面向社会化媒体用户评论行为的属性推断

刘 云[1] 孙宇清[1),2)] 李明珠[1]

[1)](山东大学计算机科学与技术学院数字媒体教育部工程研究中心 济南 250100)

[2)](山东大学软件学院 济南 250100)

摘 要 针对用户网络行为进行属性推断,在个性化推荐、市场营销和提升平台服务质量等方面具有重要应用价值.现有工作主要针对浏览行为、社交行为等可追踪用户身份的网络行为进行属性推断,而评论性网站用户多为匿名身份,其网络评论行为数据具有碎片化、信息价值含量低和不平衡的特点,且用户群体的属性分布严重不均衡,这些问题给用户属性推断带来挑战.文中引入客体信息、环境信息和语义知识库,辅助用户特征建模,增加了用户评论行为的语义内涵,缓解了用户行为数据量不平衡性和稀疏性问题;基于信息增益度量特征,提出了面向概率性特征选择的两种代表性算法的改进策略:概率包裹式特征选择和启发式概率特征搜索,在解决特征空间高维问题,提高效率的同时,降低了数据噪音影响;提出了面向小比例类型数据的差异性特征选择和迭代式增强学习算法,集成多个特征相关的分类器,既保留了重要特征信息,也给低价值特征提供了小概率选择机会.分别使用真实的中文和英文数据集验证该文方法,包括不同的行为建模方式和特征筛选方法,以及不同参数和用户属性分布不平衡问题对属性推断的影响,并和其他方法进行了对比,实验结果表明该文方法更为有效.

关键词 社会化媒体;属性推断;语义分析;用户行为;概率特征选择;社交网络

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2017.02762

User Attributes Inference Based on Reviews on Social Media

LIU Yun[1] SUN Yu-Qing[1),2)] LI Ming-Zhu[1]

[1)](*School of Computer Science and Technology, Engineering Research Center of Digital Media Technology, Ministry of Education, Shandong University, Jinan* 250100)

[2)](*School of Software, Shandong University, Jinan* 250100)

Abstract The user attribute inference problem occupies an important role in practical applications such as personalized recommendation, marketing and promotion on quality of web service. The current works mainly aim at the identity related user online behaviors, such as a user query history, user relationships etc., which are not applicable for the case on social media since users are often anonymous. Additionally, user reviews are not only fragmented and noisy, but also imbalanced on both the quantity and distribution. In this paper, we propose a series of methods to solve the above challenging problems. We take into account the item information user commented and the context as the supplements for solving the imbalanced problem on quantity distribution, which reveals a user's preference and behavior trajectory. In addition, we introduce an ontology database to enrich inner semantic features of user comments, which summarizes and generalizes the relevant knowledge of words and organizes it into a hierarchical structure. User comments are partitioned into words and mapped to the nodes in the ontology that represent conceptions of the

收稿日期:2016-12-29;在线出版日期:2017-09-13.本课题得到国家自然科学基金(91646119)、山东省重点研发项目(2017GGX10114)、山东省科技发展计划(2014GGX101046)、山东省自主创新及成果转化专项(2014ZZCX03301)和 SAICT 专家项目资助.**刘 云**,女,1989 年生,硕士研究生,主要研究方向为数据挖掘和隐私保护. E-mail: y13583130209@163.com. **孙宇清**(通信作者),女,1967 年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为协同计算与隐私保护. E-mail: sun_yuqing@sdu.edu.cn. **李明珠**,女,1994 年生,硕士研究生,主要研究方向为数据挖掘和隐私保护.

same meaning words. The hierarchical features reveal semantic relationship existed in words and effectively reduce the negative influence of fragmented data and imbalanced quantity problem. The feature dimension is high after modeling and the fragmented information has low value. To solve this problem, we adopt information gain to measure the importance of features. It can be used to measure the influence of the variety of features on user attributes inference result. It reflects the amount of information that a feature contains. In the information theory, the entropy is used to measure the uncertainty of a random variable. For user attributes inference, the uncertainty change of user attributes after adding a feature is called information gain, which indicates the amount of information brought about by this feature. The larger the difference, the more the ability of the feature to distinguish users who have different attributes. In order to reduce the influence by high dimension problem, based on information gain, we improve the two representative methods of probabilistic feature selection: Probability Wrapped Features Selection algorithm and Heuristic Probability Feature Selection algorithm. Both methods adopt feature importance as the probability in feature selection either in pre-classification or iterative learning process. These two methods reduce the search space and improve the convergence rate of feature selection. By taking into account the correlations between features and classifiers on the small scale type data, we proposed the Unbalanced Data Enhancement Learning algorithm to integrate multiple feature-related classifiers. It retains the important features while selects trivial features with low probability. It is more advantageous in the problem of unbalanced attributes inference. Several real datasets are adopted to validate our methods on attribute inference from several aspects, including behavior models, feature selection methods, parameters influence and the degree of imbalanced data on user attributes. The experimental results show that the proposed approach not only relieves the negative influence of fragmented and noisy data, but also effectively solve the difficulty of attribute classification under imbalanced user attribute distribution. The results also show that our methods outperform the related algorithms.

Keywords　social media; attribute inference; semantic analysis; user behavior; probabilistic feature selection; social networks

1　引　言

社会化媒体平台是指为用户提供诸如评论、投票、反馈、分享等功能的在线媒体，像凤凰网等新闻网站、亚马逊和淘宝等电商网站、豆瓣等电影评论网站. 用户网络评论是社会舆论的一种表现形式，具有公开性和可用性特点，群体意见为其他用户在决定购买产品或使用服务时提供了参考，形成一种动态的群体协同环境. 同时，网络服务提供者也希望通过这些信息理解用户行为，分析用户属性分布和偏好，为其产品设计、管理决策和营销活动提供参考，为第三方机构、其他商家或者政府的服务水平提升、公众舆论细粒度评估提供帮助，也为侦查网络犯罪、锁定目标人群提供辅助分析. 这些工作近期也引起了国际学术界的关注，在 AAAI、ICDE 等人工智能和数据挖掘方面的重要国际会议和学术期刊上均有工作讨论用户行为和属性推断[1-5]. 因此，基于社会化媒体用户评论行为进行属性推断具有重要的现实意义.

在社会化媒体平台上，大多数用户出于对隐私安全的考虑，倾向于隐瞒个人信息，而使用匿名身份. 现有针对社交网络用户行为进行属性推断的相关工作不再适用，如 Kosinski 等人根据电话和短信记录并结合 Facebook 的关注信息来推断个人属性[6]；Shane 等人依据社交网络平台上的用户评论、网络结构、用户关系等信息进行属性推断[7-8]，上述工作均是基于社交网络中稳定的用户关系数据的分析方法. 在更为宽泛的社会化媒体平台上，用户之

间没有明确关系,许多学者研究直接利用用户网络行为推断属性[9-10],如 Holbrook 等人研究用户年龄和性别等因素对网络行为偏好的影响,发现娱乐新闻更吸引女性,而男性更喜欢浏览体育新闻[11];Torres 等人通过用户的网页浏览记录如点击行为或者查询日志推断用户的人口统计特征[9-12].这类工作仅从用户访问行为本身进行分析,不适用于具有用户评论数据碎片化、价值含量低等特点的社会化媒体平台.社会化媒体用户通常以匿名身份进行评论,用词较为随意,评论行为数量统计也呈现长尾分布,并且用户属性分布严重不均衡,为属性推断工作带来挑战.

针对上述问题,本文分析了社会化媒体用户评论行为数据的特点,引入客体信息和环境信息,辅助用户特征建模,降低评论行为不均衡性带来的影响;针对评论数据的噪音和碎片化问题,借助分词工具和语义知识库提取语义特征,挖掘用户行为的语义关联关系,并提取了评论的样式特征;针对建模后的用户特征维度大、碎片化数据价值含量低的问题,提出了面向概率性特征选择的两种代表性算法的改进策略:概率包裹式特征选择算法和启发式概率特征搜索算法,提高了学习效率;针对用户属性不均衡问题,提出了面向小比例类型数据的差异性特征选择和迭代式增强学习算法,集成多个特征相关的分类器,既保留了重要特征信息,也给低价值特征提供小概率选择机会,提高了小比例类型数据的分类准确率.

本文第 2 节对现有的相关研究工作进行分析;第 3 节给出问题描述与分析框架;第 4 节讨论用户行为建模方法;第 5 节基于信息增益度量特征重要性,提出两种代表性概率性特征选择的改进策略;第 6 节提出面向不平衡属性分布的增强学习算法;第 7 节在实际数据集合上针对本文方法进行多层面的验证和相关工作对比实验;第 8 节总结全文和讨论未来工作.

