；第 4 节介绍了国内的研究现状；第 5 节总结了该领域广泛使用的数据集并对其进行了相应的分类；第 6 节介绍了这个领域的挑战，并预测该领域的未来研究方向；最后总结全文。

1 相关定义

本节综合以往文献，首先统一介绍社会经济地位和社会经济属性，然后讨论两者的意义以及研究方式。

1.1 社会经济地位的定义

社会经济地位指一个人在社会系统中的个人地位，包括社会地位（声望、权力和经济福利）以及获得金融、社会文化和人力资本资源的机会。早在 20 世纪 40 年代，社会经济地位就引起了人们的重视。在早期的研究中，学术界还在探索社会经济地位的构成指标。以基础医学为例，多采用可能构成 SES 的单个指标在相关影响因素之间进行探索研究，如收入与健康的关系[17-18]、教育程度和健康的关系[19]、职业类型[20-21]和健康的关系。随后，Warner 等[22]进行了住房与社会经济地位间关系的探索。20 世纪 50 年代，各研究领域逐渐达成了共识：社会经济地位指标由职业、教育水平和收入构成。

SES 会影响许多关系的质量和满意度，如成人恋爱或亲子关系[23]，而且可以进一步影响一个人可获得的群体成员的数量[24]。此外，对个体和相邻区域 SES 的评估还有助于政府和社会志愿机构开展精准社会福利、城乡基础设施建设及区域发展规划等。因此，SES 在心理学和社会学工作中都有很大的意义。除了具有科学意义外，SES 将有助于对教育、收入和职业方面相似的人（状况相似的人）进行分组，对定向广告或检测全球商业趋势等应用也具有潜在价值。

1.2 社会经济属性的定义

社会经济属性是与 SES 密切相关的一系列属性（包括但不限于收入、教育水平等），能够侧面反映出个人的社会经济地位。因此，研究 SES 往往可以从研究 SEA 入手。本文基于现有的研究成果，提出关于社会经济地位的研究可以从社会经济地位推断与社会经济属性推断这两方面入手，如图 1 所示。

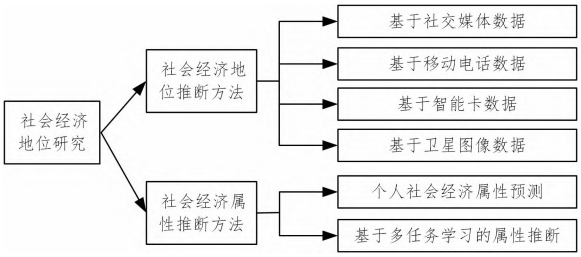

图 1 SES 大数据分析研究方法示意图

Fig. 1 Schematic diagram of SES big data analysis research methodology

2 社会经济地位推断

SES 是社会科学领域中一个被广泛研究的概念，尤其是在健康和教育分析方面[25]。近年来，公司和研究人员越来越关注 SES 的推断，因为它在许多高价值的应用中具有巨大潜力，如个性化推荐和网上银行贷款决定等。虽然目前在推断人口属性如年龄、种族和性别方面有很大的进步[2,26]，但 SES 的推断仍面临着众多挑战。其中一个主要的障碍是 SES 的关键真实数据（往往覆盖大量人群）比年龄和性别等属性更难

获得。相较而言,用户不愿意透露他们的教育、职业和收入信息,出于隐私考虑,拥有这些数据的组织也很少向公众开放数据。因此,近年来研究人员开始使用一些基于大数据的间接SES指标,这些数据源通常涵盖数百万人,记录了他们生活方式的不同方面。

2.1 基于社交媒体数据的分析方法

众所周知,社交媒体一直是重要的、由网络空间用户生成的数据源,研究人员对其给予了很大的关注。Preotiuc-Pietro等首次提出了根据社交媒体上用户生成的数据来推断个人层面的职业等级的大规模系统研究,这与SES非常相似[10]。在这项工作中,他们主要关注用户在社交媒体上的语言使用。他们收集了5 191名在用户描述栏中提到自己职业的英文用户,而且这些用户都至少有200条以上的推文。然后,他们根据用户聚合的推文集,通过奇异值分解(Singular Value Decomposition,SVD)单词嵌入、归一化点互信息(Normalized Pointwise Mutual Information,NPMI)集群、神经嵌入和神经集群来设计用户级文本特征。最后,他们使用非线性高斯过程(Gaussian Process,GP)框架来估计用户的职业类别。实验结果表明,用户的职业会影响其语言表达模式。

除此之外,Lampos等提出了第一批推断社交媒体用户个人层面社会经济地位的方法之一[12]。他们从Twitter上收集了1 342名英语用户的资料,这些用户是根据他们是否在资料中报告职业类型来选择的。然后,研究人员收集这些用户在2014年2月至2015年3月期间的推文,研究人员根据职业类型计算出用户的SES。与文献[3]相比,他们增加了其他非文本特征,如推文总数和关注账户数等。这些特征表征了用户基于平台的行为和他们在平台上的重要性。最后,研究人员还利用GP根据用户层面的社交媒体特征来预测人们的SES。

