大数据和结构化数据整合的方法论*

——以中国人脉圈研究为例

罗家德　高　馨　周　涛等**

提要:本文以人脉圈层研究为例,将抽样调查得到的扎根真相与在中国广泛使用的一款社交软件 A 的大数据结合,建立人脉圈分类模型。在理论、数据挖掘、回归模型和分类预测模型及其解释工具的对话中,通过一次次抽样取得扎根真相,进行一轮轮的模型校准,发展出越来越精准的预测模型。本案例展示的大数据与结构化数据整合的研究范式是社会科学理论导引下的大数据研究方法论的实践。

关键词:大数据　扎根真相　人脉圈分类模型　社会计算学

一、大数据与结构化数据结合的研究范式实证研究

(一)社会计算学方法论

从理论视角出发,将传统问卷调查的结构化数据与大数据相结合的研究范式,不仅可以验证和修正理论,还可以筛选和计算出有意义的大数据指标,并形成具有推论性的预测模型。

与抽样调查所得到的结构化数据不同,大数据指的是基于网络、社交媒体、传感器、电子化文本资料等产生的电子印迹数据,具有量大、即时快速产生、类型多样以及价值密度低的特点(Bloem et al., 2012),其数据类型包括结构化、半结构化和非结构化数据,并且非结构化数据占据的比重越来越大。从网络上的集

* 感谢清华大学社会网络研究中心以及清华大学校内自主科研项目"基于通讯数据的关系强度与社会资本挖掘"(计划编号:2016THZWYY03)、腾讯研究院"基于 QQ 大数据意见领袖识别研究"(项目编号:20182001706)的资金支持。

** 其他作者包括腾讯 CSIG 云与智慧产业事业群刘黎春、德国哥廷根大学数学与计算机学院傅晓明、清华大学计算机科学与技术系刘知远和清华大学政治学系苏毓淞。

体活动、社交媒体、即时通讯到在线交易、政府情报和数字化图书馆,越来越多的社会生活留在电子文本中(Evans & Aceves, 2016),但这些实时产生的大数据90%以上为噪声数据,大量数据的快速产生对于存储和运算都是挑战(Sagiroglu & Sinanc, 2013)。要在如此大量的非结构化数据中获取有价值的信息,不仅需要采用高效的并行分布式处理技术,还增加了应用自然语言处理、影像处理、社会网分析和机器学习等工具来进行分析的需求,通过搜索、过滤、计算,最终将无结构化的数据处理为有价值的信息。值得注意的是,除了需要算法层面和技术层面的支持,理论和行业知识同样需要发挥重要作用。因此,社会计算学(social computing)作为一门发明处理复杂数据新工具的结合社会科学、计算机科学、数学建模和统计学的新学科,对于利用大数据来完成知识发现、理论探索和验证的意义重大(Lazer et al., 2009; Lazer & Radford, 2017)。

大数据的产生和计算机技术的发展使社会计算学得到极大的关注,新兴的社会计算学以前所未有的广度、深度和规模利用、收集和分析数据(Lazer et al., 2009),从而产生了新的理论与数据混合驱动的研究范式,大数据和调查数据整合的研究方法成为其中一个重要的部分。社会科学理论在大数据时代所发挥的作用只增不减,理论为大数据开启了很多新议题,为构建的预测模型赋予了推论的能力,因为只有形成理论才可以在不同领域、时间、空间下进行推论(罗家德等,2018)。

(二)社会学理论在大数据研究中的意义

通过整合传统调查数据与大数据建立的预测模型具有良好的推论性,可以有效降低调查规模和成本,同时对社会治理、政策制定的实时评估和地区资源供给具有重要意义。例如,在卢旺达的一项研究曾通过收集856个个人社会经济地位的调查数据,结合他们使用手机的数据形成的模型,就可以推论到卢旺达的150万人,即通过手机数据便可以预测他们的家庭经济情况以及所在社区的经济水平(Blumenstock et al., 2015)。因此,将获取成本相对较低的实时卫星遥感、手机通讯、社交媒体等大数据与调查数据整合作为扎根真相,①使用有监督机器学习、数据挖掘、网络分析等计算方法形成指标或模型,可以有效提升定量

① 扎根真相原是遥感学界的用语(Seager,1995),指的是将高空或卫星成像用于分析地面上(ground)被拍摄的真实物件到底是什么(truth)。此概念用于数据挖掘过程中则指挖掘出来的预测因子或行为模式在现实生活中到底存不存在,以及挖掘出来的目标现象和真实世界的"事实"到底有多少差别。

因果推论和预测社会经济态势的能力(Gao et al., 2020)。

在这种从理论视角出发的大数据与结构化数据整合的研究中,社会科学调查可以确立问题意识并提供扎根真相。理论概念需先有操作性定义,成为经验世界中可被观察的对象,或近似于可观察概念间关系的陈述,其中"可被观察"意味着在理论和社会科学的方法上是可测量的,而"近似"意味着构建的研究单元的本质是不能被直接观察的(Elragal & Klischewski, 2017),比如关系强弱无法被直接观察或标签,需要理论的定义使其可近似测量,然后开展调查以确定用户之间是由理论定义的哪一种关系强度。类似的无法直接测量的概念定义以及扎根真相的收集使得社会科学理论为大数据开启了很多新的议题。

(三)科学理论、调查方法和大数据挖掘方法的结合

同时,大数据算法使得一些"近似"的概念可以基于大数据指标建立模型得到局部最优解。例如,对于中国人的人脉圈层到底可以划分为几层这个问题,需要开展调查收集到用户之间的关系强度作为扎根真相,再从这些用户在社交软件中互动的印迹化数据中整理出指标,结合扎根真相建立分类模型,不断寻找准确率最高的划分方式,从而得到最合适的圈层结构,让这样的研究对象在一定的情境(context)下变得可被测量,这就体现了大数据和结构化数据相互对话的价值。

机器学习、深度学习等算法为处理高维、海量、多样化的大数据提供了解决方案。由此,综合考虑大数据的易获取性和收集扎根真相的难度,可选取一小部分群体做问卷调查收集扎根真相,建立和计算这部分人相应的大数据指标,然后用机器学习模型进行训练。选取的这部分的样本方式和质量决定着最终模型的推论范围,实践表明,如果该模型可以推论到其他具有相同维度的大数据指标的群体,就可以有效降低大规模问卷收集的成本和难度。在此过程中,通过抽样、问卷调查和标签,最终实现社会科学理论、调查方法和大数据挖掘方法的深度融合(Gao et al., 2019)。

因此,这种研究范式可总结为:从理论视角出发,带着"问题意识",合理且最大化地利用非结构化、碎片化的大数据来解释、预测个体与组织的行为以及社会现象,从而建立预测模型并对模型进行不断校准,最终得到最优的预测模型。在数据挖掘、预测模型不断校准的过程中与理论进行对话,并在学术社群中取得一定的"主体间性"(inter-subjectivity)的共识后,才能进行推论。原来社会科学实证研究领域主要从研究者各自提出的行为预设出发提出理论假设,由于数据

集或者其他条件的限制,一些理论和经验现象很难被重复,因此较难为所有学术社群所认同,往往存在很多争论。现在大数据可获得性的提高使得很多现象和理论可以被验证和再现,因此,在一些情况下,我们不需要一开始就基于严格的因果关系的假设建立强模型(strong model),完全依照模型的指定去测量变量、确定因果机制、设计问卷、收集数据、进行验证,而是可以先建立弱模型(weak model)(Evans,2020),①在数据挖掘与理论的对话中,逐渐发展和完善理论和模型,完成演绎和推理,再结合因果模型进行理论的发展或修正。②