2 相关工作

与本文最为相关的工作是基于用户网络评论行为推断用户属性,例如,Otterbacher 等人使用发布在互联网电影数据库(IMDb)中的用户评论数据,分析评论风格、评论内容和相应的补充信息(例如打分),使用逻辑回归模型推断用户性别,结果表明男性和女性在写作风格和内容上存在显著差异[13].Yang 等人运用在线消费者的评论,提出了用户情绪识别主题模型(USTM),结合人口统计信息,探索评论内容和用户属性之间的关系,改进商品推荐[14].上述工作单纯从用户的评论信息出发,没有挖掘用户评论蕴含的内在语义关系.为此,Bergsma 等人和 Ardehaly 等人分析用户行为的语义特征[7,15-16],使用概念类特征来预测用户属性,例如从姓氏和称谓分析女性的婚姻状况.Garera 等人结合社会语言学特征,使用线性支持向量机模型进行属性推断,适用于各种口语会话记录和正式的邮件语料库[17].Rao 等人使用社会语言学的混合特性以及语法模型对 Twitter 用户进行属性推断[18].以上工作虽然使用了评论中的某些特定语义信息,但是没有考虑语义之间的层次关系和碎片化信息中隐藏的含义,不适用于解决社会化网络用户评论行为数据的碎片化和噪音问题.

另一类非常相关的工作是基于用户的搜索或浏览行为进行属性推断.Torres 等人通过学习儿童和青少年的搜索记录和页面点击行为,建立了用户查询行为和用户属性之间的关联关系,依据查询记录进行年龄属性推断,从而限制推荐的页面[9].Hu 等人学习了网页浏览记录和人群属性分布之间的关联模型,识别用户的人口统计属性[2].还有一些工作是根据购买数据推断用户属性,例如,Wang 等人提出一种结构化嵌入表示学习模型,根据零售市场中用户的购买行为数据提取特征,推断人口统计属性[19].然而这类用户行为数据一般具有结构化格式,信息含量较为丰富,而社会化媒体匿名用户的评论用词常常比较随意,带来数据噪音和不均衡问题,上述工作不能完全适用本文属性推断问题.

基于社交网络进行属性推断也和本文工作相关[37],一类工作是基于用户好友关系和网络拓扑结构推断用户特征和兴趣爱好,例如,McPherson 等人和 Dong 等人提出同质性的概念,利用好友关系和部分用户的公开特征推断其他用户属性,揭示了具有好友关系的用户更可能拥有相似兴趣[8,20-21].也有工作根据社区关系进行推断,认为同一社区的用户常常具有相似特性.例如,Mislove 等人发现具有相同用户资料的用户更有可能形成一个紧密的社区[22].Yin 等人利用社团用户的公开数据和成员信息推断目标用户属性[2],Lamba 等人分析同一社交圈内的用户兴趣相似,从而进行个性化推荐[23-25].还

有一类工作考虑了 Twitter 用户的关注行为，例如，Culotta 等人通过收集网站上的网民统计数据（例如浏览 gizmodo.com 网站上的用户 50%有学士学位），结合每个网站的 Twitter 账号的关注者信息，训练回归模型预测一组 Twitter 用户的群体属性统计分布[1]. 这些工作的基础是使用社交网络上的用户关系，但是社会化媒体用户之间并没有明确关系，所以基于社交网络结构进行属性推断的方法不适用于分析社会化媒体平台的用户评论数据.

3　问题描述与分析框架

3.1　问题描述及挑战

针对社会化媒体用户的评论行为进行属性推断，相对于以往工作有两方面困难：数据方面，用户行为数据具有噪音和碎片化的特点，用户行为数量分布也严重不均衡；分析对象方面，在不同类型的网站，用户属性分布差异非常大且不均衡.

首先，社会化媒体用户行为数据具有噪音和碎片化特点. 用户发表评论时用词较为随意，用户评论的长度大多比较短小，据统计用户评论的平均长度只有 31 个字，蕴含价值信息有限，从而为用户行为建模带来困难.

用户行为数据的分布不均衡性主要体现在两个方面，用户评论数量和用户评论长度均具有长尾特征. 如图 1 所示，对爬取的新浪新闻 8000 多个用户在 2007 年 1 月到 2015 年 8 月发表的评论数据进行统计，结果显示用户发表的评论数量呈长尾分布，其中大多数人的评论数量分布在 0～500 条之间，呈现了用户评论数量极不均衡的特性. 用户评论长度方面，根据统计，用户评论最少的只有 3 个字，最多的有四百多个字，体现了用户评论长度的不均衡性. 这些特点为用户建模带来困难，尤其是评论数量小的用户，很难学习到属性分布规律.

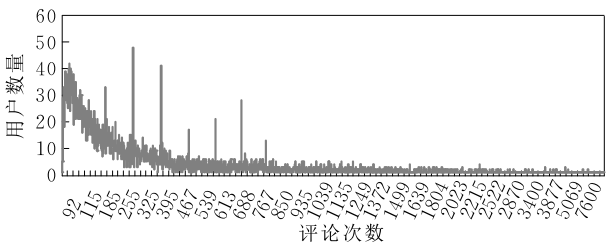


图 1　用户评论分布

再者，不同社会化媒体网站的用户属性差异很大，分布不均衡. 悉尼大学的 Fiona Martin 的研究指出，新闻网站通常更吸引男性用户，如《赫芬顿邮报》网站上 79%的评论是男性留下的，女性只占 20%；《纽约时报》网站上 72%的评论是男性留下的①. 上述研究指出了社会化媒体评论性网站特别是新闻评论网站上用户属性分布不均衡问题. 本文爬取的真实数据集同样也具有属性不均衡的问题，如性别属性中女性样本只占总体样本的 1/12.

3.2　分析框架

针对上述问题，本文提出了融合客体信息、环境信息和语义知识库的用户建模和分析框架，如图 2 所示，具体内容自下而上包括 4 个部分：

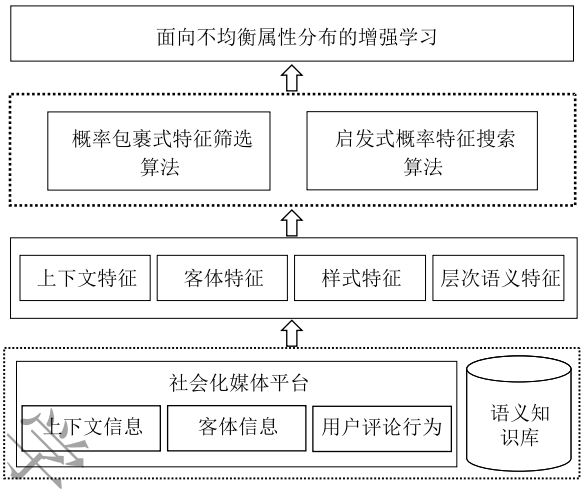


图 2　基于用户网络评论行为的属性推断框架

首先，针对用户行为数据分布不均衡性，引入客体信息和上下文信息作为对评论数量较少用户的行为数据补充，丰富用户行为特征，例如，更多的男性用户关注国际新闻，而关注娱乐新闻的用户女生居多；学生有可能在深夜浏览新闻，而工作人员则更多在上班路上浏览新闻. 针对用户数据碎片化和评论用词随意性问题，本文结合语义知识库分析用户评论的语义信息和行为偏好.

然后，从不同层面抽取用户特征. 例如，从用户行为的时间分布、地理位置、访问通道类型等提取上下文特征；基于用户行为涉及的新闻关键词、标题、类型等提取能够反映用户行为偏好的客体特征；基于评论信息，提取评论长度、表情符号、语气词等能够表明用户书写习惯的外在特征，以及结合语义知识库挖掘用户评论中的隐式特征.

再次，针对用户特征分布的高维和数据噪音的

① http://www.theguardian.com/profile/fiona-martin

问题,基于信息增益度量特征重要性,提出两种代表性概率特征筛选算法的改进策略:概率包裹式特征选择算法和启发式概率特征搜索算法,分别在分类学习前和迭代式学习过程中进行概率特征选择,既保留了重要特征信息,也给低价值特征提供小概率选择机会,筛选密切相关特征,以降低搜索空间,提高收敛速度和学习效果.

最后,针对用户群体属性分布严重不均衡带来的分类学习偏差问题,提出了面向小比例类型数据的迭代式增强学习算法,集成多个特征敏感的分类器,考虑不同特征组合和分类器适用性的同时,提高了小比例类型数据分类准确率.

4 用户行为建模

针对社会化媒体用户评论行为进行属性推断面临的挑战,本文从数据和分析方法两个角度对用户行为数据建模.首先,数据方面,增加和用户评论行为相关的上下文信息、客体信息等数据,作为用户评论行为数据的补充,解决社会化媒体用户行为数量分布不均衡的问题.其次,分析方法方面,在提取用户评论外在特征的同时,结合层次化语义知识库深度挖掘用户评论的内在语义特征,以解决评论数据的噪音、碎片化问题.

4.1 用户评论样式特征

样式特征是基于统计的用户评论的显式特征,包括评论长度、使用表情符号和标点符号情况等.这些特征可以反映用户偏好和书写风格,例如,年轻人喜欢使用表情符号,而中老年人偏好使用标点符号.

4.2 上下文特征

上下文信息是指用户评论行为相关的环境信息,包括评论时间、所在地、用户 IP、访问渠道类型、评论时间和客体发布时间差等信息,可以表示用户对新闻或者话题的关注频率和程度,反映用户行为轨迹和属性特点.例如,上班族可能喜欢在上下班的路上登录网站,而周末时学生在社会化媒体网站上的行为才变得活跃.