之后,Ding等分析了SES和从Twitter中提取的人们的活动模式之间的关系[15]。研究人员收集了7 660名居住在华盛顿特区的用户,他们有超过40条地理标签的推文。然后,根据这些地理标记的推文的地理信息和时间信息,可以推断出这些用户的家庭和工作区域。然后,研究人员分析用户的活动模式,主要包括活动区的数量、家庭和活动区之间的距离、标准偏差椭圆等。他们从这些活动模式中发现,虽然SES非常重要,但城市空间结构在影响不同社区用户的活动模式方面也起到了关键作用。

Abitbol等[27]提出了一种推断Twitter用户SES的方法,该方法结合了许多来源的信息,包括Twitter、人口普查数据、LinkedIn和谷歌地图。首先,他们收集了超过9 000万条推文,由130万法国用户在一年内发布。然后,他们根据地理标记的推文找到用户的家庭位置。通过这种方式,他们将用户映射到人口普查区,每个人口普查区的收入中位数由法国国家统计和经济研究所(INSEE)公布。人口普查区的收入中位数被用作居住在其中的Twitter用户的收入水平的近似值。用户在推特或资料中提供他们的LinkedIn账户,就可以找到其职业数据。研究人员还通过谷歌地球的用户街景来估计用户生活区的社会经济特征,他们邀请专家通过观察街景来注释用户生活区的水平。一个用户的SES水平是人口普查收入数据、职业数据和住房价格数据的组合,这些特征与文献[4-5]中的特征相似,包括用户资料和从推文中提取的文本特征。最后,研究人员使用3种经典的机器学习方法(AdaBoost,Random Forest和XGBoost)来预测用户的SES水平。

2.2 基于移动电话数据的分析方法

另一个重要的用户生成的数据类型是移动电话数据,然而现有的研究大多只关注群体层面的SES推断。Soto等探讨了如何利用从手机记录的汇总使用中获得的信息来确定人口的社会经济水平[6]。具体来说,他们的工作可以得到一个社会经济水平到每个基地收发站(BTS)塔的覆盖区域。在城市中,一个基站可以覆盖大约1 km^2的区域,研究人员只需研究经常打电话的用户,不经常打电话的用户的信息不需要进行分析。他们设计了各种用户呼叫行为的特征来区分每个基站塔,这些特征包括一个基站区域的综合呼叫行为,如呼叫或短信息的总数。一个基站区域的SES是根据政府公布的家庭收入、职业来计算的。最后,Soto等使用标准的经典机器学习方法,如支持向量机(Support Vector Machine,SVM)和随机森林,来预测每个基站区域的SES。虽然这种方法是最早从手机数据中预测(群体层面)SES的方法之一,但它不能估计每个人的个人层面SES。

基于相同的数据集,Soto等探讨了手机使用的各种特征(包括手机消费、社会信息和移动模式)和社会经济指标(包括收入和教育)之间的关系[6]。他们发现,一个人的SES与他/她的平均通话物理距离、手机相关费用、通信的交换频率和经常旅行的地理位置有适度或强烈的关联。

Blumenstock等提出了一种基于手机数据估计卢旺达人细化的群体层面SES(即家庭层面)的方法[7]。研究人员首先根据卢旺达人是否有照明用电、冰箱、电视及其他高级物品,为他们设计了一个综合财富指数。这些数据是通过电话调查收集的,然后他们从移动电话数据中提取特征,最后使用一个标准的经典机器学习方法,从这些特征中估计人们的财富指数。实验结果表明,从移动电话数据中估计的财富分布与卢旺达政府测量的实际财富分布有很强的相关性。这项工作考虑了电话使用的多种因素,包括通信、结构和联系网络。移动模式作为一个支持性特征被讨论。

Zhao等[28]将基于半监督超图的因子图模型(Hypergraph Based Factor Graph Model,HyperFGM)用于个体SES预测,HyperFGM能够有效地捕获SES与单个手机记录之间的关联,以处理单个记录的稀疏性。对于稀缺的显式关系,HyperFGM在超图结构上对用户之间的隐式高阶关系进行建模。此外,HyperFGM以半监督的方式探索有限的标记数据和未标记数据。

Almaatouq等[8]构建了一个简单的模型,仅从人们的移动通信模式就可估计一个地区的失业率。他们还发现,汇总的呼叫活动、通信网络与失业率密切相关。

Xu等基于新加坡和波士顿两个城市的移动电话数据集,分析了多种流动性特征和SES之间的关系[9]。在新加坡,他们把生活区的住房价格作为SES;在波士顿,他们使用人口普查数据作为SES。他们发现,流动性和SES之间的关系可能在不同的城市之间有所不同,而且这种关系相当复杂,它可能受到几个不同因素的影响,如住房的空间安排、就业机会和人类活动。例如,一般比较富有的电话用户群体往往在新加坡

