文章编号：1003-0077(2018)04-0120-10

基于社交媒体的用户情绪建模与异常检测

孙 晓[1]，张 陈[1]，任福继[1,2]

（1. 合肥工业大学 计算机与信息学院，安徽 合肥 230009；
2. 德岛大学 工程学院，日本 7700855）

摘 要：为了对新浪微博用户的异常情绪进行检测和分析，该文提出一种基于多元高斯模型和幂律分布的异常检测方法，根据联合概率密度值判断用户是否出现情绪异常。在实验部分，按照不同用户的异常检测准确率为83.49%，按照不同月份为87.84%。分布测试表明，单个用户的中性、快乐和悲伤情绪服从正态分布，而惊讶和愤怒情绪则不服从；群体发布的微博的情绪服从“幂律分布”，而单个用户则不服从。该文引入多元高斯模型来进行社交媒体的异常情绪的检测，通过联合概率密度值量化了异常情绪检测。当数据充足时，该方法可以检测用户或者某个社交平台每一周甚至每一天的异常情绪，这对个体异常情绪检测、网络舆情挖掘、大规模爆发事件预防以及公共安全监测有一定意义。

关键词：社交网络；异常检测；多元高斯分布；联合概率密度

中图分类号：TP391 **文献标识码**：A

User Emotion Modeling and Anomaly Detection Based on Social Media

SUN Xiao[1], ZHANG Chen[1], REN Fuji[1,2]

(1. School of Conpater and Information Hefei University of Technology, Hefei, Anhui 230009, China;
2. Faculty of Engineering, University of Tokushima, Tokushima, 7700855, Japan)

Abstract: For abnormal emotional detection among micro-blog users, this paper proposes ananomaly detection method based on the joint probability density of multivariate Gaussian model and power-law distribution. In the experiments, the anomaly detection accuracy is 83.49% in terms of individual user, and 87.84% in terms of month. Statistics reveals that individual users' neutral, happy and sad emotions fall into the normal distribution, but the amazed and angry emotions are not. Emotions of micro-blogs released by groups confirm to the power-law distribution, but not those by the individual.

Key words: social network; anomaly detection; multivariate gaussian distribution; joint probability density

0 引言

国内外社交平台正在迅速发展，根据新浪官方发布的2016年第三季度营业报告[1]，截止到2016年9月30日，微博的月活跃人数已达到2.97亿。其中，9月份的日活跃用户达1.32亿，较2015年同比增长32%；微博活跃用户中，拥有大学以上高等学历的用户始终是微博的主力用户，占比高达77.8%，他们发表的语言往往表征了一定的情感倾向性[2]。图1和图2是本文对100位用户在2011年5月—2016年5月间总计10 275条微博进行统计后，得到的五类情感分布以及情感极性比例情况。“伤心、生气”作为消极情绪占比19%，这一数据值得关注和研究。微博异常情绪检测是微博情感分析的一个重要领域。通过提取用户微博语料中的有价值的部分，可以很好地进行网络舆情的监督，甚至公共安全的监测。防止非理性情绪在网络中蔓延，对可能出现负面情绪的事件及时做出反应。防止某些不法分子企图通过微博平台传播谣言[3]，以维护社会的稳定与和谐。[4]进而还可以帮助企业根据用户对产品的评价和情感倾向，做出正确的决策，提高产

收稿日期：2017-03-03 **定稿日期**：2017-05-09

基金项目：国家自然科学基金(61432004)；安徽省自然科学基金(1508085QF119)；中国博士后科学基金(2015M580532)