4.3 基于客体的用户偏好特征

对不同类型的客体进行评论反映了用户的偏好,例如,男生对国际新闻和国际局势更为关注,就会对相关信息进行点赞或是留言;而女生对美容化妆内容更感兴趣.所以,借助客体内容丰富用户行为特征,能够解决用户行为信息不足的问题,特别是评论行为少的用户建模.本文借助用户关注的客体内容、客体类型、发布时间、发布单位、新闻事件发生地和用户所在地之间的关联关系等特征辅助用户行为建模.

4.4 基于知识库的行为语义特征

为了提取用户评论数据更为丰富的语义内涵,本文引入语义知识库,将用户评论内容映射到知识库上,抽取可理解的层次化语义特征,解决用户评论数据信息含量低、有噪声和数据碎片化的问题.

语义知识库是把词语的相关知识进行总结和概括,并依照一定的思想和结构组织成的一种便于理解的层次知识库[26],包括上下位关系、同反义关系、整体与部分等关系,在句法分析、分词系统、词义消歧等方面有成功应用.语义知识库中的每个义原节点表示一个基本的词汇概念,概念之间的包含关系构成语义层次结构,不同层次的结点对应的语义粒度不同,层数越低,语义描述越详细.本文选用 HowNet 和 WordNet 分别进行中英文数据集的用户语义特征分析.

基于知识库的用户评论建模包括评论分词、基于知识库的词汇映射、层次化用户描述三个步骤.中文评论分词采用 ICTCLAS 汉语词法分析系统,英文评论基于空格进行分词;然后查找每个单词对应的语义层次结构中的义原节点;针对每个用户评论数据集,统计其对应的语义层次结构的节点匹配次数.

现有工作主要采用两种方法提取用户的层次化语义特征,一种是均衡节点权重原则[27],就是直接将所有语义层次节点的统计分布记为用户的行为向量;另一种是统一层次描述原则[28],把用户评论映射的节点和节点对应的匹配次数投影到选定的语义层次上,选择的层次越高,语义越抽象和宏观,可以降低特征维度;而选择的层次越低,粒度越细致.

上述方法没有考虑语义路径,不能分析评论数据的多层面关联含义.本文提出了用户语义特征表示的改进方案,增加考虑用户评论词语在语义树中的匹配路径和层次关系,作为用户评论的语义特征向量.这种方法的代价是语义特征维度高,针对这一问题下一节将基于信息增益度量特征重要性,并作为特征选择的概率,目的在于挖掘内在语义特征并提高学习效率.

上述建模方式产生的特征数量共有两千多维,表 1 列出了部分关键特征,包括样式特征、层次化语义树特征、上下文特征、客体特征等.

表 1　部分关键特征说明

特征分类	特征	定义
样式特征	表情	用户各类表情符号的使用频率.
	语气词	用户在不同语气词类别的使用频率.
	符号	标点符号、空格、回车的使用频率.
	评论长度	用户发表评论的平均长度.
	评论次数	用户回复/评论的条数.
	评论评价	发表的评论被点赞、被踩的次数.
上下文特征	评论时间	用户发表评论的时间.
	用户所在地	用户所在地区.
	用户类型	区分是客户端浏览，还是网页浏览.
	时间分布	评论样式在时间上的分布.
客体特征	信息类型	如发布新闻的类型.
	发布时间	信息发布的时间.
	发布单位	信息是由那个单位发布，如新华社.
	标题	用户浏览信息的标题.
语义树特征	用户评论	结合语义知识库的层次化语义特征.

5　基于特征重要性的概率选择

用户建模过程中，引入上下文特征、客体关联特征和语义特征等信息虽然可以消除评论数据的噪音和碎片化的影响，但是会导致特征维度非常大，用户行为的不平衡性也使得特征数据稀疏，为了降低由此带来的学习代价、过度拟合等问题，需要对高维特征降维，增加分类准确性.

现有特征选择有两种代表性策略[29-30]，一是独立于分类学习的包裹式特征选择，从候选特征集合中随机选择子集，根据其分类学习结果判断特征子集的优劣，这种选择策略具有较大的不确定性，若初始特征很多，门限不变的前提下，迭代次数将会快速增多. 二是融合于分类学习过程的启发式特征搜索策略，将特征子集的评价和学习过程相结合，这种策略容易陷入局部最优解. 为此，本文引入基于信息增益的特征重要性概念，并转化为特征选择的概率，改进上述两种代表性算法的特征选择策略，提出了概率包式特征选算法和启发式概率特征搜索算法.

5.1　定　义

给定数据集 $D=(U,V,B)$，其中，用户集合 $U=\{u_1,u_2,\cdots,u_m\}$，m 表示总的用户数量，V 是客体集合，B 是用户在客体上的评论行为. 用户特征包括样式特征、上下文特征、客体特征和语义树特征，特征集合表示为 $F=\{f_1,f_2,\cdots,f_n\}$，n 表示特征数量，用户 $u_i\in U$ 的行为向量为 $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{in})\in R^n$. 所有用户的特征向量矩阵为 $\boldsymbol{X}=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_m)^{\mathrm{T}}\in R^{m\times n}$. 学习目标是依据用户的评论行为，推断用户在该属性上的类别，采用 $C=\{C_1,C_2,\cdots,C_\alpha\}$ 表示用户在该属性上的标签集合，$\alpha\in N^+$ 是类别数量，则推断目标为 $\boldsymbol{Y}^{\mathrm{T}}\in C^m$，即推断用户 $u_i\in U$ 的属性类别 $y_j\in C$. 例如，性别属性，$\alpha=2$，$C=\{F,M\}$.

5.2　特征重要性度量

特征重要性是指不同特征取值对于属性推断结果置信度的影响，反映了使用该特征对于属性分类学习带来的信息含量. 信息论中采用熵度量随机变量的不确定性，而增加一个特征后用户属性这一随机变量不确定性的变化即为信息增益，表示这个特征带来的信息量，这个差值越大，反映了该特征区分不同属性用户的能力越强. 本文基于信息增益度量特征重要性.

对于给定的用户集合 U，属性取值为 C_i 的用户比例为 $P(C_i)$，信息熵 $H(C)$ 表示用户属性取值的不确定性，$H(C)=-\sum_{i=1}^{h}P(C_i)\log_2 P(C_i)$.

针对特征 $f_j\in F$，设 $F_j=\{f_{j1},f_{j2},\cdots,f_{j|F_j|}\}$ 表示特征 f_j 的有界离散取值空间. U_j^t 表示取值空间为 $f_{jt}\in F_j$ 的用户集合，$|U_j^t|$ 为该集合用户数量，$P(C_i|f_{jt})$ 表示特征为 f_{jt} 的用户集合中属性为 C_i 的用户所占比例，条件熵 $H(C|f_{jt})$ 表示在特征取值为 f_{jt} 的条件下用户属性推断的不确定性，如式(1)所示. 因此，特征 f_j 整体上对于属性推断带来的不确定性度量为 $H(C|f_j)$，如式(2)所示.

$$H(C|f_{jt})=-\sum_{i=1}^{h}P(C_i|f_{jt})\log P(C_i|f_{jt}) \quad (1)$$

$$H(C|f_j)=\sum_{t=1}^{|F_j|}\frac{|U_j^t|}{m}H(C|f_{jt}) \quad (2)$$

条件熵 $H(C|f_j)$ 反映了在已知特征 f_j 的情况下，用户属性取值的不确定性. 相比于原有属性推断的不确定性，$H(C)$ 与 $H(C|f_j)$ 的差值越大，说明特征 f_j 整体上对于属性推断带来的信息量越大，即 f_j 带来的信息增益 $g_j=H(C)-H(C|f_j)$.

由此得到所有特征的信息增益集合 $IG=\{g_1,g_2,\cdots,g_n\}$，其中 n 为特征总数. 信息增益越大，特征对于属性推断就越重要，分类学习过程中就应更多考虑. 为此，本文基于特征的信息增益进行概率特征选择，重要性越大的特征被选择的概率越大，特征 f_j 被选择到的概率为 $p_j=\dfrac{g_j}{\sum_{i=1}^{n}g_i}$. 以此提出了面向概率性特征选择的两种代表性算法的改进策略：概率包裹式特征选择和启发式概率特征搜索. 既保留

了重要特征,又使得价值含量小的特征也有被选择到的机会,避免了因使用单一门限,遗漏相关特征的问题,提高了学习效率.

5.3 基于信息增益的概率包裹式特征筛选

基于信息增益的概率包裹式特征选择算法(Probability Wrapped Features Selection,PWFS),整体思路如下:首先,依据特征选择概率,选择包含 N_k 个特征的特征集合 S_k,并从候选分类器中如逻辑回归、随机森林等,选择一个分类器 h_k. 然后,通过分类结果 M_k 评价 S_k 和 h_k 组合,评价指标 M_k 根据不同的需要,可以选取精度、准确率、召回率等. 最后,通过 K 轮上述工作的迭代选择和评价,选择其中 T 个分类效果好的特征集合和分类器的组合 $\{(S_t,h_t)\}_{t=1}^{T}$. 输出的 T 个特征集合和分类器的组合将用于第 6 节面向不均衡分类样本的增强学习. 具体算法如下.