的旅行时间较短,但在波士顿则较长。

2.3 基于智能交通卡数据的分析方法

近年来,自动收费(AFC)系统在世界各地得到了越来越广泛的应用[29]。部署 AFC 系统的最初目的是在没有人工干扰的情况下使收费更快、更便宜。然而,研究人员意识到,每天记录的大量和连续的智能交通卡数据(Smart Card Data,SCD)可以使许多领域受益,例如智能交通卡数据可以被用来了解公共交通的需求模式,这些知识对规划新的公共交通系统有很大帮助[29]。智能交通卡数据也可以被用来调查乘客的旅行模式[30]。然而,有关 SES 和智能交通卡数据之间关系的工作相当有限。

Ding 等[15]根据上海数百万用户的 SCD 来记录他们的时间和空间流动行为。在具体实现中,他们提出了 S2S,即一种基于深度学习根据人们的 SCD 来估计其 SES 的方法。从本质上讲,S2S 建立了两类与 SES 相关的特征,即时间序列特征和一般统计特征,并利用深度学习进行 SES 估计。最终,实验结果表明,深度学习方法的效果优于传统的特征建模方法。

Zhang 等[31]根据伦敦 33 026 名公共交通用户的智能交通卡数据,来调查他们的多周活动模式。研究人员首先将每位乘客表现为几周内的有序活动序列,他们可以从这个序列中捕捉到与旅行者的时间模式有关的信息。然后,研究人员根据每个用户的长期活动序列,使用 K-Means 算法对用户进行聚类。通过这种方式,他们找到了伦敦公共交通旅行者的 11 个聚类。每个聚类的长期流动性特征完全不同。例如,与其他群组不同,前 4 个群组的用户在工作日期间更可能在主要和次要地点之间移动。然后,研究人员调查了少部分用户(1 973 人)的人口统计学属性,然后分析了每个集群的人口统计学属性。他们发现,一些集群的平均收入高于其他集群。这项工作表明,收入可能与人们的智能卡流动数据有关。

Antipov 等介绍了一种方法,根据居住在雷恩(法国)的乘客的时间习惯对他们进行聚类[32]。他们研究了票价类型在不同集群中的比例分布[30]。雷恩的 SCD 数据集包括年轻用户、普通用户、老年用户等票价类型。他们发现,不同票价类型之间存在一些流动性差异。例如,主要由学生组成的群组倾向于在周三提前回家,因为法国周三的课程时间提前结束,而其他群组则没有这种模式,这也表明 SCD 记录可能与用户的年龄和职业有关。这些工作表明,基于 SCD 的流动性和 SES 之间可能存在一些关系。

2.4 基于卫星图像数据的分析方法

城市化和社会不平等是当今时代的两个主要政策主题,在富人和穷人并存的大城市中交织在一起,减少不平等现象是全球可持续发展议程的重点,也是许多城市的政策目标。然而,目前为这些政策提供信息和衡量其实际影响的数据集来自不连贯和低效的监测系统。例如,以高维空间和时间分辨率测量社会经济地位至关重要,但即使在世界发达地区这也是一个重大挑战。新兴的大规模数据来源,如遥感、图像和 GPS 轨迹,有可能大大推进测量城市特征和人口特征的迁移速度、流动频率以及受影响的地域范围。

来自领域科学和计算机科学的研究人员对解决与应用先进的学习技术相关的问题越来越感兴趣,这些技术侧重于自动特征提取的测量和数据收集任务。机器学习与图像的相关应用包括:来自卫星数据的贫困检测[33-35]、收获大小和作物产量[36-37],以及来自谷歌街景(GSV)图像的收入[38]、感知安全[39-40]、绿色度和开放度[41-42]。

3 社会经济属性推断

3.1 个人社会经济属性预测

在一些发展中国家,个人 SEA 推断是收集经济或社会统计数据的一种替代方法[42]。估计的个人 SEA 也可用于改善个人推荐和精确营销。鉴于其重要性,研究人员提出了大量的方法来估计收入水平、职业和教育。根据本文的统计,大多数方法都试图从人们的网络空间行为数据中预测 SEA,如移动电话[10]和 Twitter 内容[11]。以个人收入预测为例,研究得最多的两个数据源类型来自在线社交网络(Online Social Networking,OSN)和移动电话(主要包括通话记录和使用数据)。如表 1 所列,相当多的研究都集中在基于 OSN 的个人收入预测上。请注意,本文也汇集了部分预测个人社会经济地位的论文,因为 SES 可以被看作是 SEA 的一个特殊版本。

表 1 个人社会经济属性预测相关工作

Table 1 Work related to prediction of individual socioeconomic attributes

相关工作	数据源	预测属性
文献[7]	tweets	SES
文献[34]	tweets	收入
文献[4]	tweets	收入
文献[44]	tweets	教育、收入
文献[1]	tweets	职业、收入
文献[45]	tweets	收入
文献[46]	tweets	教育、收入
文献[47]	tweets	教育、收入
文献[49]	tweets	家庭收入
文献[50]	Facebook Likes	收入
文献[42]	mobile phone metadata	个人收入
文献[8]	mobile phone records	SES
文献[51]	mobile phone call detail records	收入
文献[10]	mobile phone metadata	收入
文献[52]	mobile phone metadata	收入
文献[53]	cookie	收入、教育水平
文献[14]	retail transaction records	收入、教育水平
文献[13]	retail transaction records	收入、教育水平
文献[6]	smart card transportation records	SES
文献[54]	WiFi log	教育、收入