在与大数据相关的研究中,在社会学理论与方法指导下以定性、定量调查得到的资料的依然价值不减,因为它是在理论导引下收集的和理论验证与发展紧密结合的数据,可被用来作为扎根真相,弥合了价值密度低、数据结构多样的大数据与理论间的鸿沟。换言之,理论指导下的调查可以提供检验数据挖掘结果的扎根真相(罗家德等,2018)。因此,大数据和调查产生的结构化数据的整合是社会科学理论导引下大数据研究的重要方法——它要求从理论视角出发提出问题,定义和收集扎根真相,然后设计和建立大数据指标来实现理论的证明或修正,让大数据的分析结合理论并创生价值。

二、中国人的人脉圈的分类模型的建立

(一)问题的提出以及研究设计

关系可以分为几类?在西方,邓巴提出了以功能划分的五种不同的互动模式(Dunbar,1993)。黄光国则提出在中国语境中存在三种不同的社会关系行为法则(Hwang,1987)。那么,在当下的社交媒体中,依照黄光国的三种关系法则,中国人的社会关系具体可以划分多少种?大数据指标是否能为关系强度的划分

① 由于调查数据成本高,必须有非常严格的模型指定(model specification)才能去收集资料。在收集资料前需要用理论定义所有行为变量,指定测量方法、因果机制、可验证假设(一个理论可能可以指定出许多可验证假设),因此称为强模型。而大数据中已经存在大量的行为资料和印迹化数据,所以初始模型在行为变量指定上可以不那么严格,而是在以后一轮又一轮的大数据挖掘与理论对话中逐渐完善模型,因此可以先建立弱模型。

② 在大数据挖掘后得到对理论的启发,展开和现有理论的对话或创新理论,需要建构理论假设和因果模型加以验证。范例可参考罗家德等(2018)。本文主要展现大数据结合结构化数据、多轮理论与资料挖掘间的对话过程,因此对于假设推导和因果模型建立以及验证过程不过多阐述。

增加新的维度？本文旨在应用大数据和结构化数据整合的方法，建立关系强度预测模型以及可解释和回归模型，对这一问题进行探索性的实证研究。

简言之，本研究的主要问题是利用社交网络数据来计算关系强度，刻画此情境下中国人的人脉圈可分为几层（Gao et al., 2020）。

本研究步骤具体包括：（1）在邓巴圈的理论（Dunbar, 1993）启发下，基于中国本土人情、关系和面子理论（Hwang, 1987），对中国人人脉圈层进行猜想和定义。（2）设计问卷收集扎根真相。（3）根据理论选取和计算大数据指标，将用户使用社交软件 A 产生的大数据变成有意义的预测因子，同时用数据挖掘的结果进一步与理论对话，校准和补充这些大数据指标的维度。在此过程中发现和修正理论。（4）根据大数据指标初步建立人脉圈分类模型，并依据数据挖掘结果评价已有理论的解释性。当解释性还有待提高时，尝试以新的数据挖掘结果或通过建立解释模型、回归模型来修正预测模型。（5）分析预测结果，总结错误样本分错的原因，进行新一轮的问卷调查收集扎根真相、改进大数据指标、建立解释模型、校准分类模型来提高预测准确率……如此一轮又一轮，不断逼近最优模型，找到最优的人脉圈的分层模型。

邓巴为解决人脉圈到底分为几层的问题开展了较为详尽的研究（Dunbar, 1992, 1993; Dunbar & Spoors, 1995; Hill & Dunbar, 2003; Zhou et al., 2005; Pollet et al., 2011）。邓巴根据心理学和社会学的构念，依据功能原则将人脉圈层划分为五层，从内到外依次是：（1）亲族支持团体（3 – 5 人）：内部的成员，往往是网络核心（ego）的姻亲成员或能给核心提供直接的意见和物质帮助的成员。（2）共情群体（15 人）：给核心提供情感支持的群体。（3）共宿群体（50 人）：由多个共情群体相结合，是一组保证成员安全且具有分工的群体（affinity group）。（4）社群或族系（150 人以上，可视为强关系）：核心所直接认识的社会群体。（5）部落群体（300 人，可视为弱关系）：可以被共同文化或非成文的规制约束的群体。

随后，邓巴使用脸书和推特数据去验证他的理论，将 185000 个用户与其互相关注的用户之间点对点发消息的频率抓取下来，针对每个用户使用无监督的 k-Means 进行一维聚类，得到推特中最佳的分类结果为 5 类，脸书为 4 类，然后去寻找与不同类簇对应的理论假设来为其打标签。假如平均每个用户有一个 5.28 个好友的类簇，在数量上相近的即为 3 – 5 人的亲族支持群体。这样，邓巴获得了内层的三个群体与聚类结果数量上的对应关系，但最外两圈的社群和部落则并不明确（数量关系没有得到对应）。而且该研究使用无监督

的方式进行聚类,以寻找与理论猜想数量上的对应关系,没有扎根真相进行验证,在脸书和推特上结果也不一致,所以并没有完全证明其理论(Dunbar et al., 2015)。

在中国,对个人人脉圈层的研究可以追溯到费孝通先生(1998)的差序格局理论,它阐述了中国人区分亲疏远近的人际关系格局,成为个人人脉可以由内而外按照不同原则划分不同圈层的概念雏形。黄光国(Huang,1987)进而将中国人的人脉按照人际互动法则分为需求法则、人情交换法则、公平法则三种。人脉圈最内层遵循需求法则,即不论成员贡献多少,彼此都会无条件地支持对方。第二层遵循的人情交换法则依靠的是人与人之间存在"人情账",是一种兼具情感和工具型关系的混合关系(mixed tie)。最外层遵循的公平法则基本上是短期的工具交换型的关系,其核心依据规则分配资源,不存在长期的人情交换。但黄光国的研究并没有指明依据三种法则可以将关系划分为几类,而当下社交网络中关系可以被划分为几类的问题也有待解决。

黄光国的人情、关系和面子理论阐明了中国人进行人际交往的行为法则,那究竟是每一法则即对应一种关系类别,还是存在邓巴式的更加细致的划分方式?为了解决这一问题,本研究先按照邓巴的划分方式将熟人关系划分为亲密熟人、一般熟人和潜在熟人,但这还需要进一步验证各层之间是否存在明显的区别。基于此,我们在问卷设计上首先将中国人人脉圈层划分为五层——符合需求法则的家人视同 3 – 5 人的亲族支持团体,工具性关系则视同最外圈的部落群体(弱关系),熟人则具体分为亲密熟人、一般熟人和潜在熟人三个圈层。当然,后续需要通过预测结果与理论对话,以进一步确定最佳的分层方式。以下我们就借此问题来探索理论导引下大数据与结构化数据的整合过程。

(二)问卷调查收集扎根真相

本研究的目的是探索社交网络情境下的划分关系圈层的模型,以回答关系可以分为几层的问题,并识别出区分关系强度的大数据指标。在社交软件中存在着大量的用户间互动的印迹数据,要如何识别和筛选这些指标才能识别出两个人之间的关系强度?这就需要用扎根真相进行验证。得到扎根真相最好的方法就是去做调查,调查被访者与他人的真实的关系类别。换言之,社会科学理论不仅设定议题、指导算法,同时,社会科学的实证方法,不管是定性访谈还是定量调查,都为进一步验证算法的预测准确度提供了有效的手段。