算法 1(PWFS). 概率包裹式特征选择算法.

输入:数据集合$\{\boldsymbol{X},\boldsymbol{Y}\}$,特征集合 F,特征选择概率集合 $\{p_1,\cdots,p_n\}$,迭代选择次数 K,选择特征集合数 T,候选分类器集合 H

输出:T 个特征子集和对应的分类器组合的集合 $\{(S_t,h_t)\mid S_t\subset F,h_t\in H\}$

1. $SP_0=0$;
2. FOR $j=1:n$
3. $SP_j=SP_{j-1}+p_j$ //Set the threshold SP_j
4. ENDFOR
5. FOR $k=1:K$ //K times iteration
6. $N_k=rand(1,n)$;
7. FOR $i=1:N_k$ //Select feature f_j
8. $r=rand(0,1)$;
9. IF($SP_{j-1}<r\leq SP_j$) $S_k=S_k\cup\{f_j\}$
10. ENDFOR
11. Select one classifier h_k from H;
12. Train $h_k(\boldsymbol{X},\boldsymbol{Y},S_k)$;
13. $M_k=$test $h_k(\boldsymbol{X},\boldsymbol{Y},S_k)$;
14. $\{(S_t,h_t)\}_{t=1}^{T}=Top_T\{(S_k,h_k)\mid k=1..K\}$;
//Select and Update Top_T map of classifiers and feature sets based on M_k
15. ENDFOR
16. return $\{(S_t,h_t)\}_{t=1}^{T}$;

上述算法的整体时间复杂度为 $O(Kn)$,n 为全部特征的个数,其中分类器学习过程的时间复杂度独立考虑,没有计算在内. 迭代次数 K 的选择直接影响特征选择的效果和选择效率等竞争性指标,K 取值小则迭代次数少,但选取的分类器和特征集合的质量受限,考虑到需要筛选 T 个分类器,K 至少取值为 T;如果迭代次数 K 大则效率受影响,但是考虑到特征选择集合的空间为 2^n 的规模,K 的取值还是远小于 2^n,如实验中特征数量 $n>2000$,而迭代次数达到 100 时分类结果就能趋于稳定,效率提升非常显著. 所以本文提出的基于信息增益的特征筛选算法可以有效地在多项式时间内筛选特征. 概率包裹式特征选择算法能够得到基于学习器的最优特征子集,解决由于随机选择特征子集带来的不确定性以及迭代次数过大的问题.

5.4 基于启发式概率特征搜索算法

另一类特征选择方法采用了融合于分类学习过程的启发式特征搜索策略,其代表性算法为粒子群算法,该算法将特征子集评价和优化目标相结合,并用多个搜索粒子同时搜索多个特征子集,根据搜索粒子之间的交流搜索目标[31]. 现有工作有许多对粒子群算法的改进,如广义粒子群优化算法[32]、简化粒子群优化算法[33]等,分别解决离散及组合优化问题和使用更小的种群数和进化次数获得更好的优化效果. 本文在上述策略基础上,提出了基于启发式概率特征搜索算法(Heuristic Probabilistic Feature Selection algorithm,HPFS),使用基于重要性的特征概率初始化多个搜索粒子的特征子集,并在学习过程中把搜索粒子速度归一化,概率性地选择特征子集.

基于上述思想,算法分为两部分,首先基于特征重要性概率初始化搜索粒子的值,然后在迭代过程中将搜索粒子的速度归一化,概率性地筛选特征,并通过粒子群体协作和信息共享来寻找最优解. 在 n 维特征空间中,a 个粒子组成一个搜索群,每个粒子为二进制的空间,其中第 i 个粒子在空间的位置表示为一个 n 维向量,$\boldsymbol{b}_i=(b_{i1},b_{i2},\cdots,b_{in})^{\mathrm{T}}\in\{0,1\}^n$. 每个粒子的位置就是一个潜在的特征筛选的解,基于选到的特征子集 $\boldsymbol{b}_i$ 使用分类器对相应的样本集进行分类,得到的分类准确率作为衡量 $\boldsymbol{b}_i$ 优劣的依据. 并且第 i 个粒子迄今为止搜索到的最优特征组合记为 $\boldsymbol{p}_i=(p_{i1},p_{i2},\cdots,p_{in})^{\mathrm{T}}\in\{0,1\}^n$. 整个粒子群体发现的全局最优特征组合记为 $\boldsymbol{p}_g=(p_{g1},p_{g2},\cdots,p_{gn})^{\mathrm{T}}\in\{0,1\}^n$. 第 i 个粒子的搜索速度也是一个 n 维的向量,记为 $\boldsymbol{v}_i=(v_{i1},v_{i2},\cdots,v_{in})^{\mathrm{T}}\in[0,1]^n$. 第 $l+1$ 轮粒子位置速度更新公式如下所示:

$$v_{id}^{l+1}=v_{id}^{l}+\psi(p_{id}^{l}-b_{id}^{l})+\psi(p_{gd}^{l}-b_{id}^{l}) \tag{3}$$

$$b_{id}^{l+1}=\begin{cases}1, & rand()<sig(v_{id})\\ 0, & \text{其他}\end{cases} \tag{4}$$

其中，$d=1,2,\cdots,n$；$i=1,2,\cdots,a$，a 为种群规模，ψ 是常数，称为学习因子，l 为迭代次数，$rand()$ 函数随机产生取值区间为[0,1]的随机数，

$$sig(v_{id})=\frac{1}{(1+\exp(-v_{id}))}.$$

根据上述迭代公式和适应值对粒子的最优特征集合和全局最优集合进行更新，得到筛选的特征子集和对应的分类器. 这种基于启发式概率特征搜索算法是一个全局搜索，通过在迭代过程中概率性地选择特征可以防止陷入局部最优解.

6　用户属性推断学习

6.1　基于分类学习的属性推断

用户数据表示为$\{\boldsymbol{x}_i,\boldsymbol{y}_i\}_{i=1}^{m}$，$m$ 为用户数量，用户属性推断问题可以描述为以下形式，考虑额外隐藏变量即噪声 ε，希望找到分类函数 $\hat{\boldsymbol{y}}=f(\boldsymbol{x},\varepsilon)$，使得 $\boldsymbol{y}$ 和 $\hat{\boldsymbol{y}}$ 之间的差异最小. 用户属性取值有两种情况：一种为枚举类型，如性别、种族. 另一种属性取值属于有界的离散区间，例如年龄，需要对整个离散区间按照用户属性分布情况和实际需求分成多个区间. 对于选定的多个属性区间，每段属性的推断任务被视为一个单独的推断问题，目标函数是这些任务的总和. 本文级联多个分类模型组合成为一个多分类模型，用于预测具有多个分类的用户属性. 参数 $\boldsymbol{\theta}$ 为一个矩阵，矩阵的行数对应分类的类别数量，每一行元素为一个参数向量，向量的大小和特征维度相同. 多分类的公式如下：

$$p(y_{ij}=C_j\mid\boldsymbol{x}_i;\boldsymbol{\theta})=\frac{\exp(\boldsymbol{\theta}_j\boldsymbol{x}_i^{\mathrm{T}})}{\sum_{l=1}^{\alpha}\exp(\boldsymbol{\theta}_l\boldsymbol{x}_i^{\mathrm{T}})}\tag{5}$$

$$J(\boldsymbol{\theta})=-\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{\alpha}1\{y_{ij}=C_j\}\log\frac{\exp(\boldsymbol{\theta}_j\boldsymbol{x}_i^{\mathrm{T}})}{\sum_{l=1}^{\alpha}\exp(\boldsymbol{\theta}_l\boldsymbol{x}_i^{\mathrm{T}})}\right]+\lambda\|\boldsymbol{\theta}\|_2^2\tag{6}$$

其中 C_j 表示用户属性类别，$j\in\{1,\cdots,\alpha\}$，α 为类别属性数量，式(5)表示预测的属性等于 C_j 时的概率. 式(6)前一项中 $1\{\cdot\}$ 是一个指示性函数，即当大括号中的值为真时，该函数的结果就为 1，否则其结果就为 0；后一项为规则项，参数 λ 为规则项的系数，$\lambda>0$，加入规则项的目的是使得损失函数变为严格的凸函数，得到唯一的解并可以惩罚过大的参数值. 使用梯度下降法最小化损失函数，最终得到多分类器.

6.2　面向不均衡属性分布的增强学习

解决样本属性分布严重不均衡性问题，通常有两种方法，一是有条件采样策略. 有条件采样是通过改变训练数据的样本分布来消除或减小数据的不平衡性[34]，但是此方法容易造成信息缺失或者过拟合问题. 二是不平衡数据增强学习集成策略[35]，现有工作中或是特征选择过程和集成学习相互独立[36]，当训练数据由于交叉验证而改变时，所选择的结果可能会遗失相关特征；或是独立讨论分类器在每一个特征上的显著度[35]，忽略了特征之间的相互影响. 为此，本文集成多个特征相关的分类器，综合考虑分类器在不同特征群组上的分类表现，能够兼顾不同特征之间的相互影响和分类器在小比例类型数据上的分类效果，在不均衡属性推断问题方面更具优势.