(1)基于社交媒体数据的 SEA 推断。近年来,著名的 OSN 平台(如 Twitter 和 Facebook)发展迅速。许多重要的工作表明,人们的 SEA 可以通过分析他们的推特、社交链接或 OSN 记录的资料来进行预测[3-4,7,11,43-47]。Preotiuc-Pietro 等进行了该领域的第一个大规模研究,从人们产生的社交媒体数据中预测个人层面的收入[11]。他们收集了居住在英国的 5 191 名 Twitter 用户,涵盖 55 种职业类型,每种职业的平均年收入可以在英国政府发布的《年度工时和收入调查》[48]中找到。然后,研究人员根据用户档案数据和推文内容设计了一系列特征,如感知的心理人口学、情绪和情感等。最后,研究人员应用高斯过程来预测用户收入,预测的收入与实际的用户收入达到了 0.633 的相关度,表明推文可以用来预测收入。他们还分析了不同特征与用户收入的关系。他们发现,与恐惧或快乐有关的词的百分比、转发的比例,以及推文的主题是最重要的特征。例如,高收入的推特用户可能会表达更多的恐惧和愤怒,而低收入的用户则表达更多带有情绪的观点。

Volkova 等在一系列的工作中研究如何预测 Twitter 用户的收入和教育水平[44,47]。在文献[47]中,研究人员要求 Amazon Mechanical Turk 上的工人手动检查 5000 个 Twitter 用户的在线内容和资料(这些 Twitter 用户至少发布了 200 条推文)。工人需要猜测:1)用户的年收入是否超过 35000 美元;2)用户是否有大学学位。然后,研究人员从用户的推文中提取文本特征。最后,利用一个对数线性模型来预测这些用户的收入和教育水平。Volkova 等[44]在一个更大的数据集上改进了文献[47]中的方法。他们收集了来自美国和加拿大的 123513 名用户的推文。他们使用文献[47]中训练的模型来预测用户的收入和教育水平,预测的收入和教育水平被作为估计的标签来利用。然后,他们通过提取描述用户和他们的邻居用户之间情感对比的特征发现,收入和教育都可以根据该用户表达的情绪和用户的社会环境来预测。

Filho 等基于 Foursquare 和 Twitter 的用户交互信息,来估算收入和财富拥有情况,从而通过机器学习算法自动产生在线社交网用户社会地位数据,并用于预测其准确率[49]。他们的分析结果发现,如果将社会地位分成 2 个、3 个或者 4 个类别,则可以达到介于 57%~73%之间不一样的预测准确率。

最近,Matz 等提出了一种预测 Facebook 用户收入水平的方法[50]。研究人员开展了一项付费的在线调查,收集美国 Facebook 用户的收入信息。研究人员选择了 2623 名参与者,他们更新的推文拥有超过 10 个赞或 500 个字。两种数据被用于特征提取:Facebook 上的用户赞和状态更新的内容。一种广泛使用的降维方法,即奇异值分解被应用于初始特征。最后,研究人员利用一种常用的机器学习算法——岭回归模型,来预测 Facebook 用户的记录收入。

(2)基于移动电话数据的 SEA 推断。另一个重要的用户生成的数据类型是手机数据。许多现有的工作试图根据多种与手机相关的数据来预测人们的收入水平,如通信、联系网络的结构、用户的移动模式等。

文献[8]显示,手机通话行为、社会网络和流动性数据可以被用来识别生活在社区的人口的财富水平。地面真实数据是由国家统计局提供的,它考虑了 134 个指标,包括手机数量、电脑、综合收入、家庭成员的职业等的研究水平。

文献[51]通过对某电信公司电话及短信通信记录的挖掘,建立用户通话的社交网络图,并以部分用户的已知银行收入信息作为 SES 的代理变量,建立贝叶斯模型来预测未知收入用户的 SES 级别。

文献[52]提出了一种根据各种手机相关数据来区分一个人的家庭是否贫穷的方法。他们首先在一个低人类发展指数的国家进行了一次大规模的全国性调查,调查结束后,得到了 8 万多人的收入数据及其 3 个月的原始手机数据。然后,他们设计了 150 个特征,涵盖基本的手机使用数据(如通话时长)、充值交易(如每次交易的充值金额)、社交网络、手机类型(如手机品牌)、收入(如互联网的收费)和高级手机使用(如互联网容量)。最后,研究人员使用标准的多层前馈方法来预测人们的收入水平。

Blumenstock 等[42]通过从移动电话通信提取的特征和移动模式来估计卢旺达人和阿富汗人的家庭收入。研究人员发现,基于在一个国家收集的数据的模型不能直接用于另一个国家。

