DOI:10.13718/j.cnki.xsxb.2018.12.019

社交网络群中用户活跃度分析与预测①

张效尉[1]，　余云霞[2]，　王　伟[1]

1. 周口师范学院 网络工程学院，河南 周口 466001；
2. 荆楚理工学院 计算机工程学院，湖北 荆门 448000

摘要：对于社交网络中不同的群组，由于用户属性(性别、年龄等)、群类别、群成员之间关系等因素的影响，其活跃度各不相同. 本文首先从社交网络用户数据中提取人口信息、群的类别、社交关系、群用户黏性(分享消息数、图片数)等特征，然后利用 logistic 回归、支持向量机、BP 神经网络等机器学习算法对不同群中用户的活跃度进行预测. 结果表明，BP 神经网络针对社交网络群中用户活跃度分类判断时具有较高的预测性能，社交关系特征对群用户活跃性具有重要影响.

关　键　词：社交网络群；用户活跃度；人口信息学；社交关系

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1000－5471(2018)12－0115－07

随着 Web 2.0 的发展，各种社交网站和应用软件，比如 Facebook、人人网、QQ、微信、Flickr 等得到广泛的应用，促进相关企业获得了巨大的经济和社会效益. 人们利用社交软件发表观点和生活动态、互加好友维持社交关系、多个成员组成群组等方式将线下人际交往搬到互联网线上，形成各种社交网络[1]. 由多个用户聚集在一起形成的小型群组在社交网络中起重要作用，它既是用户获取信息的重要来源，又是用户实现社群生活的重要平台. 用户可以查看群内发送的文字、声音、图片等内容，也可在群内认识新朋友，讨论话题，找到组织认同感，满足人们的交往和精神需求[2].

社交网络群刚被创建时，新成员不断加入，群内发言、发送图片与视频的数量较多，有些群用户活跃度较高，成员之间能持续找到相互交流的话题，愿意通过互动实现娱乐和增强友谊，这些群能在很长一段时间内保持活跃并存在；而有些群活跃度较低，群在创建一段时间后，成员趋于稳定，互动交流较少，群进入沉寂状态[3]. 针对群内用户活跃度存在差异的问题，当前研究成果大多由从事统计、经济与社会管理的学者利用用户问卷调查的数据定性地开展研究[4－6]. 文献[7]根据用户加入群的行为具有规律性，构建了双向马尔科夫随机场模型，并对用户加入群的行为进行了预测；文献[8]利用生存理论模型分析用户在 3 个知识分享社区中的活跃周期和参与时间；文献[9]分析了人口信息学、用户社交关系等特征对群内用户活跃度判断的重要影响. 本文将以实际的社交网络数据为基础，分析影响社交网络中群内用户活跃度的因素，构建模型判断群的活跃度，为人们理解群活跃度差异的原因，制定提高社交网络群中用户活跃度的策略提供依据.

① 收稿日期：2017－12－21
基金项目：国家自然科学基金项目(U1504602)；河南省科技攻关项目(172102210089，162102210396)；河南省高等学校重点科研项目(17A520019，15A520114，16A520107).
作者简介：张效尉(1982－)，男，讲师，硕士，主要从事数据挖掘及社交网络等方面的研究.

1 问题定义

根据社交网络中群组的特征，群内用户活跃度分析与预测问题的定义如下：对于社交网络群 $G=\{V, E\}$，其中 V 为用户的集合，E 为用户之间社交关系的集合，在已知群内用户人口学信息、社交关系、采样时间内发布图片数和消息数等数据的情况下，选取一段时间内群中用户平均发布消息的数量作为衡量群用户活跃度的指标，将群内用户活跃度问题定义为2种分类任务：①二分类问题，首先计算单个群内用户发送消息数的总和除以当前群内用户成员数，得到该群内成员发送消息数量的平均值，然后将整个数据集中所有群组计算得出的平均值求和除以群组的数量，以此获得数据集中群组发送消息数量的平均值，将其作为区分点(数据集上采样时间内所有群组发送消息数量的平均值为146条)，假如一个群内成员平均发表消息的数量大于146，将该群标记为活跃群，否则为不活跃群，以此实现对群活跃度的标记，通过该分类任务找出适合预测群内用户活跃度的机器学习算法，同时判断各类特征影响群内用户活跃度的程度；②多分类问题，在该分类任务中预测群内用户发送消息数量的平均值，因为预测其精确值难度较大，且根据实验结果在一段时间范围内的用户活跃度大致相似，所以改为预测值的范围，将值小于20的作为第0类，值在20～300之间的作为第1类，大于300的作为第2类，以此定义多分类问题，进一步验证二分类中的机器学习算法在多分类问题中的预测效果，比较2种分类方法的优缺点，用以找到提高群内用户活跃度预测算法性能的策略. 针对2类社交网络群用户活跃度预测的问题，利用标记好的数据，分析造成和影响不同群内用户发放消息数量显著差别的原因，构建用户活跃度的判别模型，给出群内用户在5个月采样时间内是否活跃的分类判断.

