信号课程SRT实验报告

周若凡 汪忆辰 叶紫

2013.9 － 2014.1

引言 12

项目概述 12

现有研究分析 12

研究方法 13

数据收集 14

问卷制作 14

数据抓取 14

特征提取 14

部分特征说明 16

数据分析 18

机器学习算法 18

拟合总体效果 18

相关性检测 18

结论 20

应用系统和未来探索 21

应用系统 21

未来探索 21

文件说明和运行说明 22

感谢 23

参考文献 24

# 引言

### 项目概述

现在使用社交网络的人越来越多，而一些学术研究表明人们在社交网络上的一些行为和自身的性格有一定的联系。我们通过对社交媒体（目前锁定为人人网）中的文本、图片数据、音频信息的收集和分析，设计出了一个通过社交网络信息来分析用户性格（大五类性格）的模型，并通过人人网上的应用来展示成果。

### 现有研究分析

Golbeck J, Robles C, Turner K在Predicting personality with social media[C]中介绍了他们关于利用Facebook的用户数据预测用户性格的工作。他们使用的是近年来公认为比较有参考价值的Big-Five人格特征作为性格模型，包括Openness to experience（开放性）、Conscientiousness（严谨性），Extraversion（外向性），Agreeableness（宜人性），Neuroticism（神经质）。一般Big-Five性格测试是问卷测试的形式，得到五个方面的分数比重，类似右图。

他们通过Facebook上的应用向279位用户收集了性格调查结果和个人数据，总结出六大类特征，包括语言特征、个人信息、活动和偏好、Facebook内部数据、线性回归结果、朋友圈结构（见左图）。然后分析了各种特征和性格之间的相关度，筛选出74项特征项，并用这些数据作为训练集训练出来两种机器学习的模型：M5’ Rules和Gaussian Processes，作为预测用户性格的模型。对比10折检验的效果，他们认为效果最好的是M5’ Rules模型。文章最后还讨论了用户性格预测的潜在市场价值。

另一篇文章，Quercia D, Kosinski M, Stillwell D的Our Twitter profiles, our selves: Predicting personality with Twitter，是基于Twitter用户数据的用户性格预测分析，其中也使用了前一篇中提到的M5’ Rules模型，但是只考虑了3个特征项（following, follower and listed counts）。

他们的分析结果表明，Twitter上不同的用户类型，包括listeners, popular, highly-read, and influential users，都在情感上表现的较为稳定，大部分属于外向型性格，而非神经质性格。现在使用社交网络的人越来越多，而一些学术研究表明人们在社交网络上的一些行为和自身的性格有一定的联系。

### 研究方法

根据对现有的研究成果的分析，我们讨论出了项目实行办法如下：首先收集一定数量（计划为200以上，最终收集到约90人次）的用户数据（包括其社交网络上的文本、图片和音频信息，以及性格量化值）；第二步根据已有研究和自己的分析选取特征并进行与性格相关的分析，并计算得到一个比较好的模型；最后制作能在社交网络上的应用共给用户使用。

我们选取的社交网络为人人网（中国领先的实名制社交网络，用户多为学生或年轻一辈），因为它的API可以方便地为我们提供获取各种信息（除了音频）的接口，同时也方便我们数据的收集和成果展示。

我们选取的用户性格量化方法为大五类人格特征：大五因素模型是一种描述模型的人格,心理学家们已经开发出一批理论证实大五因素模型，一共包含O、C、E、A、N五个数值，每个人的性格可以通过在这五个方向的数值来表现（每个方向的数值为0-1，越高说明性格在这个方向上越强）。

我们获取性格用户的方法为让用户自己来做计算其大五类人格特征的问卷（该问卷为经典问卷，可以认为其可靠性很高，翻译自<http://www.outofservice.com/bigfive/>）。

# 数据收集

我们的数据来源是利用参与问卷调查的用户id和人人网提供的接口，抓取的用户状态、日志、相册、分享、个人主页等信息。通过不同的方法进行特征抽取和量化处理后得到50多种不同的特征。