(3)除了手机和社交媒体数据,研究人员也开始关注基于其他类型的用户生成的数据(如零售交易记录)来预测 SEA[13-14]。例如,Wang 等[13]提出了首个基于零售场景预测用户收入和教育水平的方法。他们从中国的一家大型零售商那里收集了一个数据集,该数据集包含 120 万用户和 22 万种商品之间的 4900 多万次交易。用户是根据他们的购买历史来表示的。最后,研究人员将所有用户的表示方法输入一个对数模型中,以同时预测用户的收入和教育水平。

3.2 多任务学习的多 SEA 推断

多任务学习(Multi-task Learning,MTL)是机器学习中的一种学习范式。MTL 的主要目的是利用多个任务中共享的有用信息来提高所有任务的概括性能[55]。所有这些学习任务都被认为是相互关联的,考虑到数据收集的成本,研究人员可能需要从一个数据集中预测多个用户的属性。因此,人们在研究如何将多任务学习应用于用户属性推断方面做出了一些努力[13,56]。

最早提出的用于社会经济属性推理的多任务模型之一是结构化神经嵌入(Structured Neural Embedding,SNE)[13]。SNE 使用一个简单的密集层来生成所有输入特征的初始嵌入向量,对这些向量进行平均汇集,然后将其送入线性预测层,用于每个 SEA 估计任务。与传统的多任务方法不同,SNE 忽略了属性之间的关联性,因为 Wang 等[13]认为,如果没有明确的任务之间的关系知识,相关关系很难建模。他们没有对每个任务进行求和,而是使用一个单一的结构化预测任务来结合所有的任务。通过这种方式,他们希望能揭示属性之间的关联模式。然而,SNE 的输出空间比传统的多任务学习方法的输出空间更大,因此,如果输入数据源的规模有限,则不适合使用 SNE,否则会导致过度拟合。

Ding 等[16]从一个人的家庭位置来预测收入水平、家庭收入水平、职业类型和教育水平。研究人员通过调查收集了包含中国 9 个省和 85 个城市的人们的家庭位置和社会经济属性,并用来自房地产网站、政府统计网站、在线地图服务等的信息进一步丰富了家庭位置。为了从输入特征中学习一个共享的表征以及不同 SEA 的特定属性表征,研究者提出了 H2SEA,即一种基于因子化机器的带有注意力机制的多任务学习方法。实验结果表明,家庭位置可以明显提高所有 SEA 预测任务的估计精度,所提出的 H2SEA 模型在各种评价指标方面均优于其他进行 SEA 推断任务的模型。

最近,Kim 等[56]提出了一种新的多任务方法,从人们的交易记录中预测年龄、性别和婚姻状况。研究人员收集了 56000 名用户的购买历史和用户属性,输入的数据与 SNE[13]非常相似。与 SNE 相比,文献[56]将共享嵌入转化为特定任务的嵌入,并通过注意力机制检测更多重要信号。结果表明,关注机制不仅提高了对客户属性的预测性能,而且有助于解释客户属性与不同项目的关系。ETNA 简单地使用项目的初始嵌入作为输入,这对于有限的输入数据源来说也是不够的。

4 国内研究现状

国内这方面的研究起步较晚,直至 2005 年,才有学者[57]

首次提出我国社会经济地位指数的计算公式。Qi 等[58]引入多种健康指标来推测我国社会经济地位各指标的作用，为社会经济地位指标的构建提供了帮助。

国内的研究基本集中在社会科学领域，学者们通过已有的社会经济地位计算公式，来分析社会经济地位与健康、心理健康等内容的关系及影响方式。Zhang 等[59]从大规模的随机问卷调查数据中发现，居民所处的社会阶层地位越高，对心理健康的影响越偏向于积极。Wei 等[60]通过分析发现，收入、教育和职业这 3 种社会经济地位变量与老年人参加文化组织活动的积极性相关度很高，继而影响健康水平；Wang[61]发现，社会经济地位中体育锻炼频率这一变量对健康的影响作用比其他因素更大。

总体来看，国内关于社会计算的研究大多集中在对社会经济地位的应用方面，仅使用简单的 Logistic 回归模型及路径分析法进行计算；但针对社会经济地位本身的分析计算研究还很缺乏，利用大数据的优势进行社会经济地位分析的工作还没有受到关注，在这方面还有很大的探索空间。希望通过这篇综述，在未来可以引起更多国内学者的研究兴趣。

5 数据集介绍

众所周知，高质量的数据集对学术研究至关重要，因此，本节总结了用于社会经济地位分析的常用数据集。由于该领域现阶段使用的数据集都是自行采样的，因此本文只介绍整体的数据集情况。本文将数据集分为 4 类，即社交媒体数据、移动电话数据、智能卡数据以及卫星图像数据，涉及的具体数据集的统计信息汇总如表 2 所列。