品质量，减少不必要的损失，提高企业收益。

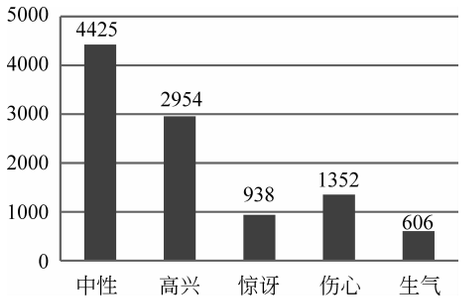

图 1　100 位用户五年微博类别数目

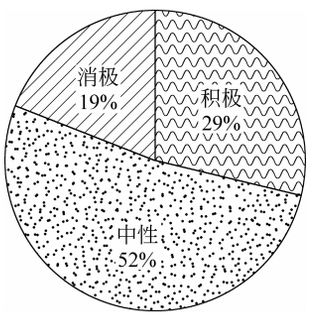

图 2　100 位用户五年微博情感极性分布

目前社交媒体的用户情绪异常检测方法主要有以下几种：Lin[5]提出了一种新的混合模型，因子图模型结合卷积神经网络，利用 Twitter 内容属性和社交信息来增强压力检测。本方法采用了大规模数据集，系统地研究用户的压力状态和社会交往的相关性。最后通过训练集预测未标注的用户压力状态，提高了检测准确性能。Zhang[6]等人以情绪因子中常用的情绪词和情绪短语为基础构建情感词典以及情绪规则库，进行微博情绪的识别和分类，取得较好结果。Wang[7]通过考虑一个文本的宾语（客体）来增加情感分类的方法，对社交数据进行异常检测，以 Twitter 的数据为样例学习，统计四天时间内推特上的中性、消极、积极情绪的数目。通过比较消极情绪比例来预测某一天网民情绪是否出现异常，得出的结论比较笼统，不能准确分析到具体事件和具体用户。李凌云[8]在基于微博的事件实时监测框架与系统中，提出用基于规则和统计的方法，用时间序列模型来监测异常，证明比普通的模型更有效果。Yin[9]等人提出一种基于提升系数的微博异常排名检测方法，有效防止了人为操纵微博提升排名的行为对微博排名的干扰。在仿真数据集的实验表明，该方法能通过微博拓扑有效地识别异常排名。

综合以上所述研究现状，目前异常情绪建模及异常检测主要是基于情感词典、文本分类、神经网络、基于统计和规则、基于时间序列分析模型和基于排名等，这些工作需要大量的标注语料作为训练集，故针对异常情绪的语料标注工作量很大。另外，目前的研究工作倾向于对一个社交平台上的所有语料进行分类和分析，从而检测出某个时间点的爆发异常事件，但针对单个用户的异常情绪的检测的研究较少。

1　准备工作

1.1　数据处理

为了检测特定用户或者特定时间段内的异常情绪，本文的工作分为三大阶段：数据处理阶段；异常情绪检测阶段；用户情绪建模阶段。其中数据处理阶段的工作在本节具体介绍，异常情绪检测和用户情绪建模将分别在本文第二节和第三节具体阐述。

数据处理阶段，首先采用多策略的大数据抓取技术[10]收集了 100 位微博用户从 2011 年 5 月到 2016 年 5 月间的 10 275 条新浪微博作为实验语料。原始语料标有相应的用户 id，发布时间等相关信息。利用 SVM 分类器[11]对获取的微博语料进行文本分类，结合人工纠正标注，得到用户微博的“中性、开心、惊讶、伤心、生气”类别数目。用户微博的五种情绪数目可以作为该用户情绪相关的变量（五维向量），考察该变量与用户情绪的相关性并建模。对于用户每一类情感的微博统计结果可以用单变量的高斯建模。而针对用户五类情感的这个五维向量，则引入多元高斯分布对用户情绪建模，并进行异常检测。对标注了情感的微博文本进行统计，本文是基于用户和时间两个角度进行异常检测，按照“用户、月份、微博类别、数目”和“月份、微博类别、数目”这两种方式进行统计。

1.2　高斯分布

高斯分布就是用高斯概率密度函数（正态分布曲线）精确地量化事物，将一个事物分解为若干个基于高斯概率密度函数形成的模型。高斯模型有单高斯模型（SGM）和多元高斯模型（MGM）[12]两种。