如上所述,我们将关系分成五层(Luo & Yeh, 2012),定义描述见表 1。

表 1 **人脉圈层本土化定义**

圈层序号	群体名称	描述
最内层	家人、拟似家人	最内层依据的是需求法则。这一层的人遵循平等原则,即不论成员贡献多少,都会无条件地予以支持,换句话说就是有福同享有难同当。
第二层	亲密熟人	第二层为亲密朋友,是一种同时存在情感关系和工具型关系的混合关系,但主要是情感关系,大多为铁哥们儿、闺蜜等。
第三层	一般熟人	第三层的人情交换法则是人与人之间的“人情账”和情感关系,是一种同时存在情感关系和工具性关系的混合关系(mixed tie)。
第四层	潜在熟人	第四层是有可能未来成为有人情交换关系的人,有继续与其交往的意愿,但现在遵循的是平等法则,不存在人情交换,为工具性关系。
第五层	认识之人	最外层平等法则是按照规则执行,依据贡献分配资源,不存在长期互换人情,只是一种工具性关系。

因为邓巴所提出的可以维持的好友的数量过于庞大,这对于问卷调查来说难度过大,实际调查中不可能调查到每个人的所有联系人,这样容易造成问卷填写质量下降、内容失真等问题,因此在本研究中每位被访者需要至少填写 26 位社交软件 A 中的好友以及与他们的关系。其中,最内层 3 人,中间三层各 5 人,最外层 8 人。考虑到问题的复杂性,调查由少数专门培训的访谈员以面访为主、电话访谈为辅的方式进行,以便解释清楚问题的内容。问卷设计①的五种关系类别分别描述如下。

1. 在生活中,我们身边会有一些与我们最亲密的人,常常是家人或被视同家人的亲密好友,大多在 5 人以内,请您浏览好友列表,至少填写三位这样的人。

2. 有一些人是我们的铁哥们儿或闺蜜,或是经常联系的亲戚,但不如上述家人与我们那样亲密,请您试着填写至少 5 个这样的人。

3. 有一些人我们愿与其长期交往,是相互欠人情、相互帮忙的关系,请您至少填写 5 个这样的人。

4. 有一些人我们尽管认识,现在会联系,但不一定与其长期交往,也并不一定会互相欠人情,可能会也可能不会发展出情感关系,请您试着填写至少 5 个这样的人。

5. 有一些人是网上的朋友,但我们暂时没有打算与其长期交往,是一种不太会发展为朋友关系的工具性关系。

另外,调查内容还包括被访者和其指认好友在社交软件 A 中具有唯一标识

① 问卷题目对关系的定义在每轮调查中保持不变,但结合每次问卷收集的情况,建议填写的人数会有所调整,本文中所列的是第四轮调查问卷建议填写的人数。

性的 ID,以便将扎根真相与用户在社交软件 A 中的大数据进行结合。合作公司在匹配资料库后进行匿名化处理,以确保分析人员无法逆向推论用户个人信息。

接下来,需要找出大数据中的各类指标并将这些指标组合成算法,但要回答最后算法与理论指涉的现实概念到底有多大落差,需要不断收集数据以验证并修正算法、结合因果模型来不断逼近现实。

本研究通过四轮问卷调查对模型进行验证和校准(如表 2 所示),同时每一轮调查都保持相同的问项和调查方式,但也会对问卷进行部分修正,例如在前三轮问卷中要求被访者在五个圈层中最内四层至少各填写 5 人,最外层至少填写 3 人,但是在训练过程中发现最外层由于样本量过少而无法学习到其特征,导致最外层准确率非常低。因此,为了修正最外圈人数不足的问题,我们在第四轮问卷中改为最内层至少填写 3 人,最外层至少填写 8 人,这样的设计也更加符合现实中的实际情况,因为实际情况中家人的人数一定少于次外层亲密熟人或一般熟人的人数,更少于最外层认识之人的人数。

由于实践中收集标签数据的无限采样的成本较高,但是大量的没有标签的大数据指标的训练是没有意义的,无监督学习受初始值、参数影响较大,并且无法仅靠数据和算法提取和识别最重要的特征,因此需要用扎根真相来打标签。但是,传统社会学采用随机抽样的方式,如图 1(b)所示,不仅需要大量人力物力成本,问卷回收率也较低。当问卷调查与大数据相结合后,我们可以采用最适化的抽样方式(optimized sampling)(Evans, 2020),构建初步预测模型后,再在新的抽样集中进行验证。需要注意的是每次抽样的关键问项要保持不变,但是样本集的比例或其他问项可能会根据每次结果进行调整和改进,然后根据每次新增的扎根真相不断改进预测模型,如图 1(a)所示,进行一轮一轮的校准,通过实验设计、算法设计和理论的对话和相互修正,最终不断逼近最优解。但在有条件的情况下还应做随机抽样进行验证,以便在更大范围内进行推论。

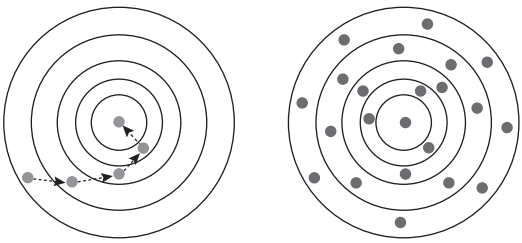


图 1 最适化抽样与随机抽样对比

(三)计算和选取大数据指标

在本研究中,依托社交软件 A 公司云计算平台提供的强大算力和存储能力,首先从每天以数百 TB 级增长的社交软件 A 的数据中找到有扎根真相的特定用户,由资料预处理人员独立提取该用户与其他人的互动记录并构建互动网络,之后再将匿名化的数据交给分析人员进行分析。社交软件 A 目前是在中国普遍使用的一款应用软件,拥有庞大的用户群体和较长时间的数据存储,因此对于本文所研究的在中国情境下社交媒体的使用者具有很好的推论性。本文抽样方式为最适化抽样[见图 1(a)],问卷收集情况见表 2。

关于大数据指标计算,在既有理论中划分关系强度需要关系久暂、互惠内容、联系频率和亲密程度等(Granovetter, 1973)指标来进行表征。伯特(Burt, 1992)及科尔曼(Coleman, 1990)等人从关系网络结构的角度出发认为关系结构塑造关系强度,也有研究(Lin & Vaughn, 1981)指出两个人之间的年龄、性别、教育程度等相似性影响关系强度。关系久暂可以通过成为好友时间来表征;出于对用户点对点聊天内容的隐私保护,对互动内容未进行自然语言分析,而是通过互发红包、互赠礼物的频率来表征;联系频率可以通过二者工作时间、非工作时间发消息的频率和工作时间、非工作时间互动频率标准差来体现;亲密程度可以根据见面频率、好友列表分组备注所体现的亲密度、点对点电话、视频频率来初步表征;同时,共同好友、共同群等结构性指标也是识别关系强度的重要指标(Shi et al., 2007)。

表 2 **问卷调查情况**

调查次序	调查时间	地点	样本量
第一次	2017. 04. 19 – 2017. 05. 21	上海市、北京市	96 名用户的 1749 条关系
第二次	2017. 10. 01 – 2017. 10. 12	北京市	29 名用户的 580 条关系
第三次	2017. 11. 05 – 2017. 11. 16	北京市	37 名用户的 736 条关系
第四次	2018. 8. 1 – 2018. 8. 6	北京市、上海市、广州市	66 名用户的 1654 条关系

理论导引下的指标体系的构建不仅可以为在海量数据中找到数据挖掘预测因子提供方向,并且可以通过理论对话和行业知识帮助研究者从原始指标中细化出新指标。例如在本研究中,由于工作时间和非工作时间联系规律在不同的圈层中往往有不同的表现,一般在工作时间内同事之间比其他关系可能产生更多联系,因此,区分工作时间和非工作时间对于关系类别的划分非常重要。在行业知识的帮助下,在时间维度上划分出更具区分性和解释性的新