我们的标准数据是由用户填写性格调查问卷后，在bigfive网站上得到的标准性格指数。抓取下来的各种原始数据以json格式存储，方便后续的分析处理。

由于我们使用了weka机器学习工具包，所以我们将抽取出来的特征数据集整理为ARFF文件格式存储，以ARFF文件作为输入进行机器学习的模型训练。

### 问卷制作

我们翻译了<http://www.outofservice.com/bigfive/>上的题目，并搭建django的web应用框架，编写了一个用于采集人人网数据的app，并在实验室搭建了服务器，让人人网上的好友来做此问卷，同时收集用户的性格测试结果（这些结果我们假设全是正确的，即这些数据被我们当作了训练的标准集合）和人人id，方便之后的数据收集。（应用的截图见右边）

### 数据抓取

我们使用python语言便捷的web工具，和人人应用提供的token和API对做了问卷的用户id进行的数据的扒取（事先已经获得了这些用户的同意），主要扒取的数据为状态、日志、分享、回复、相册、照片（由于token的使用有一定的时间限制，整个扒取的过程比较缓慢）。后期通过观察网页的url，自行扒取了用户的语音相册里的语音mp3。

### 特征提取

通过整理、分析、阅读文献和讨论，我们制定了若干个数据特征（最详细的数据收集统计表可以见analysis/personality.xls），并通过java编程完成了对数据的格式化（其中图片处理和音频处理使用了matlab相关的函数库，并在原工程中调用了导出的用matlab编写的函数的jar文件），所含特征的集合可以见下面：

1. 状态

|  |  |
| --- | --- |
| 抓取的信息 | 用户每条状态的内容，每条状态的创建时间，每条状态的分享数，每条状态的评论数量，状态是否为分享 |
| 量化的数据 | 状态数，平均被分享数，平均评论数，分享状态比例，状态长度，平均两条状态间隔，表情的统计 |

1. 日志

|  |  |
| --- | --- |
| 抓取的信息 | 每篇日志的标题，内容长度，类型（通过什么发布），创建时间，分享数量，是否加密，浏览数量，回复数量 |
| 量化的数据 | 日志数，平均被分享数，平均评论数，平均浏览数，平均长度，手机发布占比例，公开比例，平均两篇日志的间隔 |

1. 分享

|  |  |
| --- | --- |
| 抓取的信息 | 每个分享的标题、URL、回复数量、分享的时间 |
| 量化的数据 | 分享数，照片类型比例，日志类型比例，视频类型比例（只能识别url，现在识别的有youku tudou 56 youtube），其他类型比例，评论数量，平均两个分享的间 |

1. 个人信息

|  |  |
| --- | --- |
| 抓取的数据 | 姓名，是否星级用户，basicinfo的填写比例，education的填写比例，work填写的数量，like的数量，app数量，来访数，关注小站数量，喜欢的音乐数量，填写看过的电影数，好友数，好友密度关系 |
| 量化的数据 | 姓名，是否星级用户，basicinfo的填写比例，education的填写比例，work填写的数量，like的数量，app数量，来访数，关注小站数量，喜欢的音乐数量，填写看过的电影数，好友数，好友密度关系 |

1. 其他时间相关特征

|  |  |
| --- | --- |
| 抓取的数据 | status.dat share.dat blog.dat |
| 量化的数据 | 新鲜事发布的平均时间（24小时），过期的新鲜事，新鲜事平均时间间隔 |

1. 图片特征

|  |  |
| --- | --- |
| 抓取的数据 | 所有相册中的图片 |
| 量化的数据 | 图片个数、相册个数、平均每个相册所含图片个数、人脸识别、平均每张图所含的脸数、256维颜色直方图的平均值、纹理的粗糙度，对比度，方向度的平均值 |

1. 音频特征

|  |  |
| --- | --- |
| 抓取的数据 | 语音相册中的语音文件 |
| 量化的数据 | 能量熵、均值信息熵、过零率、频谱衰减、频谱流量、矩心 |