2 社交网络群用户活跃度特征分析

社交网络群用户活跃度的评价可以从多个方面来分析，如随着时间的增加是否有新成员的加入、用户发布消息、图片和视频的数量等，本文从某社交网站用户群数据中，选取在一段时间内群中用户平均发布消息的数量情况作为衡量群用户活跃度的指标，分析和提取造成群内用户平均发布消息数量差别的影响特征.

2.1 社交网络群用户黏性特征

社交网络群在一段时间内的用户活跃度是由群内成员数、消息数、图片数等体现的，其中，群内成员数量多代表该群有吸引力，有更多用户不断加入，组织庞大；群内消息和图片发布数量多代表群内成员有共同话题和交流的愿望，用户活跃度高，愿意通过相互分享和交流维持群的活力.

2.2 群类别特征

社交网络群创建后类别和用途不同，本文按照群承担的社交功能划分为不同群，将群名中包含“家”字的群归为“家庭群”，这些群也有可能是自称为一家人的群；群名中包含“班”或“同学”字的群归为“同学群”；群名中包含“精英”、“管理”、“公司”等字的群划分为“工作群”；将其它群作为用户为社交需要建立的群，全部归为“社交群”. 群的类别和用途不同，决定了群员的身份，相互维持群内关系的愿望，讨论话题和发放消息数目的差异，这是一个决定群内消息发送量的重要特征.

2.3 人口信息学特征

人作为社交网络的组成单元，其自身的特性，如性别、年龄和地域等属性对人的行为与思维方式有深远影响，人口信息学可以作为统计变量分析社交网络群中用户的活跃度. 社交网络群内男、女人数的比例对群的活跃度有直接影响，文献[10]从人口统计特征得出，相比于女性，男性在社交网站上分享新鲜事的频率更低. 社交网络群内的成员大多年龄相仿，可以用群内成员的平均年龄作为群的年龄属性，而不同年龄段的人对群活跃性影响的程度不同，如年轻人容易接受新事物，这些年龄段人多的群，用户活跃性较高，

而平均年龄较大的群受限于用户成员的精力和爱好，活跃度相对较低. MAGNUSON 等[11]研究人员发现，18～22 岁的年轻人对于社交网站的使用显得异常活跃. 由此可见群内成员的性别比例和平均年龄可以作为影响群用户活跃度的特征.

在评价地域对社交网络群用户活跃度的影响时，本文统计群中用户来自不同省份的人员数量，将用户人数最多的省份作为群所属的省，比如一个群中湖北的人数最多，则将这个群的地域属性归为湖北. 中国省份多，并且临近的省份人们生活习惯和经济条件差别不大，临近省份作为地域属性对社交网络群活跃度影响的程度应该相近，因此对地域属性进行重新概念分层，划分成华北、东北、西北、西南、东南沿海等 5 个区域.

2.4　社交关系特征

社交网络是由人与人基于内在交往需要相互联系构成的，人们在互联网上的行为，如互加好友的数量、加入群的数目、相互之间私信联系情况等可以判断用户在社交网络中的活跃情况. 社交网络群内成员之间愿意互加好友，表明群内成员交往意愿强烈，渴望相互建立关系，增加社交范围；用户加入群数量多，代表该用户愿意加入更多的组织，结交新朋友；群内成员之间，通过相互私信交流信息和想法，以及互发图片实现娱乐，以此维护友谊.