图 3 是本文对一个用户 11 个月的微博情感的数目进行高斯分布拟合，横坐标是用户每个月发布微博的数目，纵坐标是通过高斯函数计算出的联合概念密度。从图 3 可以看出，当数据集数目 N 较小时，拟合的图形并不完全符合高斯分布，但事实是，随着 N 的增大它会很快收敛于高斯分布。该结果的一个推论

是：当N趋于∞时，数据分布趋于高斯分布，所以对于一组数据，当数据量足够多的时候，可以对其进行高斯分布拟合，并进行异常检测。

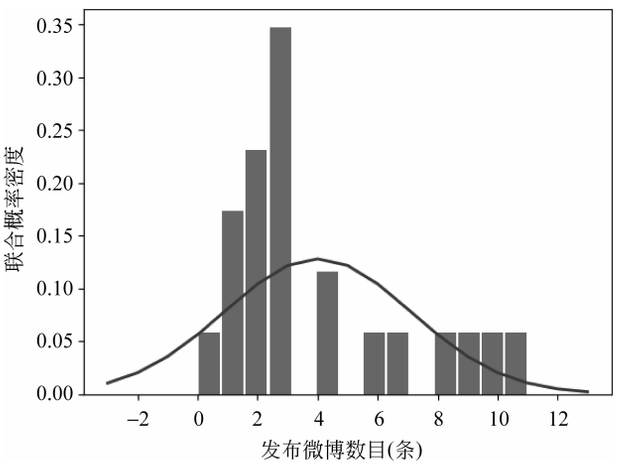

图3 正态分布拟合图

表1中，$X1$—$X5$是本文用来描述用户微博情绪的五个特征，对于传统的单高斯模型而言，这些特征需要一个一个被计算分析。单高斯模型可以在一定程度上检测一组数据的异常点。目前多数的研究也专注于单高斯模型或者二维高斯模型，几乎没有用三个甚至更多的变量来检测异常。本文通过将新浪微博用户的微博处理成五维的向量，也就是用五维的特征来表征一个用户的情感。单高斯分布将会被用于可视化用户微博的每一类情绪，多高斯模型则被用来解决异常情绪检测问题。假设用户的多维情感分布符合多元正态分布，多元高斯模型可以自动获取这些特征变量之间的联系，实现对这五维情绪的联合建模，避免了大量的计算工作。通过联合概率密度和设定合适的阈值[13]，量化了异常检测。

表1 微博五维情绪特征

变量	$X1$	$X2$	$X3$	$X4$	$X5$
特征	中性	开心	惊讶	伤心	生气

2 模型

2.1 异常检测

异常检测是从数据集合中检测异常样本。而实际生活中异常检测中的异常样本很少，传统的监督学习算法很难从这些异常样本中学习，所以大多数的异常检测方法都是基于非监督学习。目前主要有三种非监督的异常检测的方法：基于模型[14]，基于临近[15]和基于密度[16]。吴恩达教授曾以飞机发动机的异常检测[17]为例，阐述了基于密度的异常检测的基本原理：导致飞机发动机异常有多种因素。假设X＝发动机产生的热量，Y＝发动机的振动强度，给出一组数据$D=(D1,D2,...,Dn)$。由于这里发动机异常检测是基于两个变量来确定的，根据这两个变量，确定一个个坐标点(x,y)，这些数据点可以绘制在图上。如图4所示，椭圆表示的数据点密度大，可以被标记为正常；最右边的三角形标记的数据点明显偏离正常数据组，其密度比椭圆点小得多，也可看作是一个离群点(outlier)[18]。基于密度的方法来检测异常依据的是：低密度并且与邻近点相距较远的数据点，将被标记为异常点。本文基于此原理，计算出带情感的微博数据的分布密度，判断一组数据是否出现异常，进而判断该组数据对应的用户和月份是否异常。

