# 多任务回归在社交媒体挖掘中的应用

白朔天[1], 袁 莎[2], 程 立[3], 朱廷劭[4]

(1.中国科学院大学 计算机与控制学院, 100190 北京; 2.中国科学院 声学研究所,100191 北京;
3.生物信息学研究所 新加坡科技研究局,138632 新加坡; 4.中国科学院 心理研究所, 100101 北京)

**摘 要**: 随着社交媒体的迅速发展,针对网络信息挖掘的研究成为互联网领域备受关注的研究热点之一.传统的单任务回归对各个任务分别建模,在多变量预测的场合中,无法合理利用变量之间的共享信息.因此,本文通过多任务回归网络挖掘方法,分析社交媒体用户人格和网络行为的关联模式.实验通过在线被试邀请,采集了335个人人网用户样本和563个新浪微博用户样本.采用多任务回归的算法,预测精度可达87%以上.实验结果表明多任务回归对多变量建模效果要优于单任务学习算法.

**关键词**: 多任务回归;社交媒体;网络挖掘;特征提取
**中图分类号**: TP391.4 **文献标志码**: A **文章编号**: 0367-6234(2014)09-0100-05

# Application of multi-task regression in social media mining

BAI Shuotian[1], YUAN Sha[2], CHENG Li[3], ZHU Tingshao[4]

(1. School of Computer and Control Engineering, University of Chinese Academy of Sciences, 100190 Beijing, China;
2. Institute of Acoustics, Chinese Academy of Sciences, 100191 Beijing, China; 3. Bioinfomatics institute, Agency for Science, Technology and Research, 138632 Singapore; 4. Institute of Psychology, Chinese Academy of Sciences, 100101 Beijing, China)

**Abstract**: With the development of Social Media, web mining analysis has been regarded as one of hot research topics. Traditional single task regression builds models for each task, which ignores the sharing information among tasks in the occasion of multi-variable prediction. Therefore, this paper used multi-task regression mining method, and managed to analyze the pattern between user's personality and network behavior. This study collected a sample set of 335 RenRen users and 563 Weibo users through online test invitation. Using multi-task regression, the final prediction accuracy is 87% or more. The result means that multi-task regression works better then single task regression for multi-variable modeling.

**Keywords**: multi-task regression; social media; Web mining; feature extraction

网络挖掘是数据挖掘技术在网络信息处理中的应用.网络信息挖掘是从大量训练样本基础上得到数据对象间的内在特征,并以此为依据进行有目的的信息提取[1].以人人网(http://www.renren.com)和新浪微博(http://weibo.com)为首的社交媒体在国内飞速发展.据《第31次中国互联网络发展状况统计报告》统计,截止2012年底,人人网已拥有超过2亿注册用户,新浪微博注册用户数已超过5亿.社交媒体在快速改变传统网络舆论格局的同时也逐渐展现出其自身所具有的独特优势.用户在社交网络中往往可以真实、自发地表达或分享自己的情感和观点.由于网络实名制的推进,用户的网络行为和现实行为具备较强的一致性[2].这就为网络用户的行为分析研究提供大量真实、可靠的潜在数据源.针对网络挖掘建模中的多个具有相关性的任务(如用户大五人格预测[3])在同一训练集的同时学习问题,传统方法(如回归、神经网络)是在训练集上对各个任务分别建模[4].这种方法虽然考虑了各个任务的特定信息,但是忽略了任务之间的相关性,没有考虑到任务之间的某些共享信息.多任务学习不仅可

**收稿日期**: 2013-12-10.
**基金项目**: 国家自然科学基金资助项目(61070115).
**作者简介**: 白朔天(1987—),男,博士研究生;
朱廷劭(1971—),男,研究员,博士生导师.
**通信作者**: 白朔天,baishutian10@mails.ucas.al.cn.