社交网络群内用户基于好友关系建立社交网络，其中每个用户成员为一个节点，2 个用户建立了好友关系则两者之间存在一条无向边，群内成员构成一个由多个节点和边组成的网络图. 针对社交网络群成员社交网络图，需考虑群密度、群平均聚类系数、群直径等对用户活跃度的影响，其中群密度越大，表示群内成员更加倾向于互加好友，边的数量多；群聚类系数大，表示群内好友更倾向于聚集，结构更紧密；群直径越大，说明群内成员疏远程度大. 经过分析计算，发现群成员的群密度平均为 0.616 1，群平均聚类系数为 0.718 7，群直径平均为 2.695 6，三者作为特征对群用户活跃度影响较大.

综上所述，本文以社交网络群内用户消息数量作为指标衡量群活跃度，共提取了群属性、人口信息学、社交关系等 14 个特征，具体特征内容如表 1 所示.

表 1　影响社交网络群用户活跃度的特征

特征序号	特征类别	特征名称	特征序号	特征类别	特征名称
1	群属性	成员数	8	社交关系	成员互加好友数
2	群属性	图片数	9	社交关系	成员加入群数
3	群属性	视频数	10	社交关系	群内好友平均私信值
4	群属性	群类别	11	社交关系	群内好友平均互发图片数
5	人口信息学	性别	12	社交关系	群密度
6	人口信息学	年龄	13	社交关系	聚类系数
7	人口信息学	地域	14	社交关系	群直径

3　实验结果与分析

3.1　数据描述与预处理

数据集为从 2016 年 10 月 1 日到 2017 年 2 月 29 日共 5 个月时间内，某社交网站数据中有过一次或以上消息发布的群中随机采样出的 10 万个群样本. 同时，提取在 2017 年 2 月 29 日还留在群内的 150 多万个成员的基本属性信息，以及这些成员在群内发放消息和图片等情况、不同成员之间的好友关系列表. 最后，形成的数据集如表 2 所示. 针对数据集中缺失的信息，采用多种方法填充，如成员地域信息“省份”为缺失值，可以采用“unknown”来填充；群

表 2　数据集

类　　别	数　　量
社交网络群	100 000
群成员属性信息	1 517 856
群成员活跃情况信息	1 517 856
用户好友关系	87 864 352

名称缺失，可以采用“null”值来填充.

在预测过程中，首先，从数据集中提取影响群用户活跃度的14个特征和类别标记；接着，采用最小—最大标准化方法对特征值进行归一化处理；最后，将数据集划分成训练集和测试集.

3.2 预测方法

本文采用监督学习算法中的Logistic回归、支持向量机、BP神经网络等方法对社交网络群活跃度进行预测. 在预测二分类问题时，Logistic回归和BP神经网络算法得出的预测值在[0－1]之间，取0.5为活跃度分界值，将大于0.5的归为活跃群，小于0.5的归为不活跃群；支持向量机算法得出的预测值为1或－1，将值为1的归为活跃群，否则为不活跃群. 在采用上述机器学习算法预测多分类问题时，BP神经网络可以直接进行多分类判断，但是Logistic回归和支持向量机只支持二分类判断，两者要处理多分类问题，就需要构造合适的多类分类器，下面详细介绍实验中Logistic回归和支持向量机构造多分类器的实现方法.

在实验过程中，Logistic回归处理多分类问题时，直接将每个类别都建立一个二分类器，假如给定的数据集 $X\in R^{m\times n}$，它们的标记为 $Y\in R^{K}$，即这些样本有 k 个不同的类别，具体实现过程如下：①挑选出标记为$c(c\leqslant k)$的样本，将挑选出来的带有标记 c 的样本的标记置为1，剩下的不带标记 c 的样本标记置为0；②用这些标记好的数据训练出一个分类器，得到函数 $h_c(x)$(表示针对标记 c 的logistic回归分类函数)；③重复上面的步骤，针对 k 个类别可以得到 k 个不同的logistic回归分类器；④针对一个测试样本，当采用上述 k 个logistic回归分类器预测标记时，需要找到这 k 个分类函数输出值最大者，即为测试样本的标记.

