**信号课程SRT实验报告（精简版）**

**——人人网用户性格分析与预测**

**2011011300 计15 周若凡（组长）**

**2011011308 计15 叶紫**

**2011011279 计14 汪忆辰**

**1 数据说明**

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。

**2 实验流程**

**2.1 标准数据获取**

我们先在人人上提供了一个APP，一调查问卷的形式获取用户的性格标准数据和id号。

2.2 特征选取

我们通过研究专业论文和小组讨论，列出了需要抽取的特征列表，见features.txt。包括文本类特征、图片类特征和音频类特征。

**2.3 数据抓取**

叶紫完全负责项目的所有数据抓取工作，用python编写了所有数据抓取的代码，见crawl文件夹。抓取的数据分用户和类别用json格式存储，以备后续分析。

**2.4 特征提取**

汪忆辰负责全部的文本类特征提取，包括用户的状态、日志、分享、个人主页的内容。

周若凡同学负责图片和音频的特征提取。

随后将所有特征整合到一起，形成五个性格特征对应的五个ARFF数据文件。

**2.5 相关性分析和模型训练**

使用SPSS统计学软件对提取出的特征数据和性格指数进行了相关性检验，结果见analysis/correlation.xls文件。

利用weka工具包中已有的机器学习方法，进行了线性回归、M5规则和高斯过程三种机器学习训练，并进行了效果的比较。

**2.6 性格预测**

最后我们完成了一个性格预测的人人APP，用训练出的模型对使用该APP的用户进行性格的预测。

**3 实验结果**

目前我们基本完成了模型训练的工作，三种模型中高斯过程（Gaussian Process）的结果最好，这一点与之前研究的论文结果不符，论文中说M5Rules的结果相对更好。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 |

我们最终的人人性格预测APP尚未完成，敬请期待！