以保留任务的特定信息，更可以计算出任务间的共享信息，建立更准确的预测模式.最早的多任务学习方法由 Caruana[5] 提出，采用前馈神经网络进行建模，打破每次训练只针对一个任务的限制.由此得来的训练结果使得输入结点和隐藏层结点的连接权包含任务之间的共享信息，隐藏层结点和输出结点之间包含了各个任务的特定信息.虽然该方法并不复杂，但这种思路启发了学者们采用多任务学习的思路进行建模，并在机械自动化、医疗诊断等其他领域得到了应用.

本文创新性地提出采用多任务回归的方法在社交媒体中采集用户行为数据，并挖掘网络用户人格多维度与行为的相关模式[6].通过调查网络用户的大五人格，一方面分析不同人格用户的行为模式，另一方面通过分析用户的网络行为进行其大五人格的预测.由于人格的 5 个维度之间存在相关因素[7]，因而建立了基于多任务回归人格预测模型，并通过对被试用户的人格进行预测，验证了多任务回归模型的预测效果要优于其他模型.

## 1　社交媒体网络挖掘建模和特征提取

### 1.1　实验平台和网络数据采集

为了高效采集被试样本，开发一个基于人人网和新浪微博的在线问卷调查平台.本平台以第三方应用的形式接入到社会媒体中.用户可以通过其人人网或新浪微博帐号登录到平台并授权，在线填写心理学普遍认同的 NEO 大五人格问卷.在得到用户授权后，平台可以通过社交网站开放的 API 自动下载用户网上数据并保存到本地数据库.开放平台提供 API 调用方式，允许被用户授予权限的第三方应用以社交媒体用户的身份来读写社会媒体网站的资源（例如：用户基本资料、好友关系、照片等）.下载得到用户数据后，平台通过计算用户填写人格量表的结果可以得到用户的大五人格得分，并最终用人格得分对用户网络数据进行标注.平台工作流程如图 1 所示.

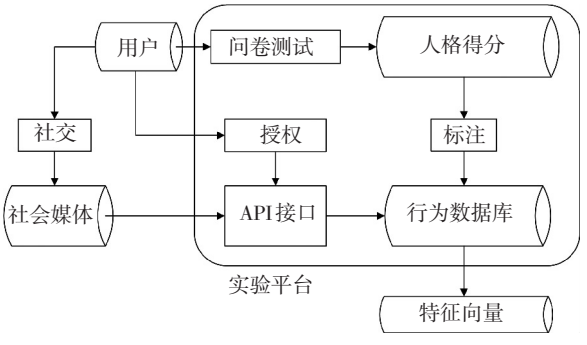


**图 1　实验平台流程图**

研究的用户实验开展于 2012 年的 1 月至 2 月.在本次实验中，只保留了活跃的用户数据进行建模与分析，非活跃用户被视为无效样本.其中，人人网非活跃用户定义为好友数少于 50，状态数少于 20，并且日志数少于 3 的用户；新浪微博的非活跃用户定义为状态数少于 50，并且在实验开始前 3 个月内有发布过微博.被试者通过社交媒体进行招募，共有 335 个人人网用户参与了实验，其中 209 名用户（141 位男性，68 位女性，平均年龄 23.8 岁）符合被试要求；共有 563 个新浪微博用户参与了实验，其中 444 名用户（171 位男性，273 位女性，平均年龄 23.8 岁）符合被试要求.

### 1.2　网络特征提取

本实验共设计 86 个用户网上特征，并计算特征与用户人格的相关度.发现在人人网和新浪微博环境中，分别有 10 个特征和用户人格具有相对较高的关联性.这些特征分别是，人人网状态数、日志数、相册数、留言数、评论数、好友数、评论人数、最近状态数、男好友比例、好友评论比例；新浪微博状态数、自我描述长度、是否默认头像、域名长度、关注数、互粉数、粉丝数、原创微博比例、互粉关注比例、互粉粉丝比例.

人人网支持用户发布短文本状态的功能，特征状态数就是用户所发表的全部状态的总数.用户可以在人人网上发表长文本的日志，特征日志数指的是用户发表日志的总数.相册数是用户上传的相册总数.不同的用户可以在彼此页面的留言板留言，特征留言数就是用户留言的总数.用户所发布的状态、日志可以被其他用户评论，特征评论数就是所有用户被评论的总数.特征最近状态数是最近一个月内用户发表的状态数.特征好友评论比例指的是所有评论中，来自该用户好友的评论占总评论的比例.