指标。因此,指标的选取就是一个理论、行业知识与数据挖掘混合驱动来确定的过程。

在指标选取过程中,先计算这些指标与关系强度之间的相关关系(如皮尔森相关系数),得到点对点发消息频率、工作时间互动标准差以及好友备注亲密度和关系强度的相关性最显著。具有理论意义和实际价值的指标还需通过预测模型、可解释模型和回归模型进行综合判断,让预测模型和理论互相补充和修正。

(四)理论指导下建立初步的人脉圈分类模型

基于上述理论回顾,本文首先建立了五层分类模型(见表1)。因为在五层、四层、三层这三种分类中,五层分类模型最严格,所以直接使用绝对准确率来比较不同分类方式的表现是不公平的,由于训练预测模型前对样本进行了过采样(over sampling)处理来解决每类样本不均衡的问题,因此三种分类方式随机猜测的准确率分别为20%、25%和33%,这里将随机分类模型作为基准模型(benchmark model),将每种分类器相对于基准模型所提升的准确率作为评价分类器表现好坏的标准。分类模型采用有监督机器学习模型,有经典的支持向量机模型(SVM)、决策树(decision tree)、logistic回归、随机森林(random forest)等算法、集成算法梯度提升树模型(gradient boosting classifier)以及XGboost模型等,最终发现XGboost模型表现最优。另外,在本文报告结果中,数据集的80%作为训练集,20%作为测试集,在实验中可以有效避免过拟合和欠拟合,同时保持了较高的准确率。最终,以该数据集划分比例,以使用XGboost模型得到的准确率相对于基准模型提升的相对准确率作为衡量分类模型表现的主要标准。

1. *建立初步的五层分类模型*

基于问卷调查中初始的划分方式,即家人/拟似家人、亲密熟人、一般熟人、潜在熟人和认识之人五种关系类型,结合大数据指标建立有监督的分类模型。最终得到五层模型最高的准确率为49.25%,较基准模型准确率(20%)提升了29.25%,准确率还有待进一步提高。

2. *建立初步的三层分类模型*

为建立三层分类模型,我们尝试将五种不同圈层归类在不同分层法则下,例如:以家人/拟似家人为最内层,三类熟人(亲密熟人、一般熟人、潜在熟人)为中层,认识之人为外层;以家人/拟似家人与亲密熟人为最内层,一般熟人为中层,潜在熟人和认识之人为外层等……共有6种排列组合的方式,最后预测准确率

最高的模型是:以家人/拟似家人为最内层,亲密熟人、一般熟人两层为中层,潜在熟人和认识之人为最外层,这与黄光国(Hwang,1987)的三种互动法则最一致。该分类模型准确率为 68.16%,较三层基准模型准确率(33%)提升了35.16%,①高于五层模型相对于基准模型准确率提升的 29.25%,从准确率提升的显著程度上看,三层模型比五层模型对扎根真相更具解释力。

3. 建立初步的四层分类模型的探索

那是否存在一种四层模型——比三层模型划分更细致,同时又没有五层模型那样严格,对扎根真相有更强的解释力?

我们基于扎根真相,针对每一圈层大数据指标进行分析,这里重点观察与关系强度相关性最高的点对点发消息频率和互动时间标准差(互动时间标准差、工作时间互动标准差、非工作时间互动标准差),在五层分类下,计算每一圈层上述指标的平均值,得到点对点发消息频率、互动标准差这两项在第四层和第五层之间平均值之间的差异显著小于其他圈层之间,说明第四和第五圈层可能具有相似的互动逻辑,并不存在明显的分类界限,这也从一定程度上解释了上述三层模型准确率优于五层模型的原因。但是在三层模型中,第一层中的家人和中层的亲密熟人(五层模型中的最内层和第二层)在这些指标上却具有较为显著的差异,亲密熟人相较于家人,不管是点对点发消息频率还是互动标准差都有一个较大的断崖式的下降。因此,理论与数据共同启发我们,构建四层模型可能会有更好的效果。由此,将可能不存在明显分类界限的最外两层进行合并,建立四层分类模型。当然,就此确定分类方式还太过轻率,还需要与其他三种组合方式进行更加严谨的比较,例如将家人/拟似家人和亲密熟人划分为一类、其他圈层保持不变等,尝试找到四层分类模型更好的划分方式。

基于上述理论猜想与数据挖掘间的对话,在五层分类模型基础上,将最外两层潜在熟人和认识之人进行合并,均遵循公平法则。同样使用有监督的分类算法建立四层分类模型并与基准模型进行对比。得到四层有监督的分类模型准确率为 52%,相对于基准模型准确率(25%)提升了 27%,并没有高于三层模型所提升的准确率。我们也尝试了其他的四层划分方式,无论是将家人与亲密熟人合并为一类,还是将亲密熟人和一般熟人合并为一类,最终准确率均低于上述划分方式。

① 将问卷的五类按不同的组合方式分成三类,最后以家人与拟似家人为最内层、亲密熟人和一般熟人为中间层、潜在熟人与认识之人为最外层的划分方式预测效果最好。

但就此得到三层模型是最好模型的结论还太过轻率,我们需要继续进行理论与数据挖掘的对话,进一步提升模型的准确率。

(五)理论与数据混合驱动下模型的修正

在模型的修正阶段,基于理论形成的指标可以结合用于解释预测模型的SHAP工具(SHapley Additive Explanations)(Lundberg & Lee, 2017)进行特征分析。SHAP可以计算预测模型中单个样本以及总体样本每一个特征的重要性,在很大程度上解决了机器学习模型预测路径中的"黑箱"导致的无法捕获样本预测的异质性、识别异常值和发掘特征之间的交互性对预测结果产生的影响等问题,将预测模型的"黑箱"变为"灰箱"(Muhamedyev et al., 2020)。但在理论层面上,我们还不能完全满足于仅对模型进行事后解释,还应该与已有理论对话,得到更多可推论、具有普遍性的经验知识来验证、补充或修正理论。

因此,本文在理论与数据混合驱动下对模型进行多轮修正,先基于初步构建的预测模型及SHAP对模型的解释,分析预测错误样本的错误原因,使用理论和行业知识更进一步地处理和筛选数据,在理论指导下增加指标,用以干预和修正预测模型。修正后继续使用SHAP来判断上述修正对模型准确率的提升效果。SHAP分析简化并加速了这一修正过程,最终在修正过程中获得了更多细化的全景式的知识以及更优的预测模型。

1. 第一轮修正

如表3所示,针对初步建立的模型,本研究通过考察分类错误样本情况对模型进行修正。实际值与预测值的混淆矩阵(confusion matrix)可以很好地区分出分类错误的样本,并可以得到具体错分到哪一圈层,进而可以对这些样本的扎根真相和大数据指标进行分析,考察分错原因,对模型进行修正。这也是以理论导引的研究范式与单纯的数据挖掘的区别。因为有扎根真相,可以考察分类错误样本的大数据指标的特征,从而有针对性地修正模型,同时就理论的不足之处进行补充和修正。表3为五层分类模型的混淆矩阵,家人为最内层,认识之人为最外层。同样,我们也生成四层、三层模型的混淆矩阵进行分析(此处略去)。