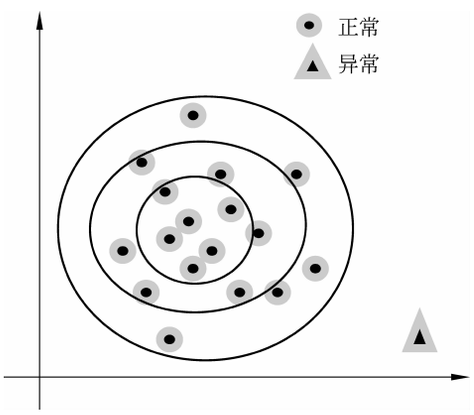

图4 异常检测案例

2.2 微博异常情绪检测

异常检测的方法有很多，目前主要是非参数[18]和非监督[19]的方法，本文选择多元高斯模型的原因是：首先，可以避免大量的数据标记和训练工作；其次，可以自动捕捉变量之间的不同特征之间的相关性；第三，实验中，微博情绪是五维的变量，而多元高斯可以对多元情绪很好地建模；最后，通过计算样本的联合概率密度，量化了异常检测。

给定数据集$\{(X_j^{(1)},X_j^{(2)},\cdots,X_j^{(m)}\}$，即一个$m\times n$的矩阵；计算样本的联合概率密度函数$P(x)$需要首先计算$\mu$和$\sum$，如式(1)～(2)所示。

$$\mu_j=\frac{1}{m}\sum_{i=1}^{m}X_j^{(i)} \tag{1}$$

$$\Sigma_j=\frac{1}{m}\sum_{i=1}^{m}[(x_j^{(i)}-\mu_j)(x_j^{(i)}-\mu_j)^T] \tag{2}$$

其中 m 是样本的个数，n 是变量的维数，j 从 1 到 n。μ(n 维) 是每一维向量的均值，由样本均值代替，$\sum$ 是协方差，由样本方差代替。假设给出一个测试样本 $x^{(k)}$(n 维的变量)，该样本的联合概率密度可以这样计算，如式(3)所示。

$$P(x;u,\Sigma)=\frac{1}{2\pi^{\frac{n}{2}}\mid\sum\mid^{\frac{1}{2}}}exp\left[-\frac{1}{2}(x^{(k)}-\mu)^{T}\sum^{-1}(x^{(k)}-\mu)\right] \quad (3)$$

图 5 是本文的用户微博异常情绪检测模型：

Step1　统计用户/月份 N 类微博

Step2　根据微博数据的 μ 和 $\sum$ 建模

Step3　计算联合概率密度值 p(x)

Step4　阈值选择，确定最优的阈值 ε

Step5　判断 p(x)＜ε 与否

Step6　p(x)＜ε，标记为异常，否则正常

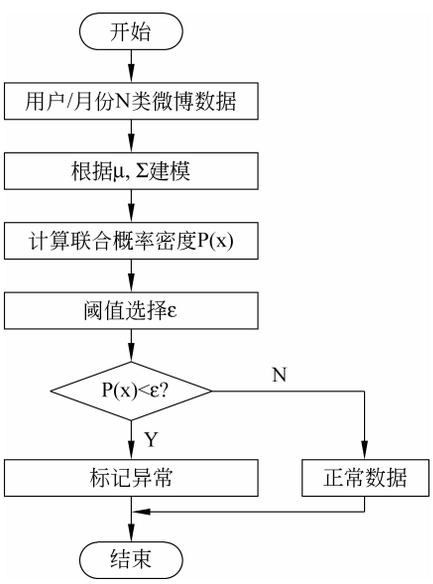

图 5　微博异常情绪检测模型

联合概率密度值是反应多个数据点中，当前样本数据点出现的概率，也是本文异常情绪检测的重要判断依据。联合概率密度值越小，说明数据出现频率越低，出现异常可能性越大。反之则较正常。如果某个用户/月份的联合概率密度值 p(x)＜ε，则被标记为异常用户/月份，阈值 ε 的选择将在 2.3 节阐述。

2.3　阈值选择

基于 1.1 节的微博数据，本文将 10 275 条微博处理成 1 700 个五维的数据集，每组数据集的联合概率密度值可以利用式(3)批量计算，减少了一部分时间。阈值的选择直接影响异常检测的准确率，阈值的选取是根据微博数据联合概率密度值的分布来确定，具体通过以下步骤实现：