表 2 公开数据集

Table 2 Public datasets

数据集类型	数据集	数据集描述	源数据出处	相关论文
社交媒体数据	新浪微博 Twitter Facebook	大规模社交网络，包含众多社会关系类型，同时还包含部分用户的职业，适用于多种社会相关理论的研究	https://weibo.com/ https://twitter.com/ https://www.facebook.com/	文献[28] 文献[4,7,34,44-47,49] 文献[50]
移动电话数据	ISP 采集数据	ISP 可以提供活跃手机用户的匿名手机上网记录，在用户同意提供与 SES 相关的个人信息（职业、教育、收入）之后，可用于社会经济地位的分析	—	文献[8,10,42,51-52]
智能卡数据	地铁记录（上海）	该数据集包含 2015 年 4 月 1 日至 4 月 16 日上海地铁的所有记录，并经过了处理，保证了用户隐私	—	文献[15]
	POI（上海）	基于高德地图对上海市 POI 数据集进行了抓取，类别包括公共设施、家政、教育、商务住宅、医院、酒店、汽车服务、体育和休闲、风景、餐厅、公共交通和金融服务	https://www.amap.com/	文献[15]
	房价（上海）	房价数据来自 lianjia.com。该网站记录了上海大部分待售公寓的价格和位置信息。该数据集包含了所有社区的平均房价	http://Lianjia.com	文献[15]
卫星图像数据	2012 European Union Urban Atlas project	该数据集提供了欧盟 28 国和欧洲自由贸易区国家约 700 个人口超过 10 万的城市地区的高分辨率土地覆盖地图。它将每个城市分割成详细的多边形，这些多边形被划分为 27 个标准土地使用类别之一	https://land.copernicus.eu/local/urban-atlas/urban-atlas-2012	文献[27]
	Aerial imagery of the complete French metropolitan territory (2013－2016)	该数据集由国家地理信息研究所（IGN）发布，数据集分布为一系列 5 km×5 km 的地理参考图	https://geoservices.ign.fr/documentation/diffusion/telechargement-donnees-libres.html#ortho-hr-sous-licence-ouverte	文献[27]

6 挑战和未来

尽管目前在社会经济地位推断领域已有丰富的研究工作，但是仍然面临着很多挑战。

（1）SES 的真实值目前没有一个统一的标准，因此针对不同分析任务定义相应的 SES 真实值并收集相关数据是一个重要的任务。如基于智能卡数据的分析方法将人们的居住地房价作为真实值，还可以将工作地区的房价水平作为真实值，但这需要更详细的 SES 调查。

（2）SEA 的数据源具有多样性，如视频、图像、文本、音频等，如何有效地融合多模态数据以计算用户的 SES 值是未来的一大关注点。

（3）SEA 数据的获取通常基于采样方法，如何保证采样的无偏性是值得关注的问题，这也能够防止得出具有“幸存者偏差”的实验结果。

结束语 大数据和机器学习技术的发展极大地促进了社会经济地位的分析以及相关应用。本文对应用于推断社会属性的大数据方法进行了全面的回顾，并系统地介绍了相应方法以及得到广泛使用的基准测试程序以及资源。本文旨在提供一份简洁、清晰的大数据方法应用于社会经济属性分析的概述，其不仅可以为对该方面感兴趣的读者提供帮助，而且可以为继续在该领域工作的研究人员和工程技术人员提供参考。

参 考 文 献

[1] ALETRAS N，CHAMBERLAIN B P. Predicting twitter user socioeconomic attributes with network and language information [C]// Proceedings of the 29th ACM on Hypertext and Social Media. 2018：20-24.