基于上述改进策略，本文提出了面向小比例类型数据的不平衡数据增强学习算法(Unbalanced Data Enhancement Learning，UDEL)，算法包含两部分，首先根据算法 1 中优选的特征集合和分类器的组合$\{(S_t,h_t)\}$，观察对小比例类型数据分类的表现，使用 $C_{\min}$ 表示小比例类型数据标签值，根据小比例类型数据正确分类的情况，更新分类器权重 w_t. 然后集成和特征相关的多个分类器，所有的集成分类器分类结果加权之和大于集成分类器的平均权重时，则把当前样本分为小比例类型数据，保证更多的小比例类型数据被正确识别. 对于多分类的情况，如果存在用户属性不平衡情况，可以迭代选择出小比例类型数据，分类结果取决于 T 个分类器最终的投票结果.

算法 2(UDEL).　不平衡数据增强学习算法.

输入：数据集合$\{\boldsymbol{X},\boldsymbol{Y}\}$，分类器集成数量 T，T 个特征子集和对应的分类器的组合$\{(S_t,h_t)\}$

输出：分类结果 $f(\boldsymbol{X})$

```
FOR t=1:T
  Predict_label=h_t(S_t,X,Y);
    FOR j=1:|X|
      IF (y(j)==C_min)&&(Predict_label(j)==
          y(j))
        w_t=w_t+1;
    ENDFOR
ENDFOR
```

8. $w_t=\dfrac{w_t}{\sum_{t=1}^{T}w_t}$;

9. $f(\boldsymbol{x})=\begin{cases}1, & \left[\sum_{t=1}^{T}w_t * h_t(\boldsymbol{x})\right]\geq\dfrac{1}{2}\\ 0, & \text{其他}\end{cases}$

10. return $f(\boldsymbol{X})$;

不平衡数据增强学习算法注重更容易分错的小比例类型数据的学习,通过集成多个特征相关的分类器,并对小比例类型数据分类效果好的分类器加以更大的权重,从而使集成分类器不仅全局性能更优,而且小比例类型数据的学习误差减少.

7 实 验

7.1 数据集和预处理

为了验证本文方法对中文和英文数据的适用性,分别验证了一个中文数据集和两个英文数据集.中文数据集为新闻评论数据集,我们使用 WebMagic 爬虫工具爬取了新浪新闻从 2007 年 1 月到 2015 年 8 月发表的 500 万条评论和对应的新闻信息,共计一万余人.考虑到需要对用户属性进行验证,本文根据用户对应的新浪微博账号个人信息得到属性真实值,去掉噪音数据后,共有 7562 个用户,其中 7483 个含有性别信息的用户,1887 个含有年龄信息的用户.男性发表的评论数量最多近 3 万条,最少 30 条,平均 600 多条.女性发表的评论最多 1 万多条,最少近 50 条,平均近 700 条.平均发表的评论数量为 661 条.基于新浪新闻数据集提取了四种不同类型的特征:样式特征、上下文特征、客体相关特征、层次化语义树特征.四类特征集合中,样式特征包含 198 维,上下文特征包含 242 维,客体相关特征包含 219 维,语义树特征 1584 维.

鉴于用户年龄属于有界的离散区间,所以依据对用户的采样结果得到对年龄的划分.图 3 描述了数据集中不同年龄的人发表的评论数量,其中,横轴为用户的年龄,纵轴为发布的评论数量.根据用户年龄分布,将年龄分为 4 个区间:0～23、24～37、38～49、50 岁以上.

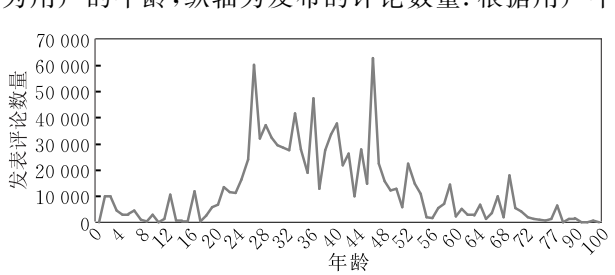


图 3 评论数量在年龄上的分布

英文数据集有两个,一个为"Data for Everyone Library"的公开数据,为 2015 年 10 月 Twitter 用户发表的评论数据集,包括近 2 万个用户发表的 2 万多条评论.数据集包括用户名、随机的一条微博、账户资料、图片、位置等信息.共 6022 个男性用户,6521 个女性用户,5800 个商家用户.对数据进行预处理,仅保留了具有性别信息的用户,共 1 万 2 千多用户.并基于 WordNet 从评论中提取了用户的层次语义特征,共包含 8 层,每一层的特征维度为 1 万维左右.

另一个英文数据集为 Reddit 发布的公开数据,Reddit 是一个社会化媒体新闻站点,用户能够浏览或者提交帖子,其他用户可以对发表的帖子投票和评论.本文使用的数据集为 2015 年 5 月 Reddit 用户发表的评论数据集,共有五千多万条评论.数据集包括用户名、评论点赞数量、评论主题所在子版块、用户属性信息、当前评论的父评论等.经过数据预处理,得到了共 10 296 名具有有效性别的用户,其中男性 6143 名,女性 4153 名,共 143 220 条评论.

7.2 性能度量指标

对于给定的用户集 $U=\{(x_1,y_1),\cdots,(x_m,y_m)\}$,可将样例根据其真实类别和学习器预测类别的组合划分为四种情形,TP 指预测为正例并且真实也为正例即为真正例,FP、TN、FN 分别表示假正例、真反例、假反例.本文采用分类学习常用的度量标准评价算法的属性推断结果,具体如下:

精度定义为每类分类正确的样本数占样本总数的比例,$accuracy(U)=\dfrac{1}{m}\sum_{i=1}^{m}|f(x_i)=y_i|$.准确率定义为 $P=\dfrac{TP}{TP+FP}$,表示在所有预测为正例的样本中分类正确的比例.召回率定义为 $R=\dfrac{TP}{TP+FN}$,表示在所有正例样本中分类正确的比例.为了综合考虑准确率和召回率,本文采用了 $F1$ 度量,$F1=\dfrac{2\times P\times R}{P+R}$.

对于不平衡数据集合,数据比例严重失衡时,如果所有的测试集都被分为大比例类型数据,用精度、准确率、召回率评价得到的结果也可以很好.例如男女比例为 12∶1,精度也可以达到 0.92.所以单纯使用精度、准确率、召回率这些静态指标评价分类器性能是不合适的.为此,对于类别不平衡数据使用

FPR 和 *TPR* 分别作为横轴和纵轴的 *ROC* 测量标准则更为合适. 本文采用了 *ROC* 曲线下面积 *AUC*. 其中，*FPR*：假阳性率，即 $FP/(FP+TN)$，*TPR*：真阳性率，即 $TP/(TP+FN)$. 例如，当男性为正类女性为负类时，所有女性被错分为男性那一部分的比例则为 *FPR*，所有男性被正确分类的那一部分的比例则为 *TPR*.

7.3　实验结果

7.3.1　属性推断方法对比

针对中英文共三个数据集，比较本文方法和三类代表性分类算法 KNN、随机森林和逻辑回归，其中数据集合中男女比例为 3∶1. 分析结果如表 2 所示，在新浪数据集上，本文方法 UDEL 整体分类结果都最好，在 Twitter 数据集上，分类结果整体表现也最好，其中 KNN 的召回率高，是因为其对大比例类型数据的预测都正确，但是 KNN 的 *AUC* 结果并不高说明小比例类型数据的预测结果并不好. 对于 Reddit 数据集，UDEL 分类结果整体最好，*AUC* 值远大于其余的对比方法，其中随机森林的召回率高，但是其余的评价指标远没有 UDEL 好，说明其只是对大比例类型数据的分类结果较好. 所以在比例失衡背景下，本文提出的方法总体表现都最好.