Step1　按所有用户和所有月份得到五维数据集，批量计算其联合概率密度。

Step2　将所有数据集分为两部分：交叉验证集[20]和测试集[21]。

Step3　通过设置不同的阈值，在交叉验证集上进行实验，对获得的准确率进行比较。

Step4　选取交叉验证集最高准确率对应的阈值作为测试集上的阈值。

3　实验与分析

3.1　联合概率密度

表 2 是基于单个用户统计的联合概率密度值，以用户 12 为例，第 2 列到第 6 列是五种情绪的微博数目。从表 2 中可以看出，该用户大部分月份的微博情感联合概率密度值是从 1E-03 到 1E-04。但在 2013 年 5 月，该用户的微博数据联合概率密度值为 3.40E-06。偏离其他组，是一个离群点，被标记为异常。

表 2　基于用户的联合概率密度

用户 12	X_1	X_2	X_3	X_4	X_5	联合概率密度
Dec-15	2	0	0	0	0	8.15E-04
Nov-15	1	1	0	0	0	7.70E-03
Oct-15	1	0	0	0	2	3.79E-04
Sep-15	1	2	0	0	0	7.38E-03
Aug-15	1	4	1	0	1	5.01E-04
Jul-15	1	0	1	2	0	2.78E-04
Jun-15	1	2	0	0	0	7.38E-03
May-15	0	0	0	2	0	6.92E-04
Apr-15	0	1	0	1	0	5.40E-03
Mar-15	0	0	1	0	0	9.16E-04
Oct-14	0	1	0	0	0	9.18E-03
Jul-14	0	0	1	0	1	9.96E-04
Jun-14	0	1	0	0	0	9.18E-03
May-14	0	0	0	0	1	6.22E-03
Feb-14	2	0	0	0	1	1.54E-03

续表

用户 12	X_1	X_2	X_3	X_4	X_5	联合概率密度
Jan-14	0	1	0	1	1	3.10E-03
Sep-13	1	1	0	0	0	7.70E-03
Aug-13	1	2	0	0	0	7.38E-03
Jul-13	0	3	0	1	0	7.61E-04
Jun-13	3	2	1	3	2	1.06E-04
May-13	10	11	2	3	1	3.40E-06

图6是表2对应的用户12的微博原文本，从图中可以看出，用户在2013年5月，确实出现“不舍得，累，失眠”等异常情绪字眼。

表3是基于月份统计的联合概率密度结果，第2列到第6列是五种情绪的微博数目，可以看出2016年5月，user1，user 8，user 19，user 20这四个用户的数据联合概率密度值明显小于其他组，这些密度值远小于其他组的用户将被标记为疑似异常。

user12 洗个澡，明天拍照然后就要着手打广告卖了，真心不舍得呀…[泪]
user12 星期二 05/07 00:28:10 2013
user12 刚刚看完 nba 正在消亡的垃圾话..哈哈..球场精英骂垃圾话也是天才,小编很有才.http:/t.cn/zTYMObn
user12 星期一 05/06 19:59:20 2013
user12 打完球大晚上 ya 个越南牛肉粉 laska。不可以再饭前拍一个了，太 90 后了[汗]
user12 星期一 05/06 12:32:49 2013
user12[馋嘴]早你一个餐
user12 星期六 05/04 15:55:36 2013
user12 还有 20 分钟啊啊啊…今晚出去 city 吃个小饭然后迅速回家
user12 星期六 05/04 00:22:51 2013
user12 今天肯定是很累了，好好睡一觉
user12 星期五 05/03 04:26:25 2013
user12 天都亮了，真是的发神马呆想神马想，还没有睡着。。肚子都饿了，唉...日子不好过，度日如年啊啊啊。。

图6 验证异常情绪文本(用户12)