从实际值与预测值的混淆矩阵中发现,三种划分方式均存在大量实际在最内层、第二层样本被错分到最外层的样本。为此进行两方面探究:一是基于上文初步得到的预测模型计算SHAP值进行特征分析,以识别有效特征;二是分析这些分错样本的互动行为,从数据中还原个人特征,以得到其分错的原因,对这部分人进行更细致的划分,生成更多有意义的指标来提高预测准确率。

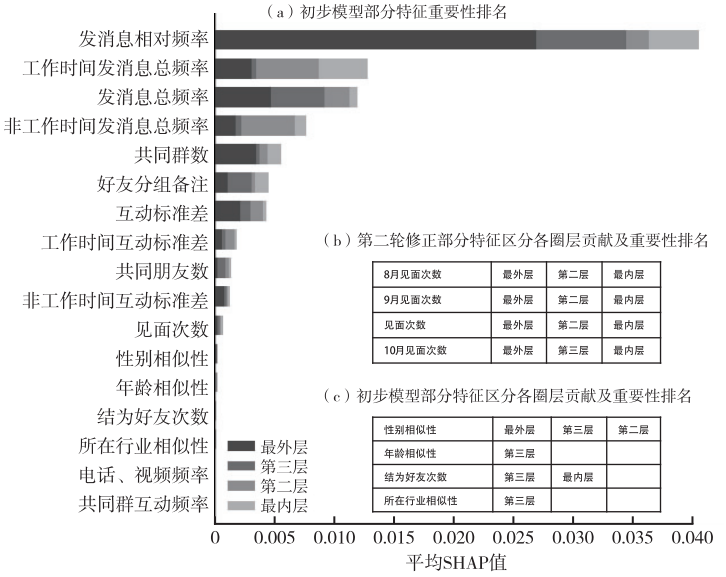
表 3 五层模型实际值与预测值的混淆矩阵

真实值 \ 预测值	最外层	第四层	第三层	第二层	最内层
最外层	122	1	13	2	2
第四层	30	36	4	1	0
第三层	44	6	13	6	3
第二层	29	7	13	17	7
最内层	25	2	6	3	10

注:竖列为真实值,横行为预测值,对角线为分类正确的样本量。

使用 SHAP 对于特征值重要性进行分析,由于指标中视频通话次数、点对点发红包、转账频率缺失值过大,虽然采用缺失值插值技术进行插补,但放入预测模型对于模型预测准确率的提升反而存在干扰,因此舍去。如图 2(a)所示,得到发消息相对频率对于模型整体预测准确性的贡献最大,尤其是针对最外层模型,这与已有理论相一致(Marsden & Campbell, 1984)。另外,工作时间发消息频率对提高最内层预测准确率贡献最大,非工作时间发消息频率对第二层亲密熟人的准确预测贡献最大,即在社交媒体上,人们更倾向于在非工作时间与亲密朋友进行互动,即使在工作时间也会与家人互动。另外,有一些指标虽然对于提升模型整体准确率贡献不大,但对于特定圈层的识别是有意义的。例如性别相似性指标有利于识别出第二圈层,而第二圈层多为铁哥们儿、闺蜜,即人们更愿意与性别相同的人建立亲密熟人关系;年龄相似性、行业相似性有利于识别出第三圈层,而第三圈层多为同学和同事关系,因此年龄和行业对该圈层识别有意义[详见图 2(c)]。SHAP 有助于更精准地捕捉到这些变量,它们对全体样本的解释贡献不大,却可以提高某些特定圈层的预测准确率。这也启示我们,预测模型每一圈层划分所用的指标和模型存在差异。

我们在分析错误样本时还发现见面频次对于模型预测准确率影响较小,这与已有理论不一致(Marsden & Campbell, 1984)。在对见面频次进行更细粒度的划分后,发现不同时间段的见面频次存在较大不同。通过混淆矩阵分析发现,由于社交软件 A 使用者平均年龄较小且多为学生,在假期的见面频次会更有助于识别其内层亲密关系,而上学期间与朋友见面的频次会增加。因此我们对见面频次做了更细的划分,分为假期的 8 月、开学的 9 月以及国庆假期三个时间段。使用 SHAP 分析发现,划分假期、非假期的见面频次可以提高模型预测准确率,8 月对于最内层家人预测准确率贡献最大,9 月对于第二层亲密朋友和最外层认识之人预测准确率贡献最大[详见图 2(b)],结果与上述分析一致。在建立



注：大图（a）为可视化的初步模型部分特征重要性排名，右下角（b）为第一轮修正后模型8月、9月见面次数与10月见面次数等部分特征重要性排名，（c）为初步模型中部分特征局部放大图，不同颜色代表不同圈层，条形框越长表示该特征在总体样本中平均SHAP值越高，即越重要。

图2 第一轮修正前后预测模型特征重要性排名（以四层模型为例）

大数据模型时，由于存在很多低价值的噪音数据，因此利用已有理论知识和信息将这部分数据排除或拆分出更细粒度的指标也非常重要。扎根真相和预测算法的结合利用混淆矩阵以及SHAP分析可以有效地提供这些先验知识来完善算法，然后用新数据来对算法效果进行验证，避免单样本的回音效应。

修正后最终得到新的五层、四层和三层的模型的准确率分别为54.70%、73.17%、76.66%，相对于随机分类，准确率分别提升了34.70%、48.17%和43.66%。本轮模型修正后，四层模型准确率提升最为显著，因此，就第一轮修正结果看，四层模型对扎根真相的解释力最强。

2. 第二轮修正

进一步来看，上一轮修正的模型是否还有改善空间？新的数据挖掘结果又一次为我们带来启发。我们继续绘制3×3、4×4、5×5的混淆矩阵，分析错误样本分错原因，对每一圈层之间和每一层内不同属性、特征的关系进行更细粒度的划分，以期找到更有意义的指标，寻找其互动模式的异同。通过计算这些大数据指标的标准差，观察每个圈层的标准差的变化情况，得到最内层相对频率等指标的标准差最大，并且从最内层到最外层标准差依次降低，结合本轮实际值与预

测值混淆矩阵进行分析,发现同一圈层的家人也有较大的差异,那这种差异是什么引起的呢?邓巴曾指出,家人是否住在一起会影响其互动模式(Dunbar et al., 2015)。由于社交软件 A 中有用户实时的位置信息记录,通过混淆矩阵的分析,对错误样本的位置信息与扎根真相对应的关系强度进行比对,发现住在一起的家人和不住在一起的家人有较为不同的互动模式。因此,这启发我们先用地理位置信息来计算两个用户在 00:00 – 5:00 是否经常出现在同一地点,以判断二者是否居住在一起,并使用 SHAP 进行交互关系的分析,得到该变量与点对点互动指标(相对发消息频率、工作时间互动频率标准差、非工作时间点对点发消息频率)以及好友分组备注的交互项都具有显著交互作用,并且这种交互作用可以有效提高预测准确率。因此,在使用社交网络互动的情境下,是否住在一起和联系频率、亲密程度具有显著的交互作用,有利于关系强度的识别。结合预测模型发现,加入该变量对于模型预测效果也有提升作用,同时家人这一层的准确率有了明显的提高。

经过本轮修正,得到新的五层、四层和三层的模型的准确率分别为 54.36%、74.22% 和 77%,相对于随机分类,准确率分别提升了 34.36%、49.22% 和 44%。本轮修正后四层模型依然解释力最强。