新浪微博提供了发布短文本微博状态的功能，特征状态数指的是用户发表的全部状态总数.自我描述长度指的是用户自我描述中的字符数.用户可以根据自我偏好设置个性域名，特征域名长度指的是用户个性域名的字符数.新浪微博支持单向的好友关系构建，这个人人网的双向好友关系有所不同.微博用户可以自由关注他人，也可被他人关注；因此特征关注数表示用户关注其他微博用户的总数，特征粉丝数表示某用户被其他用户关注的总数，特征互粉数表示既在关注列表又在粉丝列表的用户数.

### 1.3　特征评估

为验证特征的有效性，本实验通过计算特征和标注之间的皮尔逊相关系数作为特征有效性的检

验指标.表 1、2 给出了用户网络特征集与其大五人格的皮尔逊相关系数$\rho$,及其对应的显著性 $p$ 值.

$$\rho = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

其中$\rho \in [-1,1]$代表两个变量之间的相关程度.若$\rho > 0$,则随着 $X$ 的增长,$Y$ 也呈现增长趋势,且$\rho$ 越大,这种趋势越明显,反之亦然.$p$ 值为显著性水平,其值越小,表示相关结论偶然发生的可能性越小,结果的可靠性越高.

**表 1 人人网用户网络特征与大五人格的相关系数**

| 特征 | 宜人性 | 尽责性 | 内外向 | 神经质 | 开放性 |
|---|---|---|---|---|---|
| 状态数 | 0.08 | 0.12 | −0.06 | 0.07 | 0.15* |
| 日志数 | 0.12 | 0.01 | 0.02 | 0.05 | 0.10 |
| 相册数 | 0.26** | 0.12 | 0.17* | 0.18** | 0.11 |
| 留言数 | −0.11 | −0.07 | −0.04 | −0.07 | −0.02 |
| 评论数 | 0.12 | 0.05 | −0.01 | 0.07 | 0.13 |
| 好友数 | 0.11 | 0.06 | 0.07 | −0.08 | −0.17* |
| 评论人数 | 0.21** | 0.12 | −0.04 | 0.01 | 0.16* |
| 最近状态数 | −0.19** | −0.01 | −0.18** | 0.12* | −0.15* |
| 男好友比例 | 0.30** | 0.07 | 0.06 | −0.05 | 0.19** |
| 好友评论比 | −0.01 | 0.01 | −0.16* | 0.16* | 0.02 |

* * $p < 0.01$. * $p < 0.05$. $n = 209$.

**表 2 新浪微博用户网络特征与大五人格的相关系数**

| 特征 | 宜人性 | 尽责性 | 内外向 | 神经质 | 开放性 |
|---|---|---|---|---|---|
| 状态数 | 0.13* | −0.01 | −0.04 | 0.02 | 0.13* |
| 自我描述 | −0.02 | 0.05 | 0.11 | 0.01 | 0.10 |
| 默认头像 | 0.06 | −0.01 | 0.02 | −0.07 | 0.15* |
| 域名长度 | 0.09 | 0.16** | −0.08 | −0.05 | −0.05 |
| 关注数 | 0.04 | 0.03 | −0.16** | 0.02 | 0.06 |
| 互粉数 | 0.10 | 0.08 | 0.13* | −0.06 | 0.06 |
| 粉丝数 | −0.01 | 0.12* | −0.02 | 0.03 | 0.04 |
| 原创微博比 | 0.03 | −0.02 | 0.13* | −0.01 | −0.03 |
| 互粉关注比 | 0.08 | −0.01 | 0.01 | −0.07 | −0.13* |
| 互粉粉丝比 | 0.03 | 0.04 | −0.16* | −0.10 | 0.07 |

* * $p < 0.01$. * $p < 0.05$. $n = 444$.