表 2　不同方法对属性推断性能对比

数据集	方法	准确率	召回率	*F*1	*AUC*
新浪新闻	KNN	0.83	0.77	0.80	0.55
	RF	0.86	0.93	0.89	0.60
	Logistic	0.84	0.60	0.70	0.55
	UDEL	**0.87**	**0.96**	**0.91**	**0.64**
Twitter	KNN	0.70	**1.00**	0.82	0.50
	RF	0.71	0.84	0.77	0.52
	Logistic	0.70	0.76	0.73	0.50
	UDEL	**0.75**	0.93	**0.83**	**0.54**
Reddit	KNN	0.74	0.81	0.77	0.54
	RF	0.73	**0.99**	0.84	0.52
	Logistic	0.73	0.53	0.61	0.51
	UDEL	**0.82**	0.88	**0.85**	**0.68**

7.3.2　用户行为建模分析

为了验证本文提出的用户行为建模是否能够解决用户行为数据的噪音、碎片化和分布不均衡问题，本节分析融合了上下文信息、客体相关信息、评论样式信息和语义知识库之后的特征对分类结果的影响. 采用从新浪新闻数据集中提取的四种类型特征推断用户性别和年龄. 表 3 是 4 种类型特征之间的属性推断结果对比，UL 表示样式特征，UC 表示上下文特征，UR 表示客体特征，UT 表示层次语义特征. 从中我们可以看出对于性别预测，上下文特征和语义特征的准确率和 *AUC* 值相对较高. 对于年龄预测，样式特征和上下文特征更能区别用户的年龄. 表 4 为 4 种类型特征相结合的属性推断结果对比，显示语义树特征和样式特征相结合的性别分类和年龄分类的效果都最好，其次为语义树特征和上下文特征的结合. 语义树特征和其余特征结合的分类效果有明显提升. 对于年龄预测除了语义树特征和样式特征、语义树特征和上下文特征结合预测精度较高外，还有语义树特征、样式特征和上下文特征三者相结合的效果较好. 说明针对用户评论噪音和碎片化问题，本文引入上下文信息、客体相关信息和语义知识库，相对于没有引入语义知识库和环境信息的情况来说可以有效地提升属性分类的效果.

表 3 不同类型特征的属性推断结果对比

特征类型	性别					年龄
	精度	准确率	召回率	*F*1	*AUC*	精度
UL	0.79	0.81	**0.92**	**0.86**	0.62	0.77
UC	0.76	**0.83**	0.85	0.84	0.67	0.74
UR	0.71	0.79	0.82	0.81	0.60	0.74
UT	**0.80**	**0.83**	0.90	**0.86**	**0.69**	**0.78**

表 4　不同类型特征相结合的属性推断结果对比

特征类型	性别					年龄
	精度	准确率	召回率	*F*1	*AUC*	精度
UL+UR	0.70	0.68	**0.89**	**0.77**	0.65	0.74
UL+UT	**0.72**	0.75	0.75	0.75	**0.72**	**0.78**
UC+UT	0.65	**0.83**	0.53	0.65	0.68	**0.78**
UR+UT	0.65	0.71	0.72	0.71	0.63	0.77
UL+UT+UC	0.64	0.72	0.68	0.70	0.65	**0.78**

另外，对比了基于知识库的 3 种用户评论建模方案：均衡节点权重原则、统一层次描述原则、全语义路径原则. 如表 5 所示，其中统一层次描述中语义树层次共有 12 层，第 7 层实验结果较好，所以表 5 中基于统一层次原则的实验结果使用的是映射到第 7 层的语义特征. 由表 5 可以看出本文使用的基于全语义路径原则的方案整体表现较好. 因为保留语义路径可以挖掘出用户评论行为潜在的语义关系，获得更多的有价值信息.

表 5　基于知识库的用户评论建模方法对比

方案	精度	准确率	召回率	*F*1	*AUC*
均衡节点权重	0.68	0.83	0.59	0.69	0.69
统一层次	0.65	0.78	0.58	0.66	0.66
全语义路径	**0.71**	**0.86**	**0.62**	**0.72**	**0.72**

7.3.3　特征筛选算法对比及参数分析

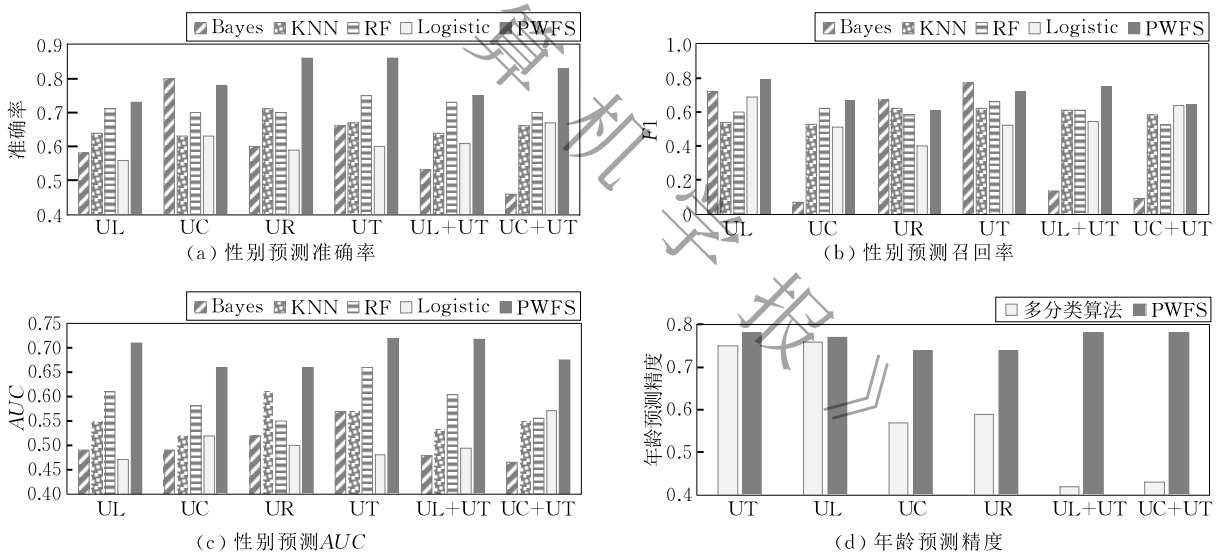
本节包括三部分：首先是不同特征算法对比；其次是使用了特征选择算法和没有使用特征选择算法的对比；最后讨论了不同迭代次数对结果的影响.

第一部分对概率包裹式特征筛选算法(PWFS)和随机包裹式特征选择算法(LVW)进行对比,对基于启发式概率特征搜索算法(HPFS)和启发式的粒子群算法(PSO)进行对比.实验选取了新浪新闻数据集中男女比例为1∶1的层次语义特征用于算法对比,分别根据PWFS和HPFS两个算法达到稳定时的迭代次数100和150进行实验.由表6可知,针对独立于分类学习的包裹式特征选择策略,PWFS在整体上的表现都比LVW要好.针对融合于分类学习过程的启发式搜索策略,本文改进后的HPFS算法相比于PSO算法来说总体的表现也较好.从两类代表性策略的对比来看,PWFS在精度、准确率的表现要好,而HPFS的召回率和$F1$值较高,说明HPFS把男性误判为女性的数量比把女性误判为男性的数量少,有可能是因为男性的评论行为数量比女性多造成的.

表6 特征选择算法对比

代表性策略	算法	精度	准确率	召回率	$F1$
包裹式	PWFS	**0.71**	**0.86**	**0.62**	**0.72**
	LVW	0.68	0.76	0.60	0.67
启发式	HPFS	**0.68**	**0.77**	**0.75**	**0.81**
	PSO	0.64	0.72	0.56	0.63

第二部分的实验对比使用PWFS算法和没有使用特征筛选算法在不同特征组合上对分类结果的影响.使用新浪新闻数据集,其中男女比例为1∶1.特征选择使用逻辑回归分类器,实验结果如图4所示,和相关工作中使用的评论结合逻辑回归方法[13]对比,使用了特征筛选算法之后的分类结果有明显提升.并且和具有特征筛选功能的随机森林算法对比,分类结果也较好.对于不同的特征类型组合,本文的算法对于性别预测的准确率和AUC都平均提升了约10%,在年龄预测方面,对于使用了上下文特征和客体关联特征的预测精度也有很大提升.说明了本文提出的特征筛选算法,可以有效选取有价值的信息,降低数据噪音的影响.



(a)性别预测准确率 (b)性别预测召回率 (c)性别预测AUC (d)年龄预测精度

图4 算法分类效果度量

最后基于新浪新闻数据集,分析特征筛选中参数的选取对于属性推断结果的影响.结果如图5所示,K表示特征筛选算法迭代次数,随着K的增加,PWFS和HPFS整体性能成上升趋势,当K大于一定门限时,分类结果趋于稳定.如前面对于算法PWFS的复杂度分析,与指数规模的特征空间相比,迭代次数非常少,实验表明特征数量>2000时,迭代次数接近100次,分类结果就能趋于稳定达到最优,说明本文提出的特征筛选算法非常高效.其中,PWFS算法的收敛速度整体较快,并且稳定后的