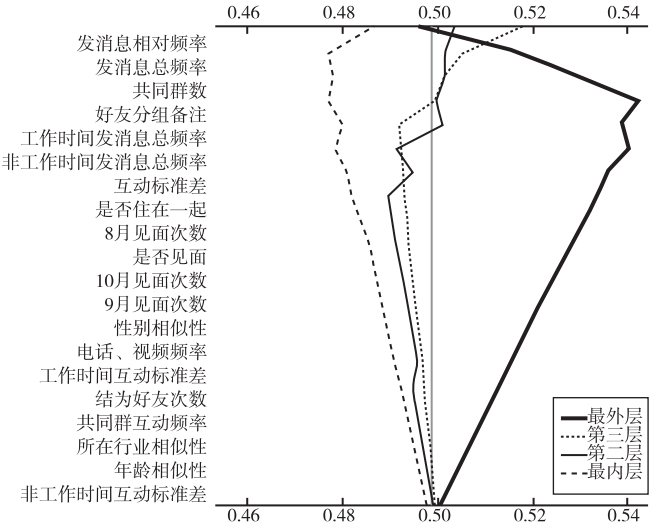
3. 第三轮修正

继续分析混淆矩阵得到最外层和次外层之间也存在较多混淆。使用 SHAP 分析分错误样本的决策路径,以某一分错样本为例,该样本本应该为最外层,却被错分为第三层(如图 3 所示),发消息相对频率和发消息总频率这两个指标对最终样本被错误分类为第三层影响较大,因此我们计算了更长时间段的大数据指标,尝试让二者在更长的时间段内(2017 年 8 月 –2018 年 10 月)进行区分,计算除地理位置信息外的其他指标。结果表明更长时间段的指标更有助于区分最外面两个圈层,准确率进一步得到提升。最终五层、四层、三层分类模型准确率分别为 60.19%、76.39% 和 77.78%,相对于随机分类,准确率分别提升了 40.19%、51.39% 和 43.78%。本轮修正后四层模型依然解释力最强,对于模型的解释性最强,预测模型最逼近扎根真相。

(六) 总结

1. 最佳的分类模型

经过几轮修正(如表 4 所示),可以得到如下结论:(1)三、四、五层模型相对于无监督分类模型以及随机分类的基准模型准确率都有很大的提升,这说明经



注:纵轴为特征,图中最粗的实线为该样本实际分类,即最外层,最终样本被错误预测为第三层(输出值最大)。根据决策路径可知,发消息总频率和发消息相对频率两个指标对最终样本被归类错误影响较大。

图 3 某分错样本决策路径图

过理论导引下扎根真相和大数据的整合,在理论与预测模型不断的对话中,模型越来越接近最优解;(2)在修正过程中,四层模型的准确率相对于随机分类模型提升最为显著[51.39%(四层)>43.78%(三层)>40.19%(五层)],说明四层模型是三者中最优的模型,在此情境下最具解释力。

表 4 四轮修正准确率对比

	五层模型	四层模型	三层模型
随机分类模型	0.20	0.25	0.33
初步模型	0.4925	0.5200	0.6816
第一次修正	0.5470	0.7317	0.7666
第二次修正	0.5436	0.7422	0.7700
第三次修正	0.6019	0.7639	0.7778

从表 4 几轮修正得到的准确率结果来看,四层模型更加适合本研究所针对一线城市 18 – 25 岁青年的人脉关系分类。当然,模型的准确率还有相应的提升空间,将来我们还需要不断进行新的问卷调查,继续开展理论和预测模型的对

话,以进一步修正模型。同时本研究也是一个解释性和预测性相结合的研究——在解释性研究中拆解预测模型“黑箱”。

以上分析启发我们,获得模型的可解释性主要有两个方面:一方面是从算法上获得更多的可解释性,例如本文所使用的 SHAP 值分析,使得黑箱逐渐变得可见;另一方面是与理论不断进行对话,例如在混淆矩阵中找出分错的样本,分析分错的原因,看现有结论是否与理论相一致,思考理论还可以为我们提供哪些洞见。结合理论来分析现有预测结果,计算更多有实际意义、解释性强的指标,尤其对于不同人群更细粒度的划分、不同互动模式的识别以及不同特征之间的交互作用等,有助于启发我们构建更完善的区分关系强度的指标体系。我们不能以牺牲解释性为代价来追求预测准确性。正如霍夫曼和沃茨等(Hoffman & Watts, 2017)所说,解释和预测是相互补充而不是相互替代的,我们不能忽视解释的重要性——加深对数据的理解,澄清概念上的分歧。因此,在几次不断对话的过程中,本文主要完成了以下几个方面的工作。

(1) 进行多轮的问卷调查,在新的数据集上不断对模型进行检验和修正。

(2) 以理论导引指标筛选。基于已有理论中测量关系强度的指标,在中国情境下考察社交软件中哪些指标可以明确识别关系强度以及指标之间是否存在交互作用,从而完成对理论的补充发展,并在预测结果后与理论进行对话。

此外,为进一步探究这些指标在统计上的显著水平,我们使用回归模型对基于理论建立的各项指标进行综合分析,最终发现在社交软件 A 中联系频率、亲密程度、结构性指标对关系强度的影响具有统计意义上的显著性。关系久暂中成为好友时间没有统计显著性,好友次数也不具有显著性,说明在社交网络 A 中成为好友时间不与关系强度线性相关。在调节项分析中,同样得到是否住在一起和联系频率的交互项与某些圈层的关系强度存在统计上的显著性,年龄、性别、行业相似性等指标在统计上不显著(详见 Gao et al., 2020)。

综上,我们在数据挖掘分析大数据指标的特征中,结合了上述 SHAP 对于预测结果的解释;对每一圈层各变量平均值、方差与关系强度的相关性进行了分析;利用回归模型探索了部分指标在统计上的显著性;建构了关系强度的计算算法和分类模型。在该分析过程中,我们找到四类指标:第一类为具有统计显著性又可以提高预测准确性的指标,在分析中可探究这些大数据指标所代表的理论意义;第二类是有统计显著性但对于提高预测准确性作用有限的指标,可进一步分析,拆分、重组或计算交互项以提高分类准确率,由此可以得到既具有显著相关性又可以提高预测准确率的指标;第三类是不具有统计显著性但可提高预测

准确率的指标,这类指标具有应用价值,但理论意义待发现,可结合 SHAP 进一步分析了解其指标是否具有推论性和普遍性;第四类为既不具有统计显著性又不能提高预测准确率的指标,分析中可以分辨为噪音,并予以排除(如表 5 所示)。

表 5 **指标类型**

	提高预测准确率	未提高预测准确率
具有统计显著性	兼具理论意义和应用价值的指标	具有理论意义,但可能与其他变量存在共变等关系,应用价值待开发的指标
不具有统计显著性	具有应用价值,理论意义待发现的指标	噪音指标或指标计算方法、质量有待提高。

因此,在指标的计算、验证、解释、排除中,本文最终证明或补充、修正了已有理论。这里需要说明的是,本文主要探究的是与关系强度存在相关性的指标,在其他研究中,当需要探究更严格的因果关系时,也应将预测模型(Galdo et al., 2019; Ghoddusi et al., 2019; Gu et al., 2020)及可解释模型(Lundberg & Lee, 2017; Strumbelj & Kononenko, 2014; Muhamedyev et al., 2020)结合因果推断模型(Granger, 1969; Runge et al., 2017; Athey & Imbens, 2019)进行综合分析(Ghose & Ipeirotis, 2011),让预测模型具备更广泛的推论能力。

(3) 对分类后预测有错误的样本进行分析,可建立实际值与预测值的混淆矩阵,结合 SHAP 来比对分错样本的扎根真相与大数据指标的特征,探究错误原因,进行与理论的对话,重新设计和改进分类模型,进一步提高模型准确率。