结果表明,人人网用户大五人格中的宜人性和相册数、评论人数、男性好友比例有着明显的正相关,和最近发表的状态数有明显的负相关;微博用户的宜人性则和状态数)有着显著正相关.宜人性低的个体,容易和他人产生矛盾,对他人冷漠麻木,容易在网络的非面对面环境中产生过激的言行,激发网络安全问题[8].

人人网用户的尽责性和状态数、相册数和评论人数有着正相关的联系;微博用户的尽责性表现在域名长度和粉丝数.尽责性可理解为自律,对他人、事物的责任心等.低尽责性的用户容易和其他用户因意见分歧而产生过激行为[9].

内外向指的是个体自我魅力展示的程度,是社交能力的重要表现.在人人网中,它和相册数呈正相关,和最近状态数、好友评论比例呈明显负相关;在微博中,它和关注数呈现负相关,与互粉数和原创微博比呈现正相关[10].外向人会上传更多的私人照片来展示自我魅力.

人人网中神经质维度和相册数、好友评论比例呈现明显的正相关;微博中神经质与互粉粉丝比呈现较弱的正相关.在大五人格理论中,神经质被定义为情绪稳定性.通常而言,神经质维度得分越高的人情绪越不稳定.这样的人容易喜怒无常,容易让自己陷在抑郁或焦虑的状态中[11].

人人网用户开放性和状态数、男好友比例、评论人数呈正相关,和好友数、最近状态数呈负相关.微博中开放性和状态数正相关,同时高开放性的用户更倾向使用个性头像.开放性反映了个体想象力的丰富度,对新事物的好奇度[12].高开放性的用户在网络中会表现得较为随和亲切,不会因为是陌生人而产生怠慢、粗鲁的行为.

## 2　回归建模

尝试两种回归方法：增量回归和多任务回归.

增量回归是一种使用多个线性模式的组合，以拟合复杂的非线性问题的方法(算法 1).

算法 1. 增量回归

输入：样本集($X$ | $Y$)，误差阈值 $e_0$，最小样本数 $n$

　定义：

　　样本队列：对样本集的特征进行归一化并排序

　　模型队列：[模型参数，定义域]

　　点数组：若干个样本点的集合

　　Regression ($X$)：用数据集 $X$ 进行线性拟合

　REPEAT

　　从样本队列中提取前 $n$ 个样本放进点数组

　　模型=Regression(点数组)

　　对样本队列中下一个点计算测试误差

　　if(测试误差小于 $e_0$)

　　　将该点放入点数组

　　　重新拟合模型，模型=Regression(点数组)

　　else

　　　将模型和定义域保存至模型队列

　　　清空点数组

　UNTIL 样本队列无待训练样本

输出：模型队列

增量回归首先对样本集合进行排序，选取少量点进行局部建模. 随后用这个局部模型对新的训练样本进行测试. 当测试误差超过阈值时，则理解为模式的跳变，并把当前模型保存重新执行算法. 此方法可将复杂的模式通过多个简单的模型表达出来，在处理非线性问题时能显示出极强的优势. 然而在建模过程中，其参数需要严格控制. 首先，面对排序策略的不同，模型的效果可能差距极大. 通常情况下，根据归一化样本的模从小到大排序. 其次建模的最小样本数 $n$ 也会对结果产生很大影响. 若 $n$ 的值过大，则模型退化为线性回归；若 $n$ 过小，则局部模型的准确度降低. 一般而言，可设置 $n$ 的值为训练集样本的维数. 例如一个在两维空间中的回归问题，$n$ 可设置为 2.

增量拟合虽然可以处理非线性的问题，但它只能对各个任务分别建模. 在处理多任务学习的过程中，无法考虑任务间的共享关系. 多任务学习的主要目标是在同一场景下采用多个任务学习的策略来提高性能以超越单任务学习的效果. 假设有 $T$ 个回归任务，对于每个任务 $t$，都有一个独立的训练集合 $\{(x_{tn}, y_{tn})\}$，$t = 1,2,\cdots,T$，$n = 1,2,\cdots,N$. 式中，$(x_{tn}, y_{tn}) \in X \times Y$ 代表任务 $t$ 中第 $n$ 个实例标签对，$N$ 表示任务实例的个数(假设所有任务拥有相同的实例数目)，$\boldsymbol{x} \subseteq \boldsymbol{R}^d$，$\boldsymbol{y} \subseteq \boldsymbol{R}^{\mathrm{T}}$. 假设每个样本表示为列向量，则