最终我们得到的量化后的数据如下（这里由于大小的原因只列了很小的一部分，也就是说表格的宽和长都比下表要大很多）：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| StatusNum | AvgStatusShare | MaxStatusShare | AvgStatusCmt | MaxStatusCmt | StatusTimeDiff | MinStatusTimeDiff |
| 143 | 0.314685315 | 6 | 7.902097902 | 66 | 1500752.676 | 5 |
| 894 | 0.070469799 | 5 | 5.559284116 | 91 | 200112.8152 | 20 |
| 342 | 0.260233918 | 4 | 6.143274854 | 41 | 480388.261 | 57 |
| 800 | 0.71125 | 8 | 4.8175 | 45 | 72249.92991 | 43 |
| 192 | 0.572916667 | 14 | 10.36458333 | 68 | 618187.5602 | 93 |
| 14 | 0.642857143 | 2 | 1.857142857 | 9 | 1580921 | 51 |
| 510 | 0.117647059 | 3 | 2.511764706 | 45 | 317290.2672 | 4 |
| 1733 | 0.193306405 | 7 | 8.270628967 | 175 | 89089.13626 | 8 |
| 598 | 0.342809365 | 6 | 6.806020067 | 105 | 248304.1876 | 11 |
| 569 | 0.323374341 | 6 | 7.766256591 | 65 | 261681.8451 | 10 |
| 760 | 0.485526316 | 8 | 3.947368421 | 61 | 158497.7036 | 12 |
|  |  |  | … |  |  |  |

### 部分特征说明

* 正负面情感词：对于一条状态中的文本，可以利用中文文本分词工具ICTCLAS进行分词，得到带有词性标注的文本，如“人物/n 很/d 饱满/a”。分词后我们可以很容易过滤出一段文本中可能带有情感取向的词，一般为形容词、副词或名词。再通过正负面情感词词典对一个用户的所有状态的情感词进行统计，就可以得到一些非常有意义的统计量，如用户平均每条状态中正面情感词数量、负面情感词数量。进一步，对每条状态中正负面情感词数量进行比较，可以基本判定一条状态整体的情感取向，从而得到用户所有状态中表达正、负面情感的状态所占的比例情况。
* 好友密度：对于好友关系我们可以形成下面这张图：每个人是一个点，如果两个人是好友关系（在人人网上这个是相互的）则连一条边。每个人的好友（不包括他自己）形成了一张图。我们定义这张图的边数除以完全图的边数为好友密度。
* 表情统计：人人的状态中可以插入图片，比如等等。我们把表情也抽取了出来，作为一个特征计算。
* 人脸识别：通过matlab编程对每张图片进行了人脸识别，算法主要为扫描每个图中和人脸肤色相近的颜色块，并判断该颜色块的形状和面积是否满足一定条件，如果满足则认为是一个脸。下图为一个测试结果（可以看到脸被红色的方框圈了出来）



* 颜色直方图：即每个颜色在图片中所占的比例，根据一些文献中说可以一定程度反应图片表达的情感。
* 图片粗糙度、对比度、方向度：这里是选取了著名Tamura纹理特征中的三个，文献中提到这三个可以比较准确地表达纹理的图片特征。其中粗糙度为图片活动窗口像素的平均强度值；对比度为灰度图片的方差和alpha通道数值的计算；方向度为每个像素的梯度向量。
* 音频特征：这里根据一些文献选取了6个特征（分别有公式计算），并用活动窗口取方差的方式进行计算。

# 数据分析

### 机器学习算法

我们进行对比测试的机器学习方法是针对量化数据（arff格式）进行估值的模型，包括线性回归模型、M5Rules和高斯过程回归。利用现有的机器学习工具Weka，这一过程并不困难。我们将对每个模型的参数进行调整，已达到该模型的最好性能。在此基础上对比各模型的性能选择最好的模型用到最终的APP中。

M5Rules是用于解决回归问题的机器学习算法，基于分割覆盖的思想根据给定的训练集生成一系列规则，每一次迭代会建立一棵M5的决策树，并把最“好”的叶子节点添加到规则中。