(4) 尝试使用不同的分类模型,本研究在通过已有理论得到五层和三层模型后,在与大数据挖掘的对话过程中发展出了新的四层分类模型,可以得到目前为止最接近最优解的结果,从而提供了一个在中国社交媒体情境下计算关系强度可行的测量方法,也可以用于后续更多理论和应用研究。本研究共进行了三轮模型的修正,每一轮模型修正都使得模型的准确率有了不同程度的提高,最接近最优解的四层模型的准确率相对于其他模型有更显著的提升。未来还需要在跨时段、更大量的样本收集等方面进行修正,进一步提高准确率。

2. 理论发现

在理论与预测模型相互对话的过程中,本文在理论上亦有所发现,因为本案例重点展示探索性研究过程,所以省略了理论推演、建立假设、验证假设的过程,但仍有一些或经过验证或有待验证的发现提供了理论洞见。

(1) 在社交软件中联系频率、亲密程度、结构性的指标与关系强度具有显著相关性,并且工作时间的相关指标和非工作时间的相关指标对于区别不同圈层具有不同方向的影响,因此区分工作时间和非工作时间是有意义的。

(2) 在相似性指标中,年龄、性别、职业行业相似性对于整体预测准确率贡献程度较小,但在预测模型中,性别相似性可以提升第二层亲密熟人圈层预测的准确率,年龄相似性、行业相似性可提升第三层一般熟人的预测准确率。因此,性别相似性、行业相似性与关系强度虽然不存在显著线性相关性,但对于特定圈层的准确率提升具有作用,进而说明二者之间可能存在非线性的关系,后续还有待进一步探究。

(3) 在社交网络情境下,是否住在一起与联系频率、亲密程度的交互项对于关系强度识别有显著作用。

(4) 研究中最大的发现是亲密熟人和一般熟人能够区分开来,虽然都是兼有情感性与工具性的混合性关系,但明显前者的情感性较强,人情交换程度较高,在互动模式上也表现出明显不同的特征(Gao et al., 2020)。

3. 预测分类模型的阐释

本文采用的是有监督的分类模型,有监督的分类模型就要涉及使用调查问卷收集的扎根真相来训练模型,输入的是用户之间互动的大数据指标,输出的是关系类别,因此只有合适的划分方式和模型才会达到较高的准确率。过细或者过粗的分类都会使最后的准确率较低,因此需要不断地验证和改进分类模型,从而找到最优的分类方式。为了更加清晰地说明这种情况,这里利用图 4 所示的分类模型确定过程进行说明(杨鲲昊,2018)。每一个方块、圆圈、三角和平行四边形代表在特定情境下实际真实存在的几种关系强度的类型,虚线为分类方式。(a)为没有进行分类的原始数据,算法开始不知道要分几类;(b)为一个合理的

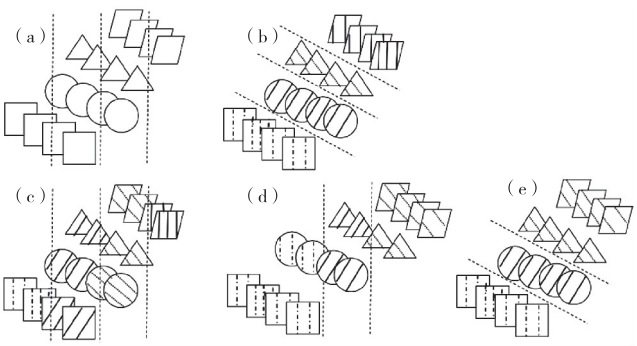


图 4 不同分类模型的比较

分类方式,这种分类方式得到了真实的数据所属的类别,具有较高的准确率(图4中为理想情况);(c)为一个过于精细的分类器,使得本来是一个类别的数据没有分到一起,准确率降低;(d)为一个过于粗糙的分类器,将不同类别的数据分到一起,也会使准确率降低;(e)同样是一个过于粗糙的分类器,与(d)不同的是,虽然这里有监督的分类标签准确率与(b)相同,但是相对于随机分类从四类随机分类25%的准确率提升为三类随机分类33%的准确率,从(b)到(e)准确率并没有明显的提升,也就是说分三类与更严格的分四类相比准确率不再得到明显改善,甚至低于四分类的准确率,故而可以判定这种分为三类的方式过于粗糙。

三、讨论:数据挖掘、理论与预测模型三者的互动

过去也曾有过类似的对话,例如非参数模型(预测导向)与参数模型(理论驱动)的交锋、贝叶斯学派(弱模型)和频率学派(强模型)的争论。这次的竞合可以视为过去的方法论对话的延续。不同的是,过去不同方法的争论是以结构化数据为基础的,此次的竞合则是在大数据的基础上探索方法融合的可能性。

本案例以基于大数据的人脉圈分类模型研究为范例,探究如何利用大数据进行初步的理论探索研究,说明了如何用理论指导设计问卷、调查、收集扎根真相,用理论指导大数据指标的计算和筛选,用大数据指标启发理论的探索以及模型的建立和改进、收集更多的扎根真相,再进行新一轮的指标计算、模型修正与理论对话,再用修正的理论指导分类模型的构建,用新资料验证新预测……如此一轮又一轮,最终找到最接近最优解的分类模型,预测精度不断提高、推论范围不断扩大,通过调查数据和大数据的整合,理论、扎根真相、大数据指标及预测模型的对话,理论和模型均处于螺旋式的上升和发展中。

需要注意,这里的研究主旨并不是证明邓巴圈理论或中国本土关系理论在中国情境中正确与否,而只是以两种理论综合得到的弱模型开始展开一系列的三角对话。到目前为止,研究得到的阶段性结论是本文所提出的四层模型——也就是在人情、关系、面子互动法则下中将熟人又划分为亲密熟人及一般熟人。随着预测精度越来越高,会不会找到更精准的模型从而修正现有的结论?这有待更多轮的三角对话。比如当四层模型的预测精准度从现在的76%能成长到90%时,基本上就可以认为三层模型划分过于粗糙,未能完善地描述出中国人的关系类型。又比如当五层模型经过多轮对话后仍无法接近四层模型的预测准确度时,就可以得出

四层模型是最能描述中国人关系的分类方式,进而可以用来修正现有理论。在此基础上,可以进一步从亲密熟人与一般熟人的实际行为中探究这两类人的交换法则分别有何不同。这个暂时的结论又可以进一步启发我们在理论上的思考:比如,亲密熟人和一般熟人虽然都被列在人情交换法则之下,但是二者在网上的互动模式存在明显不同,这些不同互动行为的背后是否有互动法则上的差异?人情交换法则是否可以更进一步进行划分?或是同为人情交换但又强弱有所不同?