[2] SZOPIЙSKI T S. Factors affecting the adoption of online banking in Poland[J]. Journal of Business Research, 2016, 69(11): 4763-4768.
[3] CHEN D, JIN D, GOH T T, et al. Context-awareness based personalized recommendation of anti-hypertension drugs[J]. Journal of Medical Systems, 2016, 40(9): 1-10.
[4] HUNG L. A personalized recommendation system based on product taxonomy for one-to-one marketing online[J]. Expert Systems with Applications, 2005, 29(2): 383-392.
[5] WU Y, CARNT N, STAPLETON F. Contact lens user profile, attitudes and level of compliance to lens care[J]. Contact Lens and Anterior Eye, 2010, 33(4): 183-188.
[6] SOTO V, FRIAS-MARTINEZ V, VIRSEDA J, et al. Prediction of socioeconomic levels using cell phone records[C]// International Conference on User Modeling, Adaptation, and Personalization. Berlin: Springer, 2011: 377-388.
[7] BLUMENSTOCK J, CADAMURO G, ON R. Predicting poverty and wealth from mobile phone metadata[J]. Science, 2015, 350(6264): 1073-1076.
[8] ALMAATOUQ A, PRIETO-CASTRILLO F, PENTLAND A. Mobile communication signatures of unemployment[C]// International Conference on Social Informatics. Cham: Springer, 2016: 407-418.
[9] XU Y, BELYI A, BOJIC I, et al. Human mobility and socioeconomic status: Analysis of Singapore and Boston[J]. Computers, Environment and Urban Systems, 2018, 72: 51-67.
[10] PREOŢIUC-PIETRO D, LAMPOS V, ALETRAS N. An analysis of the user occupational class through Twitter content[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015: 1754-1764.
[11] PREOŢIUC-PIETRO D, VOLKOVA S, LAMPOS V, et al. Studying user income through language, behaviour and affect in social media[J/OL]. PloS One. https://doi.org/10.1371/journal.pone.0138717.
[12] LAMPOS V, ALETRAS N, GEYTI J K, et al. Inferring the socioeconomic status of social media users based on behaviour and language[C]// European Conference on Information Retrieval. Cham: Springer, 2016: 689-695.
[13] WANG P, GUO J, LAN Y, et al. Your cart tells you: Inferring demographic attributes from purchase data[C]// Proceedings of the ninth ACM International Conference on Web Search and Data Mining. 2016: 173-182.
[14] OYAMADA M, NAKADAI S. Relational mixture of experts: Explainable demographics prediction with behavioral data[C]// International Conference on Data Mining (ICDM). IEEE, 2017: 357-366.
[15] DING S, HUANG H, ZHAO T, et al. Estimating socioeconomic status via temporal-spatial mobility analysis—A case study of smart card data[C]// 28th International Conference on Computer Communication and Networks (ICCCN). IEEE, 2019: 1-9.
[16] DING S, GAO X, DONG Y, et al. Estimating Multiple Socioeconomic Attributes via Home Location—A Case Study in China [J]. Journal of Social Computing, 2021, 2(1): 71-88.
[17] CULLUMBINE H. The health of a tropical people. A survey in Ceylon. 2. Environment, health and physique[J]. Lancet, 1953, 264: 1144-1147.
[18] GOVER M. Physical impairments of members of low-income farm families; 11 490 persons in 2,477 rural families examined by the Farm Security Administration, 1940; variation of blood pressure and heart disease with age; and the correlation of blood pressure with height and weight[J]. Public Health Reports, 1944, 59(36): 1163-1184.
[19] AYYAGARI P, GROSSMAN D, SLOAN F. Education and health: evidence on adults with diabetes[J]. International Journal of Health Care Finance and Economics, 2011, 11(1): 35-54.
[20] SHORTELL S M. Occupational prestige differences within the medical and allied health professions[J]. Social Science & Medicine, 1974, 8(1): 1-9.
[21] SMITH A M, BAGHURST K I. Public health implications of dietary differences between social status and occupational category groups[J]. Journal of Epidemiology & Community Health, 1992, 46(4): 409-416.
[22] MEEKER M, EELLS K. Social Class in America[J]. Journal of Consulting Psychology, 1949, 13(6): 451-452.
[23] CONGER R D, CONGER K J, MARTIN M J. Socioeconomic status, family processes, and individual development[J]. Journal of Marriage and Family, 2010, 72(3): 685-704.
[24] JETTEN J, HASLAM S A, BARLOW F K. Bringing back the system: One reason why conservatives are happier than liberals is that higher socioeconomic status gives them access to more group memberships[J]. Social Psychological and Personality Science, 2013, 4(1): 6-13.
[25] BRADLEY R H, CORWYN R F. Socioeconomic status and child development[J]. Annual Review of Psychology, 2002, 53(1): 371-399.
[26] SIRIN S R. Socioeconomic status and academic achievement: A meta-analytic review of research[J]. Review of Educational Research, 2005, 75(3): 417-453.
[27] ABITBOL J L, KARSAI M. Socioeconomic correlations of urban patterns inferred from aerial images: interpreting activation maps of Convolutional Neural Networks[J]. arXiv: 2004.04907, 2020.
[28] ZHAO T, HUANG H, YAO X, et al. Predicting individual socioeconomic status from mobile phone data: a semi-supervised hypergraph-based factor graph approach[J]. International Journal of Data Science and Analytics, 2019, 9(1): 1-12.
[29] BAGCHI M, WHITE P R. The potential of public transport smart card data[J]. Transport Policy, 2005, 12(5): 464-474.
[30] MOHAMED K, CÔME E, OUKHELLOU L, et al. Clustering smart card data for urban mobility analysis[J]. IEEE Transactions on intelligent transportation systems, 2016, 18(3): 712-728.
[31] ZHONG Y, YUAN N J, ZHONG W, et al. You are where you go: Inferring demographic attributes from location check-ins [C]// Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. 2015: 295-304.
[32] ANTIPOV G, BERRANI S A, DUGELAY J L. Minimalistic CNN-based ensemble model for gender prediction from face images[J]. Pattern Recognition Letters, 2016, 70: 59-65.
[33] STEELE J E, SUNDSØY P R, PEZZULO C, et al. Mapping poverty using mobile phone and satellite data[J/OL]. Journal of

The Royal Society Interface, 2017, 14(127). https://doi.org/10.1098/rsif.2016.0690

[34] XIE M, JEAN N, BURKE M, et al. Transfer learning from deep features for remote sensing and poverty mapping[C]// Thirtieth AAAI Conference on Artificial Intelligence. 2016.