利用支持向量机构造多分类器时，采用间接法中的一对一法，该方法避免了采用直接法实现时计算复杂度高，算法难于设计，只适合解决小型问题而不适用于处理海量大数据的缺点，同时解决了一对多间接法实现时预测结果偏好于数据量大的类别易造成不实用的问题[12]. 一对一法实现时，在任意两类样本之间设计一个支持向量机模型，因此 k 个类别的样本就需要设计 $k(k-1)/2$ 个支持向量机模型. 在实验中，采用Libsvm中的一对一间接多类分类法，假设有0,1,2三类. 在训练的时候分别选择0和1,0和2,1和2所对应的向量作为训练集，得到3个支持向量机模型. 当对一个未知样本进行分类时，首先把未知样本对应的向量放入3个支持向量机模型中得到3个结果，然后针对这3个结果进行投票，最后得票最多的类别即为该未知样本的类别.

3.3 评测指标

为了评价社交网络群用户活跃度预测模型的效果，对于二分类预测问题，采用信息检索的标准评价指标对群用户活跃度的预测效果进行评估，包含查全率、准确率、F1度量等. 针对多分类问题，给出不同机器学习算法在每一类用户活跃度级别，即平均发布消息数量范围预测上的准确率. 本文还分析了各类特征对预测结果的影响程度，评价指标采用ROC曲线，ROC曲线又称“受试者工作特征”曲线，ROC曲线以“真正例率”为纵坐标，以“假正例率”为横坐标，预测方法的ROC曲线越靠近左上角表明该方法预测效果越好.

3.4 社交网络群用户活跃度预测结果

下面将介绍不同机器学习算法对社交网络群用户活跃度的预测结果，分析不同特征对社交网络群用户活跃度预测的重要程度. 由实验结果可以看出，本文提取的特征与构建的模型在预测社交网络群用户活跃度方面具有较高的性能.

3.4.1 社交网络群用户活跃度预测

表3显示了针对二分类问题，不同机器学习算法对社交网络群用户活跃度的预测结果. 由实验结果可以看出，相对于Logistic回归和支持向量机(SVM)，BP神经网络在判断群内用户活跃度时效果较好(如F1度量的综合评价标准较高).

表 3　社交网络群用户活跃度二分类预测结果

学习方法	准确率	查全率	F1 度量
Logistic 回归	0.842 4	0.646 5	0.731 5
支持向量机	0.629 1	0.702 3	0.663 7
BP 神经网络	0.681 2	0.944 1	0.791 4

表 4 显示了针对多分类问题，不同机器学习算法对每一类用户活跃度级别的预测结果. 从实验结果可以看出，针对社交网络群中用户活跃度多分类时，BP 神经网络相对于 Logistic 回归和支持向量机(SVM)在第 0 类与第 2 类的预测时都取得了较好的效果，而第 1 类问题预测时效果好于支持向量机，但差于 Logistic 回归，从总体上考虑，BP 神经网络更适合于用户活跃度多分类的判断. 3 种机器学习算法在第 0 类和第 1 类问题预测时效果都明显好于第 2 类问题的预测结果，原因在于第 2 类社交网络群数据中作为用户活跃度指标的用户平均发送消息数量值的变化范围较大，从值为 300 开始，个别群的值为 2000 以上，这是 3 类机器学习算法预测效果都偏低的主要原因；而第 0 类群值在 0～20 之间，第 1 类群值在 20～200 之间，这两类群的值较为集中，上述算法都取得了较好的预测效果，同时因第 1 类群在现实中更为常见，数据中占比较大，各种算法预测时都取得了比第 0 类更好的效果.

表 4　社交网络群用户活跃度多分类预测结果

类级别准确率	Logistic 回归	支持向量机	BP 神经网络
0	0.66	0.69	0.70
1	0.82	0.71	0.72
2	0.28	0.36	0.52

3.4.2　不同特征对社交网络群活跃度的影响程度分析

文中采用了群属性特征(群用户黏性特征和群类别特征联合构成群属性特征)、人口信息学特征和社交关系特征等对社交网络群的用户活跃度进行预测，然而现实中每一种特征对群用户的影响程度大小是不相同的，需将不同特征对用户活跃度影响结果进行预测与比较，以此判断哪类特征对群的用户活跃性影响最大. 表 5 显示了针对二分类预测问题，在 Logistic 回归模型中，不同特征对群中用户活跃度的预测结果，为更好地提高群用户活跃度提供了依据.