表3 基于月份(2016.5)的联合概率密度

用户	X_1	X_2	X_3	X_4	X_5	联合概率密度	用户	X_1	X_2	X_3	X_4	X_5	联合概率密度
user1	14	0	0	0	0	1.41E-05	user12	0	0	0	1	0	3.21E-03
user2	1	2	0	0	1	6.50E-04	user13	0	3	1	0	1	9.26E-04
user3	0	0	0	1	0	3.21E-03	user14	0	1	0	0	0	4.72E-03
user4	0	0	0	1	0	3.21E-03	user15	0	1	0	0	0	4.72E-03
user5	0	2	1	1	1	2.71E-03	user16	2	0	0	0	0	7.15E-03
user6	1	2	0	1	0	3.01E-03	user17	1	0	0	0	0	6.61E-03
user7	0	1	0	0	0	4.72E-03	user18	1	0	0	0	0	6.61E-03
user8	1	0	2	1	1	2.59E-05	user19	5	1	1	0	2	5.45E-05
user9	1	1	1	1	0	2.09E-03	user20	8	10	2	2	0	2.13E-06
user10	0	0	0	1	0	3.21E-03	user21	2	0	0	0	0	7.15E-03
user11	7	1	0	0	0	3.21E-03							

图7是表3对应的异常微博文本，可以看出2016年5月，用户19，20出现“心酸，可怜，忘掉过去，甘愿放弃”等异常情绪状态字眼。

3.2 异常检测准确率

为了提高异常检测的准确性，本文将1 700组数据集继续划分为500组交叉验证集和1 200组测试集。根据批量计算得到所有数据集的联合概率密度，通过观察和分析，选择三个比较合适的阈值放在交叉验证集上实验。不同的阈值将得到不同的实验精度，而精度最高的对应的阈值将被选择出来用于测试集继续实验。

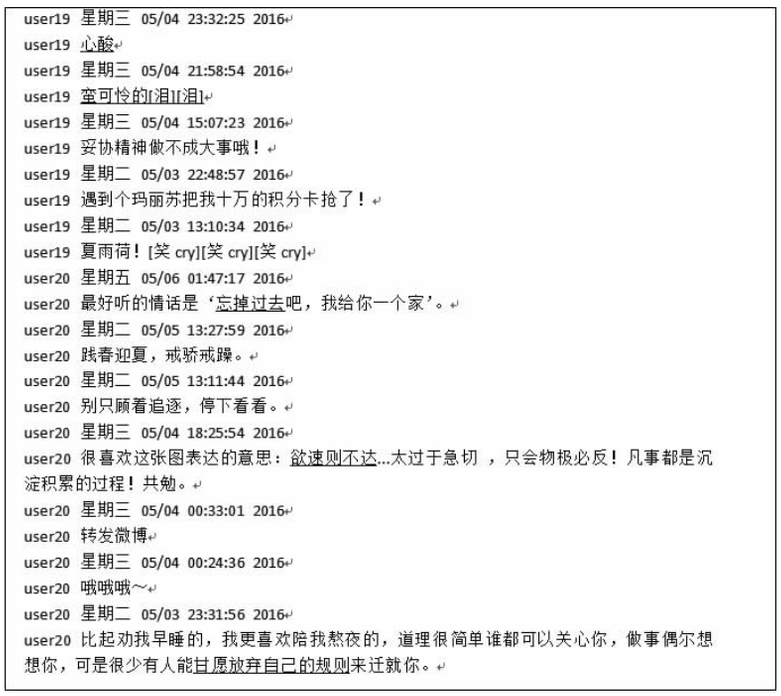
user19 星期三 05/04 23:32:25 2016
user19 心酸
user19 星期三 05/04 21:58:54 2016
user19 蛮可怜的[泪][泪]
user19 星期三 05/04 15:07:23 2016
user19 妥协精神做不成大事哦！
user19 星期二 05/03 22:48:57 2016
user19 遇到个玛丽苏把我十万的积分卡抢了！
user19 星期二 05/03 13:10:34 2016
user19 夏雨荷！[笑 cry][笑 cry][笑 cry]
user20 星期五 05/06 01:47:17 2016
user20 最好听的情话是‘忘掉过去吧，我给你一个家’。
user20 星期二 05/05 13:27:59 2016
user20 践春迎夏，戒骄戒躁。
user20 星期二 05/05 13:11:44 2016
user20 别只顾着追逐，停下看看。
user20 星期三 05/04 18:25:54 2016
user20 很喜欢这张图表达的意思：欲速则不达...太过于急切 ，只会物极必反！凡事都是沉淀积累的过程！共勉。
user20 星期三 05/04 00:33:01 2016
user20 转发微博
user20 星期三 05/04 00:24:36 2016
user20 哦哦哦～
user20 星期二 05/03 23:31:56 2016
user20 比起劝我早睡的，我更喜欢陪我熬夜的，道理很简单谁都可以关心你，做事偶尔想想你，可是很少有人能甘愿放弃自己的规则来迁就你。