[35] LOBELL D B. The use of satellite data for crop yield gap analysis[J]. Field Crops Research, 2013, 143: 56-64.

[36] YOU J, LI X, LOW M, et al. Deep gaussian process for crop yield prediction based on remote sensing data[C]// Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[37] GEBRU T, KRAUSE J, WANG Y, et al. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States[J]. Proceedings of the National Academy of Sciences, 2017, 114(50): 13108-13113.

[38] NAIK N, PHILIPOOM J, RASKAR R, et al. Streetscore-predicting the perceived safety of one million streetscapes[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2014: 779-785.

[39] NAIK N, KOMINERS S D, RASKAR R, et al. Computer vision uncovers predictors of physical urban change[J]. Proceedings of the National Academy of Sciences, 2017, 114(29): 7571-7576.

[40] SEIFERLING I, NAIK N, RATTI C, et al. Green streets—Quantifying and mapping urban trees with street-level imagery and computer vision[J]. Landscape and Urban Planning, 2017, 165: 93-101.

[41] RICHARDS D R, EDWARDS P J. Quantifying street tree regulating ecosystem services using Google Street View[J]. Ecological Indicators, 2017, 77: 31-40.

[42] BLUMENSTOCK J E. Estimating economic characteristics with phone data[C]// AEA Papers and Proceedings. 2018: 72-76.

[43] VOLKOVA S. Predicting demographics and Affect in social networks[D/OL]. John Hopkins University. https://jscholarship.library.jhu.edu/handle/1774.2/39639?show=full.

[44] VOLKOVA S, BACHRACH Y. Inferring perceived demographics from user emotional tone and user-environment emotional contrast[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 1567-1578.

[45] HASANUZZAMAN M, KAMILA S, KAUR M, et al. Temporal orientation of tweets for predicting income of users[C]// Association for Computational Linguistics (ACL). 2017.

[46] VOLKOVA S, BACHRACH Y. On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure[J]. Cyberpsychology, Behavior, and Social Networking, 2015, 18(12): 726-736.

[47] VOLKOVA S, BACHRACH Y, ARMSTRONG M, et al. Inferring latent user properties from texts published in social media [C]// Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.

[48] Annual survey of hours and earnings[OL]. http://www.ons.gov.uk/ons/rel/ashe/annual-survey-of-hoursandearnings/.

[49] FILHO R M, BORGES G R, ALMEIDA J M, et al. Inferring user social class in online social networks[C]// Proceedings of the 8th Workshop on Social Network Mining and Analysis. 2014: 1-5.

[50] MATZ S C, MENGES J I, STILLWELL D J, et al. Predicting individual-level income from Facebook profiles [J/OL]. PLoS One. https://doi.org/10.1371/journal.pone.0214369.

[51] FIXMAN M, BERENSTEIN A, BREA J, et al. A Bayesian approach to income inference in a communication network[C]// 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2016: 579-582.

[52] SUNDSØY P, BJELLAND J, REME B A, et al. Deep learning applied to mobile phone data for individual income classification [C]// Proceedings of the 2016 International Conference on Artificial Intelligence: Technologies and Applications. Bangkok, Thailand. 2016: 24-25.

[53] ATAHAN P. Learning profiles from user interactions and personalizing recommendations based on learnt profiles[M]. The University of Texas at Dallas, 2009.

[54] REN Y, TOMKO M, SALIM F D, et al. Understanding the predictability of user demographics from cyber-physical-social behaviours in indoor retail spaces[J]. EPJ Data Science, 2018, 7: 1-21.

[55] ZHANG Y, YANG Q. A survey on multi-task learning[J]. arXiv: 1707.08114, 2017.

[56] KIM R, KIM H, LEE J, et al. Predicting multiple demographic attributes with task specific embedding transformation and attention network[C]// Proceedings of the 2019 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2019: 765-773.

[57] LI C L. Prestige Stratification in Contemporary Chinese Society-Occupational Prestige and Socioeconomic Status Index Measurements[J]. Sociological Studies, 2005(2): 74-102.

[58] QI L S, WANG C W. Health status and socioeconomic status: a study based on multiple indicators[J]. Chinese Health Economics, 2010, 29(8): 47-50.

[59] ZHANG W H, YU Y M. Effects of social network, social status and social trust on Residents' mental health[J]. Journal of Fujian Normal University (Philosophy and Social Sciences Edition), 2020(2): 100-111, 170.

[60] WEI X P, WU R J. The impact of social participation of the elderly on the risk of death in China[J]. Southern Population, 2015(2): 57-69.

[61] WANG F Q. Socioeconomic status, lifestyle and health inequality [J]. Society, 2012(2): 125-143.

YAO Xiao-ming, born in 1970, technical director of big data unit at the Cloud Branch, China Telecom. His main research interests include smart cities, mobile big data and data mining.

FU Xiao-ming, born in 1973, Ph. D, professor, IEEE fellow, IET fellow, ACM distinguished scientist, is a member of Academia Europaea. His main research interests include networked systems, cloud computing and big data analytics.

(责任编辑:李亚辉)