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1N} \\ y_{21} & y_{22} & \cdots & y_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ y_{T1} & y_{T2} & \cdots & y_{TN} \end{bmatrix},\ \boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{2N} & \cdots & x_{dN} \end{bmatrix}.$$

多任务学习的目标是通过样本来预测 $T \times d$ 的传递矩阵

$$\boldsymbol{W} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1d} \\ w_{21} & w_{22} & \cdots & w_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ w_{T1} & w_{T2} & \cdots & w_{Td} \end{bmatrix},\text{通常 } T < d.$$

其中 $y_{ij} = \boldsymbol{W}_{i\cdot} \cdot \boldsymbol{X}_{\cdot j} = \sum_h w_{ih} \cdot x_{hj}$. 这种情况下，多任务学习的目标就是通过训练模型，找到使预测值和标注值之差最小的传递矩阵，也就是

$$\boldsymbol{W} = \operatorname{argmin}\{L(x, y, \boldsymbol{W}; 1:T) + \lambda \Omega(\boldsymbol{W})\}.$$

式中：$L(x, y, \boldsymbol{W}; 1:T)$ 代表训练样本中预测的经验损失函数；$\Omega(\boldsymbol{W})$ 是正则化函数；$\lambda$ 为正则项系数，通常为正.

实验采用最小平方损失和弗罗贝尼乌斯范数(Frobenius norm)的方法进行建模计算. 此时有

$$L(x, y, \boldsymbol{W}; 1:T) = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \sum_{t=1}^{T}\sum_{n=1}^{N}(y_{tn} - \hat{y}_{tn})^2 = \sum_{t=1}^{T}\sum_{n=1}^{N}\left(y_{tn} - \sum_h w_{ih}x_{hj}\right)^2,$$

$$\Omega(\boldsymbol{W}) = \| \boldsymbol{W} \|^2 = \operatorname{tr}(\boldsymbol{W}^{\mathrm{T}} \cdot \boldsymbol{W}).$$

上式进行变形得到

$$\boldsymbol{W} = \left(\lambda \boldsymbol{I} + \sum_n x_n x_n^{\mathrm{T}}\right)^{-1}\left(\sum_n x_n y_n^{\mathrm{T}}\right).$$

对应于本文的多任务人格预测，$T = 5$ 代表大五人格的五个维度，$d = 10$ 代表本文提取的网络特征；如果在人人网实验中，则 $N = 209$ 代表本实验采集的 209 个被试用户. 建立多任务人格预测模型的目标就是找到一个 $5 \times 10$ 的传递矩阵和一个可靠的避免过拟合的正则系数 $\lambda$.

## 3　模型应用实例与分析

主要探讨人格在网络社交圈中的行为表现模式，分析的目的是为建立人格计算模型选取准确的特征. 得到上述所有的分析结果后，开始用不同的机器学习算法进行大五人格的计算. 本文以高斯过程和线性回归作为基准，以绝对平均误差作为标准，证明了本文提出的方法在人格预测方面具有较好的性能.

尝试增量回归和多任务回归的学习方法，并根据预测精度将他们与高斯过程、线性回归和 M5P 回归做了对比. 在增量回归中，设置最小训练样本数为 11，误差阈值为 0.1，样本集根据模的大

小排序.在多任务回归中,经过对正则参数进行穷举计算,实验得到 $\lambda = 0.0973$ 时模型预测效果最稳定.采用 5 倍交叉验证,表 3、4 列出了采用不同算法对人人网和微博用户大五人格预测的误差率.