图 7　验证异常情绪文本(2016 年 5 月)

表 4 是基于用户的阈值选择，选取三个较合适的阈值在交叉验证集上进行实验：1E-04，4E-05，1E-05，当阈值是 4E-05 时，此时异常检测准确率最高，为 88.89%；表 5 是基于月份的阈值选择，选取了三个较合适的阈值在交叉验证集上进行实验：1E-05，1E-06，1E-07，从表 5 可以看出当阈值是 1E-06 时，此时异常检测准确率最高，为 88.89%。从表 4 和表 5 也可以看出，阈值设定的越小，准确率不一定越高，阈值的选取是根据不同统计的结果来进行初步观察，交叉验证和筛选，最后选择合适的结果。本文通过交叉验证，最后基于用户和月份的异常检测选取的阈值分别是 4E-05 和 1E-06。

表 4　基于用户的阈值选择

阈值	验证数目	真阳	假阳	准确率/%
1E-04	36	29	7	80.56
4E-05	36	32	4	88.89
1E-05	36	27	9	75.00

表 5　基于月份的阈值选择

阈值	验证数目	真阳	假阳	准确率/%
1E-05	36	30	6	83.33
1E-06	36	32	4	88.89
1E-07	36	31	5	86.11

表 6 是基于用户的异常情绪检测结果，基于不同的用户统计得到的结果可以判断出一个用户在哪一段时间内情绪出现异常。从表 6 可以看出，实验中 109 个数据集被标记为异常，通过和原始的微博情绪(半自动标记结果)比较，其中 91 条数据集是真阳性(检测正确)异常，18 条为假阳性(检测错误)，最终的准确率是 83.49%。表 7 是基于不同的月份的异常情绪检测结果，基于不同的月份统计得到的结果可以判断出一段时间内哪些用户出现异常情绪。从表 7 可以看出，74 条数据集被标记为异常，其中 65 条数据集是真阳性，9 条为假阳性，最终的准确率是 87.84%。

表 6　基于用户的异常情绪检测结果

阈值	测试数据	真阳	假阳	准确率/%
4E-05	109	91	18	83.49

表 7　基于月份的异常情绪检测

阈值	测试数据	真阳	假阳	准确率/%
1E-06	74	65	9	87.84

为了更好地对提出的模型进行评估，本文选取了前人相关的典型工作来进行比较，目前针对社交媒体上的用户情绪进行检测的研究较少，而针对一段时间内异常的检测的相关研究较多。如图 8 所示，基于 NMF(nonnegative matrix factorization)的

方法准确率较低，约为51%。基于微博的事件实时监测方法[8]达到了73.33%的准确率，基于SSDM[22]的检测方法准确率是85.20%，相对于前者有较大提高。但本方法是针对Twitter文本的垃圾信息检测，与中文微博的异常检测有一定不同。本文通过引入联合概率密度参数对异常检测进行量化，得到异常情绪检测的准确率为87.84%，取得了一定进步。由于网络事件异常多样，特征不明显，且具有意外性和偶发性，目前，异常检测的结果均不是太高，需要进一步研究。