高斯过程回归类似于贝叶斯回归，只是用RBFKernel核函数替代了贝叶斯中的线性函数进行回归。

### 拟合总体效果

我们对我们进行对比了M5和高斯过程的拟合效果，最终结果如下（可以看出高斯过程的拟合效果非常好）：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Correlation Coefficient | Agreeableness | Openness | Conscientiousness | Neuroticism | Extraversion |
| M5Rules | 0.7353 | 0.7335 | 0.8241 | 0.7196 | 0.4644 |
| Gaussian Process | 0.9602 | 0.9663 | 0.9618 | 0.9621 | 0.9583 |

### 然后，我们对我们预测的结果和原来实际的结果进行了比对，可以见下表。第二列数据为实际数据（我们收集到的数据）在该方向维度上的平均值，第三列为我们训练完的模型在原数据上跑出的结果的平均值（可以看出基本上一样）。第三列为偏差的平均值（偏差的计算方法为：|predict-actual|/actual）；第四列为偏差的最大值。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | actual | predict | Error(precentage) | max\_diff |
| Agreeabless | 0.6368 | 0.6369 | 0.115 | 0.19 |
| Conscientiousness | 0.5186 | 0.5275 | 0.146 | 0.17 |
| Extraversion | 0.5266 | 0.5253 | 0.124 | 0.15 |
| Openness | 0.6154 | 0.6154 | 0.120 | 0.22 |
| Neuroticism | 0.5209 | 0.5209 | 0.197 | 0.28 |

### 相关性检测

我们利用SPSS统计分析工具对测试的结果进行了相关性检验（Pearson检验），表格如下（由于篇幅关系，我们不可能将50多个特征的相关性都列出，具体可以看analysis/correlation.xls）：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | Agreeableness | Openness | Conscientiousness | Neuroticism | Extraversion |
| PosWordNum | 相关性 | -.091 | -.022 | -.007 | -.019 | .072 |
| 显著性 | .400 | .835 | .945 | .861 | .507 |
| N | 88 | 88 | 88 | 88 | 88 |
| NegWordNum | 相关性 | -.128 | .041 | .181 | .098 | .131 |
| 显著性 | .236 | .706 | .091 | .365 | .223 |
| N | 88 | 88 | 88 | 88 | 88 |
| PosStatusRatio | 相关性 | .062 | -.090 | -.223\* | -.088 | -.080 |
| 显著性 | .564 | .405 | .036 | .413 | .461 |
| N | 88 | 88 | 88 | 88 | 88 |
| NegStatusRatio | 相关性 | -.062 | .090 | .223\* | .088 | .080 |
| 显著性 | .564 | .405 | .036 | .413 | .461 |
| N | 88 | 88 | 88 | 88 | 88 |
| Agreeableness | 相关性 | 1 | .056 | -.006 | -.477\*\* | .272\* |
| 显著性 |  | .601 | .955 | .000 | .010 |
| N | 88 | 88 | 88 | 88 | 88 |
| Openness | 相关性 | .056 | 1 | .173 | .104 | .183 |
| 显著性 | .601 |  | .107 | .333 | .088 |
| N | 88 | 88 | 88 | 88 | 88 |
| Conscientiousness | 相关性 | -.006 | .173 | 1 | -.170 | .200 |
| 显著性 | .955 | .107 |  | .114 | .061 |
| N | 88 | 88 | 88 | 88 | 88 |

从表中可以看出，由显著性水平可以看出我们选取的一些特征和用户性格是相关的，例如表中PosStatusRatio（正面情感状态比例）对Conscientiousness（严谨性）的双侧Pearson检验在0.05的水平上显著相关，这说明PosStatusRadio越大，该用户的性格在严谨性方面就会差一些，即越多的正面情感的表达意味着越低的严谨性，这从日常生活经验出发是可以理解的。

### 结论

我们对我们收集到的数据先进行了人工分析。我们对各性格进行平均后发现，用户在A（宜人性）和O（开放性）方面的数值都比较大（为0.6，而认为人群平均值为0.5），这里也验证说明了使用社交网络的人群比普遍人群性更更开朗、开放一些，这点也是和文献中说明的是一样的。