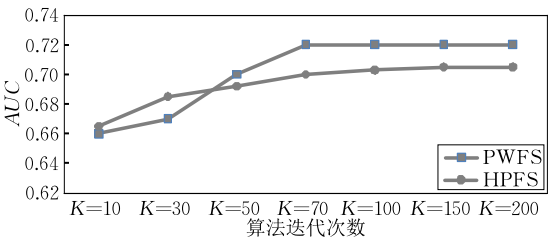


图5 不同迭代次数下的特征筛选算法度量

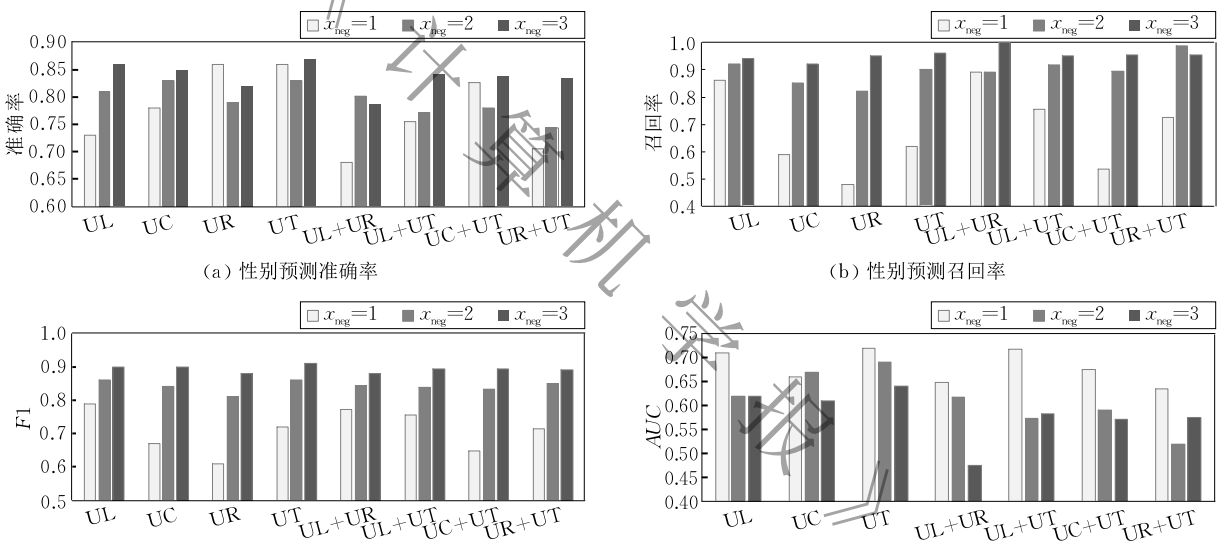
AUC值比HPFS算法要好.但是,HPFS算法在迭代次数K等于10次到30次之间时其AUC要比PWFS算法好.可能是因为HPFS算法中搜索粒子

之间可以根据当前的局部最优解和全局最优解进行相互交流，所以在迭代初期可以快速逼近搜索目标，而到了迭代后期，由于学习因子等超参数的设置使得粒子速度有偏差最终导致搜索到的最优解在搜索目标周围震荡.

7.3.4　不平衡数据处理方法对比及参数分析

为了解决样本比例不均衡的问题，本节首先进行有条件采样，选取属性比例不同的样本进行实验对比；其次，对于不同比例的样本，讨论不平衡数据增强学习中级联分类器数量对结果的影响. 最后，观察不平衡数据增强学习算法在不同比例样本下的分类表现.

对于有条件采样方法的对比，使用中文数据集，参数 x_{neg} 是指男性样本和女性样本的比例，取值为 $x_{neg}\in\{1,2,3\}$，x_{neg} 值越大表示用户属性越不平衡. 从图 6 可以看出，随着 x_{neg} 取值逐渐增加，分类的准确率、召回率和 $F1$ 都有明显增加，但是 AUC 值却呈减少的趋势. 说明随着女性样本的减少女性被正确分类的比例也随之下降. 通过不同特征集合的表现，我们发现随着样本逐渐的不平衡语义树特征和其余特征的两两结合有较为稳定的表现. 根据男女样本比例不同我们可以发现，对于不同的样本比例，在 $x_{neg}=3$ 时，准确率和 $F1$ 最高，在 $x_{neg}=1$ 时 AUC 最高. 上述有条件采样的方法可以用于样本数量很大的情况，但如果样本数量有限则会导致有条件采样得到的学习样本不足以得到好的预测结果. 所以本文又提出了集成多个特征相关分类器的方法：不平衡数据增强学习，使得在样本比例差异很大的情况之下，也能取得较好的分类性能.

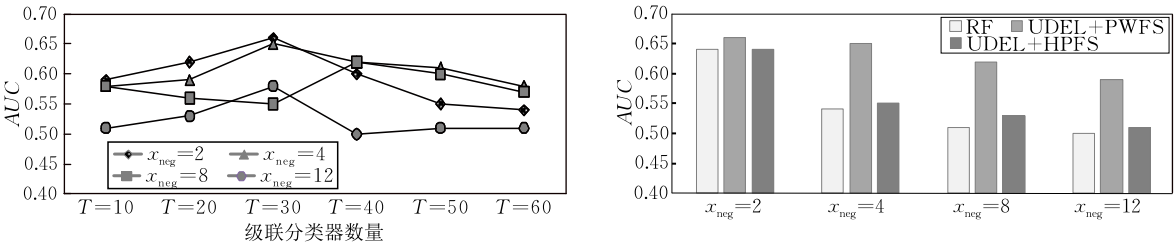


(a) 性别预测准确率　(b) 性别预测召回率

(c) 性别预测$F1$值　(d) 性别预测AUC

图 6　基于有条件采样属性推断与分析

对于不平衡数据增强学习的算法实验我们选取了男女样本比例 $x_{neg}\in\{2,4,8,12\}$，分类器级联的个数 $T\in\{10,20,30,40,50,60\}$. 从图 7(a)可以看出，随着分类器级联数量的增加，AUC 呈逐渐减少的趋势，这是因为分类器数量越多，后面的分类器在小比例类型数据上的表现并没有很好，所以影响了整体的性能. AUC 在 $x_{neg}=2,4,12$ 时级联 30 个分类器可以取得最好的性能. $x_{neg}=8$ 时，级联 40 个分类器 AUC 达到最好. 为了进一步验证两种特征筛选算法对于不平衡数据学习的适用性，图 7(b)对结合 PWFS 和 HPFS 两种特征筛选算法的不平衡数据增强学习算法与 7.3.1 节中表现最好的具有特征选择功能的对比算法随机森林(RF)的 AUC 结果进行对比. UDEL 使用了逻辑回归进行分类，在不同比例类型样本下，随着男女样本比例的增大，整体上 AUC 乘下降的趋势，但是结合不同特征筛选算法的不平衡学习要比随机森林性能要好，并且不平衡数据增强学习算法下降的趋势最慢. 其中，结合 PWFS 算法的不平衡数据增强学习要比结合 HPFS 算法的 AUC 值要高，根据 7.3.3 节实验可知是由于 PWFS 算法得到的 AUC 值比 HPFS 高导致的. 综上，本文提出的不平衡数据增强学习的算法在数据比例失衡时可以达到较好的性能.



(a) 使用不同数量级联分类器对不同比例样本分析 (b) 不同算法结果对比

图 7 不平衡数据增强学习差异分析

7.3.5 用户行为不均衡性对分类结果的影响

由于评论性网站上用户评论数量分布不均衡，本文探索了用户评论数量和预测结果之间的关系，把用户评论数量分为 0～50，50～100，100～200，200～500，500～1000，1000 以上 6 个区间，然后分析每个区间上用户属性预测性能，如表 7 所示，从中可以看出随着评论数目的增多，分类的精度和 *AUC* 逐渐增加，在 500～1000 时达到最高. 结合用户数量分析，虽然在评论数量为 200～500 之间时用户数量最多，但是评论数目比 500～1000 的评论数量要少，所以在样本数量较多的情况下信息不足导致预测性能下降. 在评论数目高于 1000 时，用户数量比 500～1000 区间的数量要多但是有可能会因为评论数目过多导致噪声增加，也会影响预测精度.

表 7 不同用户评论数和预测性能之间的关系

评论数	用户数	精度	*AUC*
0～50	52	0.25	0.50
50～100	176	0.52	0.49
100～200	270	0.56	0.52
200～500	**309**	0.59	0.59
500～1000	180	**0.65**	**0.61**
>1000	207	0.61	0.54

8 总 结

针对社会化媒体用户网络行为进行属性推断，在市场营销和个性化推荐等方面具有重要应用价值，较之现有的基于社交网络或交易行为的属性推断问题更具挑战性. 本文主要贡献为：引入客体信息、环境信息和语义知识库，辅助用户特征建模，为碎片化的用户评论行为增加了语义内涵，降低了用户行为数据量不平衡性和稀疏性带来的困难性；基于信息增益度量特征重要性，提出了面向概率性特征选择的两种代表性算法的改进策略：概率包裹式特征选择和启发式概率特征搜索，在解决特征空间高维问题、提高效率的同时，降低了数据噪音的影响；提出了面向小比例类型数据的差异性特征选择和迭代式增强学习算法，集成多个特征相关的分类器，既保留了重要特征信息，也给低价值特征提供小概率选择机会. 最后，选择中文和英文共三个真实数据集进行实验，结果表明本文方法比现有方法更优.

未来工作将探索多平台用户行为关联分析和属性推断，更好地理解用户行为和提供个性化服务. 鉴于不同类型的网站定位有差异，用户在不同平台上的行为也有很大不同，以及数据来源复杂性等问题，将为属性推断带来更多挑战. 我们将尝试平台依赖的领域知识学习，进行跨平台用户建模和属性关联性分析，增加情感因素等. 不平衡数据分类问题是另一个未来工作，将尝试引入不同的学习规则如代价敏感方法、目标函数优化方法等，提高属性推断效果.