另外一个有趣的理论议题是:本研究在问卷设计上以人情交换法则测量第四层潜在熟人,最后得到无法与公平法则下的工具性关系区分开,因此提出猜想,在问卷设计上,如果使用其他法则去定义第四层,是否可以与最外层相区分并使五层模型的相对准确率优于四层模型?那么这个法则又是什么?无疑,一轮轮的大数据挖掘结果与现有理论的对话可以提供新的发现,为改进关系分类理论不断提供洞见。综上,本文的四层分类方式也非最终结论,只是在当前问卷定义下无法区分出最外两层的潜在熟人与认识之人,但如果我们找到在需求法则、人情法则与公平法则之外的第四种法则,或许该法能将第四和第五层区分开来。这需要我们进一步做更多的定性研究,探究这两种关系类型是否存在明显不同的互动模式和不同的信任、互惠程度,这样才能完成理论的修正。

此外,大数据还可以用于开展很多过去十分困难的课题研究。真实的世界是一个复杂的社会系统,计算机技术的发展和社交网络、移动互联网的普遍使用使得这些用户在不同场景、时间、空间维度的个体动作以及个体之间的互动被记录下来。因此,这使得研究个体的动态,个体间的关系和互动,小团体的结构变化,宏观的网络的变化,集体行动的涌现(如重大创新、社会运动、革命爆发等)和复杂社会系统的非常态演化(如金融风暴、景气突转、社会变迁等)(罗家德等,2018)成为可能。大数据下的社会计算学的研究范式为研究这些社会现象提供了一个综合而动态的解释方式,为检验不同变量和因素如何共同作用于现象和行为的产生提供了新的可能。

这种理论导引下大数据与结构化数据结合的研究方法需要跨学科的整合,从而增加了学科之间的对话和结合,打破了学科之间的严格边界。相信未来通过大数据与社会科学间的对话,理论与数据驱动的混合研究方法会用于更多有趣的议题,验证并修正和发展更多的理论,得到更具推广意义的应用。

参考文献:

费孝通,1998,《乡土中国 生育制度》,北京:北京大学出版社。

罗家德、刘济帆、杨鲲昊、傅晓明,2018,《论社会学理论导引的大数据研究——大数据、理论与预测模型的

三角对话》,《社会学研究》第5期。

杨鲲昊,2018,《基于社交应用数据的用户人脉模型》,北京:清华大学硕士学位论文。

Akbas, M. I., R. N. Avula, M. A. Bassiouni & D. Turgut 2013, "Social Network Generation and Friend Ranking Based on Mobile Phone Data." Paper presented at International Conference on Communications. Budapest, June 9.

Athey, S. & G. W. Imbens 2019, "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11(1).

Bloem, J., M. V. Doorn, S. Duivestein, T. V. Manen & E. V. Ommeren 2012, *Creating Clarity with Big Data*. Groningen:Sogeti.

Blumenstock, J. E., G. Cadamuro & R. On 2015, "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350(6264).

Burt, R. S. 1992, *Structural Holes: The Social Structure of Competition*. Cambridge: Harvard University Press.

Dunbar, R. I. 1992, "Neocortex Size as a Constraint on Group Size in Primates." *Journal of Human Evolution* 22(6).

—— 1993, "Coevolution of Neocortical Size, Group Size and Language in Humans." *Behavioral and Brain Sciences* 16(4).

Dunbar, R. I. & M. Spoors 1995, "Social Networks, Support Cliques, and Kinship." *Human Nature* 6(3).

Dunbar, R. I., V. Arnaboldi, M. Conti & A. Passarella 2015, "The Structure of Online Social Networks Mirrors Those in the Offline World." *Social Networks* 43.

Elragal, A. & R. Klischewski 2017, "Theory-driven or Process-driven Prediction? Epistemological Challenges of Big Data Analytics." *Journal of Big Data* 4(1).

Evans, J. A. 2020,"Social Computing Unhinged." *Journal of Social Computing* 1(1).

Evans, J. A. & P. Aceves 2016, "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology* 42 (9).

Galdo, V., Y. Li & M. Rama 2019, "Identifying Urban Areas by Combining Human Judgment and Machine Learning: An Application to India." *Journal of Urban Economics* 103229.

Gao, J., Y. Zhang & T. Zhou 2019, "Computational Socioeconomics." *Physics Reports* 817.

Gao, X., J. D. Luo, K. H. Yang, X. M. Fu., L. C. Liu & W. W. Gu 2020, "Predicting Tie Strength of Chinese *Guanxi* by Using Big Data of Social Networks." *Journal of Social Computing* 1(1).

Ghoddusi, H., G. G. Creamer & N. Rafizadeh 2019, "Machine Learning in Energy Economics and Finance: A Review." *Energy Economics* 81.

Ghose, A. & P. Ipeirotis 2011, "Estimating the Helpfulness and Economic Impact of Product Reviews." *IEEE Transactions on Knowledge and Data Engineering* 23(10).

Granger, C. W. J. 1969, "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica* 37(3).

Granovetter, M. 1973, "The Strength of Weak Ties." *American Journal of Sociology* 78(6).

Gu, S., B. Kelly & D. Xiu 2020, "Empirical Asset Pricing via Machine Learning." *Social Science Electronic Publishing* 9.

Hill, R. & R. I. M. Dunbar 2003, "Social Network Size in Humans." *Human Nature* 14(1).

Hofman, J. M., A. Sharma & D. J. Watts 2017, "Prediction and Explanation in Social Systems." *Science* 355 (6324).

Hwang, K. 1987, "Face and Favor: The Chinese Power Game." *American Journal of Sociology* 92(4).

Lazer, D., A. Pentland, L. Adamic, S. Aral, A. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy & M. V. Alstyne 2009, "Computational Social Science." *Science* 323(5915).

Lazer, D. & J. Radford 2017, "Data Ex Machina: Introduction to Big Data." *Annual Review of Sociology* 43(1).

Lin, N. & E. J. C. Vaughn 1981, "Social Resources and Strength of Ties: Structural Factors in Occupational Status Attainment." *American Sociological Review* 46(4).

Lundberg, S. & S. I. Lee 2017, "A Unified Approach to Interpreting Model Predictions." *Proceedings of the International Conference on Neural Information Processing Systems* (*Nips*). Long Beach, May 22.

Luo, J. D. 2011, "Guanxi Revisited: An Exploratory Study of Familiar Ties in a Chinese Workplace." *Management & Organization Review* 7(2).

Luo, J. D. & K. Yeh 2012, "Neither Collectivism Nor Individualism: Trust in the Chinese *Guanxi* Circle." *Journal of Trust Research* 2(1).

Marsden, P. V. & K. E. Campbell 1984, "Measuring Tie Strength." *Social Forces* 2(2).

Muhamedyev, R., K. Yakunin, Y. Kuchin, A. Symagulov, T. Buldybayev, S. Murzakhmetov & A. Abdurazakov 2020, "The Use of Machine Learning 'Black Boxes' Explanation Systems to Improve the Quality of School Education." *Cogent Engineering* 7(1).

Pollet, T. V., S. G. Roberts & R. I. Dunbar 2011, "Use of Social Network Sites and Instant Messaging Does Not Lead to Increased offline Social Network Size, or to Emotionally Closer Relationships with Offline Network Members." *Cyberpsychology, Behavior, and Social Networking* 14(4).

Runge, J. D. Sejdinovic & S. Flaxman 2017, "Detecting Causal Associations in Large Nonlinear Time Series Datasets." *Science Advances* 5(11).

Sagiroglu, S. & D. Sinanc 2013, "Big Data: A Review." Paper presented at International Conference on Collaboration Technologies and Systems (CTS). San Diego, May 20 – 24.

Seager, W. 1995, "Ground Truth and Virtual Reality: Hacking vs. van Fraassen." *Philosophy of Science* 62(3).

Shi, X., L. A. Adamic & M. J. Strauss 2007, "Networks of Strong Ties." *Physica A* 378(1).

Strumbelj, E. & I. Kononenko 2014, "Explaining Prediction Models and Individual Predictions with Feature Contributions." *Knowledge & Information Systems* 41(3).

Zhou, W., D. Sornette, R. A. Hill & R. I. Dunbar 2005, "Discrete Hierarchical Organization of Social Group Sizes." *arXiv: Statistical Mechanics* 272(1561).

作者单位:清华大学社会学系与公共管理学院(罗家德、高馨)
电子科技大学计算机科学与工程学院、大数据研究中心(周涛)

责任编辑:杨 可