**表 3 人人网数据集上不同算法的大五人格预测误差率** %

| 维度 | 高斯过程 | 线性回归 | M5P | 增量回归 | 多任务回归 |
|---|---|---|---|---|---|
| 宜人性 | 12.49 | 12.85 | 12.69 | 12.45 | 12.39 |
| 尽责性 | 9.71 | 9.61 | 9.67 | 9.01 | 8.72 |
| 内外向 | 10.48 | 10.61 | 10.78 | 11.23 | 11.13 |
| 神经质 | 11.97 | 11.71 | 11.71 | 10.90 | 10.23 |
| 开放性 | 12.17 | 12.40 | 12.36 | 11.93 | 11.63 |

**表 4 微博数据集上不同算法的大五人格预测误差率** %

| 维度 | 高斯过程 | 线性回归 | M5P | 增量回归 | 多任务回归 |
|---|---|---|---|---|---|
| 宜人性 | 18.94 | 18.81 | 18.90 | 18.18 | 14.50 |
| 尽责性 | 18.71 | 18.99 | 17.65 | 17.65 | 14.42 |
| 内外向 | 20.36 | 20.86 | 20.76 | 19.20 | 14.00 |
| 神经质 | 21.64 | 21.96 | 21.31 | 20.37 | 14.75 |
| 开放性 | 20.26 | 20.21 | 18.46 | 18.91 | 11.50 |

从表 5 中数据表明高斯过程的平均预测误差为 15.67%,线性回归平均误差为 15.81%,M5P 平均预测误差为 15.43%.相比而言,高斯过程的预测效果略好.而增量回归误差率在 14.98%,多任务回归的平均误差率为 12.33%,预测精度好于其他经典算法.

在表 4 中,将本文的模型和相关工作中的模型从样本量、样本获取方式以及分析的方法进行对比.在方法上,通过调用 API 批量化获取用户网络数据.这种方法克服了 Kelly[13] 模型数据样本少,采集不够客观等局限;克服了 Correa[6] 模型中工作量巨大等问题.在目前的经典研究中,研究者一般注重于网络特征与人格的相关分析.本文在 Gosling[14] 工作的基础上,进一步用多种机器学习的方法建立了人格预测模型.

**表 5 本文工作与相关工作的对比**

| 研究 | 数据量 | 数据获取 | 建模方法 |
|---|---|---|---|
| Kelly[13] | 219 | 自陈 | 相关分析 |
| Correa[6] | 959 | 自陈 | 相关分析 |
| Gosling[14] | 292 | 自陈/人工记录 | 回归分析 |
| 本文 | 898 | API 自动获取 | 相关分析、人格预测 |

## 4 结 语

针对网络挖掘中,单任务建模对多变量预测的低效性,提出了采用多任务回归的思路预测社交媒体用户的人格变量.新方法可以在建模过程中合理利用多任务之间的共享信息,其预测精度要显著高于单任务算法.今后,本实验将会继续扩大实验范围,大规模采集更多的社交网站用户数据.继续设计并提取用户网络特征,进一步考虑研究心理学中的心理健康、社会态度等心理属性在社交网络中的行为表现模式.同时考虑更多的多任务学习方法,修改预测模型.

## 参考文献

[1] DOYD D, ELLISON N. Social network sites: definition, history, and scholarship [J]. Journal of Computer-Mediated Communication, 2007, 13(1): 210-230.

[2] GOBY V. Personality and online offline choices: MBTI profiles and favored communication modes in a Singapore study [J]. Cyber Psychology and Behavior, 2012, (9): 5-13.

[3] KOSINSKI M, STILLWELL D, GRAEPEL T. Private traits and attributes are predictable from digital records of human behavior [J]. Proceedings of the National Academy of Sciences, 2013, 110(15): 5802-5805.

[4] SCHWARTZ H, EICHSTAEDT J, KERN M, et al. Personality, gender, and age in the language of social media: the open-vocabulary approach [J]. PloS one, 2013, 8(9), e73791.

[5] CARUANA R. Multitask learning [J]. Machine Learning, 1997, (28): 41-75.

[6] CORREA T, HINSLEY A, ZIGA H. Who interacts on the web? The intersection of users´ personality and social media use [J]. Computers in Human Behavior, 2010, 26(2):247-253.

(下转第 110 页)