表 5　不同特征对群用户活跃度影响的预测结果

特征集合	准确率	查全率	F1 度量
所有特征	0.842 4	0.646 5	0.731 5
－社交关系特征	0.730 7	0.176 7	0.284 6
－人口信息学特征	0.833 3	0.627 9	0.716 1
－群属性特征	0.784 5	0.660 4	0.717 1
随机猜测	0.500 0	0.500 0	0.500 0
＋社交关系特征	0.755 3	0.660 4	0.704 7
＋人口信息学特征	0.428 5	0.014 8	0.028 7
＋群属性特征	0.716 4	0.223 2	0.340 4

从表 5 可以看出，针对二分类问题进行预测时，考虑采用所有特征训练出来的分类器，具有较高的 F1 度量值，综合评价效果较好，删掉任一特征，效果会相应变差，由此说明每一类特征对社交网络群用户活跃度的预测都有影响，考虑所有特征的预测模型具有最佳效果. 图 1 显示了群属性特征、人口信息学特征和社交关系特征对群用户活跃度预测结果影响的 ROC 曲线. 实验结果表明，在以上 3 种特征中，对群用户活跃度影响最大的是社交关系特征，人口信息学特征影响较小. 从表 5 和图 1 的 ROC 曲线中可以看出，相对于人口信息学和群属性特征，社交关系特征对群中用户的活跃性影响较为显著，去掉社交关系特征后，分类器的 F1 度量值从 0.731 5 下降到 0.284 6，下降幅度明显；而只采用社交关系特征预测时，F1 度量值达到 0.704 7，略低于包含所有特征的 0.731 5；这充分说明了社交关系特征的重要性，社交网络群用户以

一定的交往目的加入群，相互之间满足社交需要，这与社交网络作为一个社交平台承载社会交往和信息交流功能的认识是一致的.