我们的拟合的总体结果可以说是相当不错的，在数据的数量尚未达到预期目标的条件下，我们的偏差还是控制在了20%以内。随后我们随机选取了一些人进行测试，发现预测结果和心理测试结果也比较相符，不会出现非常离谱的预测（比如某性格方向的偏差达到了50%），说明社交网络上的使用还是能表现用户的性格的。

由于数据抓取的关系，我们主要还是对文本分析的比较多，从相关性分析中我们可以发现，事实上文字上的表达（不管是词性的分析还是量化的结果）其实就能体现出用户的性格特点。

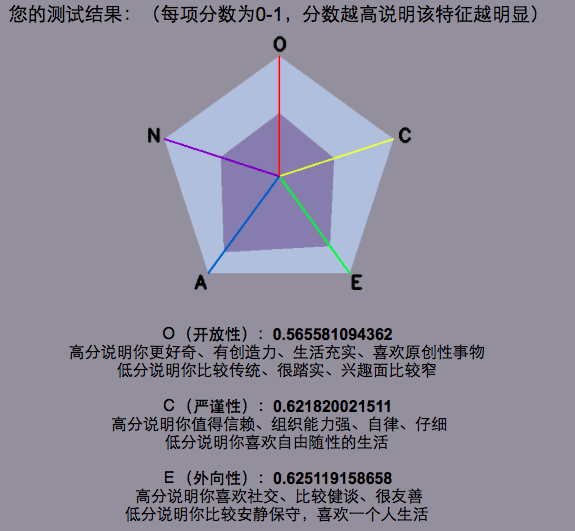
用户使用图片情况（包括对图片的分析等等）也能一定程度上表现用户的性格特点，但由于我们抓取的图片特征还比较少，所以图片在我们整个模型计算的比例也比较少。

由于人人网上使用语音功能的人数比较少，我们能收集的数据也比较少，所以考虑最终效果，所以虽然已经完成相关的算法但我们并没有将该方面的分析加入到最终的预测模型中。

# 应用系统和未来探索

### 应用系统

由于配置关系（我们的项目需要配置比较多的工具），我们在自己的电脑上配置了服务器，用户需要登录<http://apps.renren.com/testfive> 授权后即可以等待计算结果。所以服务器有的时候可能没有打开，需要测试前请提前联系小组成员。



### 未来探索

该项目已经申报了挑战杯。我们会继续在这个项目上扩展：除了更多的分析图片和音频特征之外、收取更多的数据做更多的测试之外，我们也会更新界面。

# 文件说明和运行说明

srt/ --项目总体文件夹

analysis/ --数据、训练模型程序和一些运行结果

data\_img/ --抓取下来所有用户的图片文件

data\_jason/ --抓取下来并已经量化的数据，每个用户一个文件夹

data\_voice/ --抓取下来用户的音频文件

result/ --收集的用户性格数据

src/ --数据量化的源文件

srt/ --matlab导出的jar工程

bigfivetest/ --最终app的django代码，运行服务器请输入python manage.py runserver

crawl/ --抓取数据的python代码

doc/ --文档（包括各种报告）

papers/ --参考文献

predict/ --预测模型和分析数据

server/ --性格测试问卷的django代码，运行服务器请输入python manage.py runserver

slides.pptx –课堂展示的ppt

＊＊启动项目需要保证以下工具：python、django、MCR、java

# 感谢

感谢徐老师给予项目上的支持和建议！

感谢助教每次辛苦的收取报告和对延期提交的谅解！

感谢为我们做了问卷和提供人人信息的各位同学！

# 参考文献

Affective image classification using features inspired by psychology and art theory

Facebook Profiles Reflect Actual Personality, Not Self-Idealization

Interpretable aesthetic features for affective image classification

Machine prediction of personality from Facebook profiles

Personality impressions based on personal Web sites

Personality in 100,000 Words

Personality in Cyberspace

Personality and Social Media Use

Predicting Personality with Social Media

Predicting Personality with Twitter

The Personality of Popular Facebook Users