参 考 文 献

[1] Culotta A, Kumar N R, Cutler J. Predicting the demographics of Twitter users from website traffic data//Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI 2015). Austin, USA, 2015: 72-78

[2] Yin Hongzhi, Hu Zhiting, Zhou Xiaofang, et al. Discovering interpretable geo-social communities for user behavior prediction //Proceedings of the 32nd IEEE International Conference on Data Engineering (ICDE 2016). Helsinki, Finland, 2016: 942-953

[3] Mao Jia-Xin, Liu Yi-Qun, Zhang Min, et al. Social influence analysis for micro-blog user based on user behavior. Chinese Journal of Computers, 2014, 37(4): 791-800(in Chinese)
(毛佳昕，刘奕群，张敏等. 基于用户行为的微博用户社会影响力分析. 计算机学报，2014，37(4)：791-800)

[4] Gao Quan-Li, Gao Lin, Yang Jian-Feng, et al. A preference elicitation based on users' cognitive behavior for context-aware recommender system. Chinese Journal of Computers, 2015, 38(9): 1767-1776(in Chinese)
(高全力，高岭，杨建峰等. 上下文感知推荐系统中基于用户认知行为的偏好获取方法. 计算机学报，2015，38(9)：1767-1776)

[5] Zafarani R, Liu Huan. Connecting users across social media sites: A behavioral-modeling approach//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2013). New York, USA, 2013: 41-49

[6] Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 2013, 110(15): 5802-5805

[7] Bergsma S, van Durme B. Using conceptual class attributes to characterize social media users//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, 2013: 710-720

[8] Fang Quan, Sang Jitao, Xu Changsheng, et al. Relational user attribute inference in social media. IEEE Transactions on Multimedia, 2015, 17(7): 1031-1044

[9] Hu J, Zeng H J, Li H, et al. Demographic prediction based on user's browsing behavior//Proceedings of the 16th International Conference on World Wide Web (WWW 2007). New York, USA, 2007: 151-160

[10] Bi B, Shokouhi M. Inferring the demographics of search users: Social data meets search queries//Proceedings of the 22nd International Conference on World Wide Web (WWW 2013). Seoul, Korea, 2013: 131-140

[11] Holbrook M B, Schindler R M. Age, sex and attitude toward the past as predictors of consumers' aesthetic tastes for cultural products. Journal of Marketing Research, 1994, 31(3): 412-422

[12] Torres S D, Weber I. What and how children search on the web//Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM 2011). New York, USA, 2011: 393-402

[13] Otterbacher J. Inferring gender of movie reviewers: Exploiting writing style, content and metadata//Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM 2010). New York, USA, 2010: 369-378

[14] Yang Z, Kotov A, Mohan A, et al. Parametric and non-parametric user-aware sentiment topic models//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2015). New York, USA, 2015: 413-422

[15] Burger J D, Henderson J, Kim G, et al. Discriminating gender on Twitter//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011). Edinburgh, UK, 2011: 1301-1309

[16] Ardehaly E M, Culotta A. Inferring latent attributes of Twitter users with label regularization//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologie (NAACL HLT 2015). Denver, USA, 2015: 185-195

[17] Garera N, Yarowsky D. Modeling latent biographic attributes in conversational genres//Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP. Singapore, 2009: 710-718

[18] Rao D, Yarowsky D, Shreevats A, et al. Classifying latent user attributes in Twitter//Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents (SMUC 2010). New York, USA, 2010: 37-44

[19] Wang Pengfei, Guo Jiafeng, LanYanyan, et al. Your cart tells you: Inferring demographic attributes from purchase data//Proceedings of the 9th ACM International Conference on Web Search and Data Mining (WSDM2016). San Francisco, USA, 2016: 173-182

[20] McPherson M, Lovin L S, Cook J M. Birds of a feather: Homophily in social networks. Cook Annual Review of Sociology. Review Sociology, 2001, 27(1): 415-444

[21] Dong Yuxiao, Yang Yang, Tang Jie, et al. Inferring user demographics and social strategies in mobile social networks//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2014). New York, USA, 2014: 15-24

[22] Mislove A, Viswanath B, Gummadi K P, et al. You are who you know: Inferring user profiles in online social networks//Proceedings of the T3rd ACM International Conference on Web Search and Data Mining (WSDM 2010). New York, USA 2010: 251-260

[23] Lamba H, Narayanam R. Circle based community detection //Proceedings of the 5th IBM Collaborative Academia Research Exchange Workshop. New York, USA, 2013: 1-4

[24] Qian Xueming, Feng He, Zhao Guoshuai, et al. Personalized recommendation combining user interest and social circle. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(7): 1763-1777

[25] Yang Xiwang, Steck H, Liu Yong. Circle-based recommendation in online social networks//Proceedings of the 18th International Conference on Knowledge Discovery and Data Mining (KDD 2012). Beijing, China, 2012: 1267-1275

[26] Prasojo R E, Kacimi M, Nutt W. Entity and aspect extraction for organizing news comments//Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM 2015). Melbourne, Australia, 2015: 233-242

[27] Liu Peng-Yuan, Zhao Tie-Jun. Unsupervised translation disambiguation by using semantic dictionary and mining language model from Web. Journal of Software, 2009, 20(5): 1292-1300(in Chinese)
(刘鹏远，赵铁军. 利用语义词典 Web 挖掘语言模型的无指导译文消歧. 软件学报，2009，20(5)：1292-1300)

[28] Xu Haoran, Sun Yuqing. Identify user variants based on user behavior on social media//Proceedings of the IEEE 34th International Performance Computing and Communications Conference (IPCCC'03). Nanjing, China, 2015: 1-8

[29] Liu H, Motoda H, Setinono R, et al. Feature selection: An ever evolving frontier in data mining//Proceedings of the 4th Workshop on Feature Selection in Data Mining (FSDM). Hyderabad, India, 2010: 4-13

[30] Sun Z H, Bebis G, Miller R. Object detection using feature subset selection. Pattern Recognition, 2004, 37(11): 2165-2176

[31] Kennedy J, Eberhart R C. A discrete binary version of the particle swarm algorithm//Proceedings of the 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation. Orlando, USA, 1997: 4104-4108

[32] Gao Hai-Bing, Zhou Chi, Gao Liang. General paarticle swarm optimization model. Chinese Journal of Computers, 2005, 28(12): 1980-1987(in Chinese)
(高海兵，周驰，高亮. 广义粒子群优化模型. 计算机学报，2005, 28(12): 1980-1987)

[33] Hu Wang, Li Zhi-Shu. A simpler and more effective particle swarm optimization algorithm. Journal of Software, 2007, 18(4): 861-868(in Chinese)
(胡旺，李志蜀. 一种更简化而高效的粒子群优化算法. 软件学报，2007, 18(4): 861-868)

[34] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357

[35] Viola P, Jones M J. Robust real-time face detection. International Journal of Computer Vision, 2004, 57(2): 137-154

[36] Tan A C, Gilbert D. Ensemble machine learning on gene expression data for cancer classification. Applied Bioinformatics, 2003, 2(3 Suppl): S75

[20] Pennacchiotti M, Popescu A M. Democrats, republicans and star bucks afficionados: User classification in Twitter//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2011). New York, USA, 2011: 430-438

LIU Yun, born in 1989, M. S. candidate. Her research interests include data mining and privacy protection.

SUN Yu-Qing, born in 1967, Ph. D., professor. Her research interests include collaborative computing and privacy protection.

LI Ming-Zhu, born in 1994, M. S. candidate. Her research interests include data mining and privacy protection.

Background

User attribute inference can help understand different aspects of social network and benefit related analysis, such as the principle of social value allocation, personalized recommendation, the quality of service etc. The current works mainly aim at the identity related user online behaviors, such as user query records, user relationships etc., which are not applicable for the case on social media since users are often anonymous. Additionally, user review behaviors on social media are not only fragmented and noisy, but also imbalanced on the quantity and distribution. This paper focuses on the user attribute inference problem based on user behavior on social media. We take into account user behavior related item information and context as the supplements, and introduce an ontology database to enrich inner semantic features. We adopt information gain to measure the importance of features, based on which we improve the two representative methods of probabilistic feature selection: Probability Wrapped Features Selection algorithm and Heuristic Probability Feature Selection algorithm. We proposed the Unbalanced Data Enhancement Learning algorithm to integrate multiple feature-related classifiers. The experimental results show that our methods outperform the related algorithms.

This work is partially supported by the National Natural Science Foundation of China under Grant No. 91646119, which is about Individual Value Recognition and Predication Based on Multiple Social Media Data. This program analyzes user behaviors with various probabilistic methods on character tags, social role and community, as well as intrinsic characteristics besides the traditional explicit attributes such as sex or age etc. This work helps it thoroughly understand individual value and solves the key problem of the program.

This work is also partially supported by other programs as, the Key Research and Development Program of Shandong Province under Grant No. 2017GGX10114, the Science and Technology Development Plan of Shandong Province under Grant No. 2014GGX101046, the Independent Innovation and Achievement Transformation Special Project of Shandong Province under Grant No. 2014ZZCX03301 and the SAICT Expert Program.