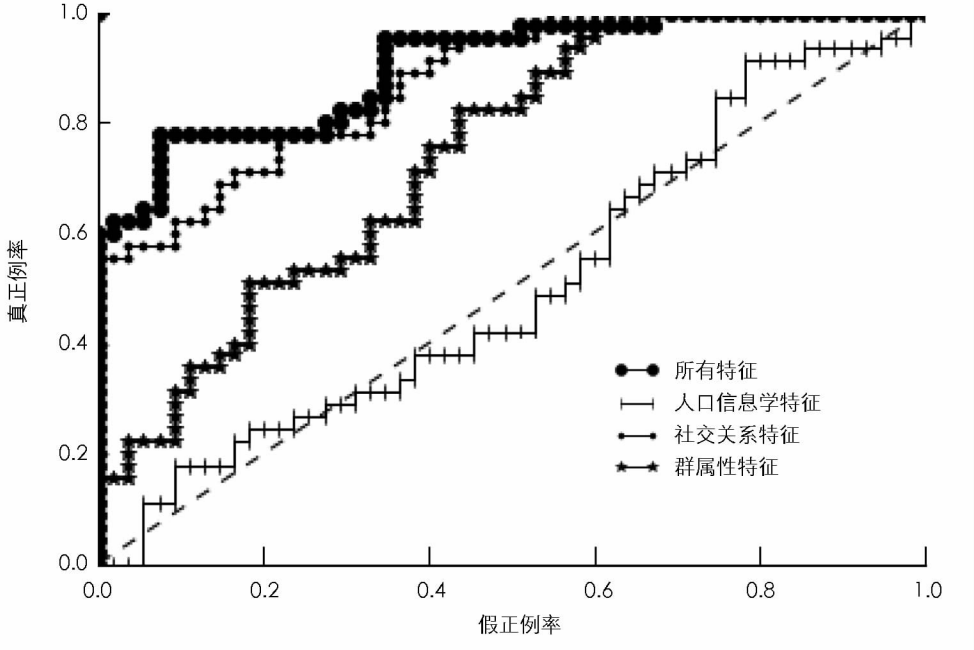

图1 群用户活跃度预测不同特征的ROC曲线

4 结 论

为了预测社交网络群的用户活跃度，从互联网上抓取海量群数据，提取影响用户活跃度的特征，采用Logistic回归、支持向量机、BP神经网络等机器学习算法构建用户活跃度分类判断模型. 实验结果显示，BP神经网络在针对用户活跃度2类判断中，整体上具有较高的预测性能. 当前预测算法存在如下缺点及主要改进之处：①用于预测的特征数量与特征组合为人工确定，缺少对特征最优组合的分析；②针对社交群用户活跃度的预测采用Logistic回归、支持向量机、BP神经网络等机器学习算法，预测准确率不高，缺少提出一个更加有效的算法. 在下一步的工作中，我们拟进一步提取更多特征，分析特征之间的相关性，找出特征最优组合，并提出预测精度更高的新算法.

参考文献：

[1] 蒋文丽，汤 庸，许玉赢，等. 社交网络中角色活跃度的好友推荐[J]. 小型微型计算机系统，2016，37(10)：2162－2165.

[2] HAN Y，TANG J. Who to Invite Next? Predicting Invitees of Social Groups [C]// 26th International Joint Conference on Artificial Intelligence. Melbourne：AAAI Press，2017：3714－3720.

[3] QIU J Z，LI Y X，TANG J，et al. The Lifecycle and Cascade of WeChat Social Messaging Groups [C]// 25th World Wide Web Conference. Montréal：ACM Press，2016：311－320.

[4] Butler B S. Membership Size，Communication Activity and Sustainability：A Resource－Based Model of Online Social Structures [J]. Information Systems Research，2001，12(4)：346－362.

[5] 李献礼. 在线社会网络多维度链式信任计算算法[J]. 西南大学学报(自然科学版)，2016，38(8)：142－147.

[6] 黄 炜，李总苛，李岳峰. 微信用户活跃度影响因素分析[J]. 湖北工业大学学报，2015，30(6)：29－32.

[7] SHI X L，ZHU J，CAI R，et al. User Grouping Behavior in Online Forums [C]//15th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Paris：ACM Press，2009：777－785.

[8] YANG J，WEI X，ACKERMAN M S，et al. Activity Lifespan：An Analysis of User Survival Patterns in Online Knowl-

edge Sharing Communities [C]//4th International Conference on Weblogs and Social Media. Washington: the Association for the Advancement of Artificial Intelligence, 2010: 186－193.

[9] ZHU Y, ZHONG E H, PAN S J, et al. Predicting User Activity Level in Social Networks [C]//22nd ACM International Conference on Information & Knowledge Management. California: ACM Press, 2013: 159－168.

[10] 周　静，李　季. 从人口统计特征和生活方式探讨 SNS 社交网站用户的使用行为及其活跃度 [J]. 科技与管理，2011，13(2)：72－77.

[11] MAGNUSON M J, LAUREN L. Gender Differences in "Social Portraits" Reflected in My Space Profiles [J]. Cyber Psychology and Behavior, 2008, 11(2): 239－241.

[12] 梁修荣，杨正益. 基于聚类和 SVM 的数据分类方法与实验研究 [J]. 西南师范大学学报(自然科学版)，2018，43(3)：91－96.

On Prediction and Analysis of Social Network Group User Activity

ZHANG Xiao-wei[1], YU Yun-xia[2], WANG Wei[1]

1. *School of Network Engineer, Zhoukou Normal University, Zhoukou Henan 466001, China;*

2. *School of Computer Engineering, Jingchu University of Technology, Jingmen Hubei 448000, China*

Abstract: For different groups in a social network, their activities are frequently influenced by a variety of factors such as user's attributes (i. e., gender, age), group classes, social relationships between group members and so on. In order to model and analyze the activities of different groups in this paper, several features which may influence the activity of a group have first been extracted from the historical data generated from a social network, such as census information, social relationships, group class, user stickiness (the number of the shared photos and information) and so on. Then, based on the extracted features, the activity of a group is predicted using logistic regression model, support vector machine and BP neural network. The results show that BP neural network has high performance on classifying group users' activity, and social relationships have a major impact on the activity of a group.

Key words: social network group; user activity; demographic informatics; social relationships

责任编辑　崔玉洁