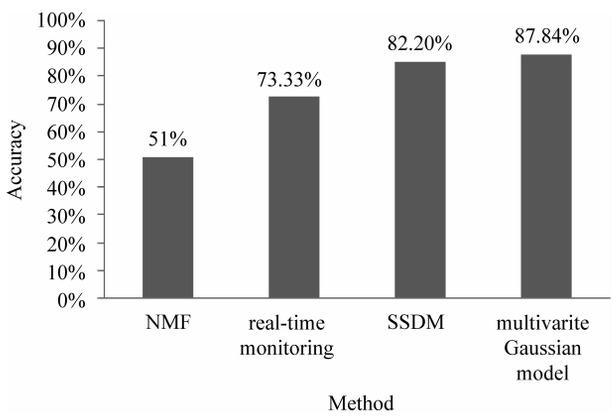

图8 准确率比较

3.3 用户情绪建模

单个用户的微博异常情绪检测可以通过对该用户的微博数据进行分析，群体用户的异常情绪检测则可以通过对一段时间内所有用户的微博数据进行建模分析。假设用户每个月的五类情绪数据可以看成一个五维矩阵。假设每一维相互独立且服从高斯分布，可以使用K-S(Kolmogorov-Smirnov)检验[23]对每一维的数据是否服从高斯分布进行检测。原假设：检验分布为正态分布。

K-S检验表格中的VAR1-VAR5分别代表“中性、开心、惊讶、伤心、生气”这五个情绪变量，N代表参与检测的样本数目。评价指标是即渐近显著性(双侧)值(P-value)，一般P-value设定为0.05，当P-value >0.05时，不能拒原假设，即可以认为数据服从正态分布；反之拒绝原假设，即认为数据不服从正态分布。

表8 单个用户微博情绪数据的K-S检验

Kolmogorov-Smirnov 检验						
		VAR1	VAR2	VAR3	VAR4	VAR5
N		20	20	20	20	20
正态参数 a,b	均值	1.400 0	1.800 0	0.400 0	0.750 0	0.500 0
	标准差	2.500 53	1.399 25	0.680 56	0.850 70	0.688 25
最极端差别	绝对值	0.288	0.157	0.422	0.261	0.366
	正	0.255	0.151	0.422	0.261	0.366
	负	−0.288	−0.157	−0.278	−0.189	−0.234
Kolmogrov-Smirnov Z		1.287	0.701	1.886	1.167	1.638
渐近显著性(双侧)		0.073	0.709	0.002	0.131	0.009

注：a. 检验分布为正态分布。b. 根据数据计算得到。

由表8可以看出“中性、开心、伤心”情绪的P>0.05，则不能拒绝原假设，这也符合前面的假设。用户大部分正常的情绪是近似符合正态分布，而生气等情绪不符合正态分布。在对单个用户高斯建模的基础上，本文研究了群体用户的微博情绪模型，对群体的情绪分布是否满足高斯分布进行了检测：表9是100位用户其中一个月的微博情绪K-S检测结果，发现P=0<0.05，即拒绝原假设。接着对100位微博用户的60个月的微博数据进行检测，发现P均小于0.05，进一步说明群体情绪并不服从正态分布。

通过实验，本文得出如下推论：

a. 微博的“中性、开心、伤心”情绪可近似为正态分布，其中“开心”类最具代表性。

b. “惊讶”和“生气”这两类情绪数据稀疏，且具有爆发性，也就是用户异常情绪，不服从正态分布。

c. 群体的情绪不满足正态分布，它更趋向于另一种指数分布：“幂律分布”[24]。

推论a，b在上述实验中已得到验证，为了验证推论c，即群体的微博情绪满足幂律分布，本文给出了如图9所示的检验过程。幂律分布的检验主要是使用matlab进行数据拟合，原始数据row data接近

长尾分布(long-tailed)形状，经过取对数得到的 log data 近似一条直线，最后用“残差和”对原始数据进行计算并评估分布的合理性。“残差和”越小，可以认为该组数据越服从幂律分布。图 9 的数据得出残差和为 0.03，可以认为该组数据近似服从幂律分布，为了实验的准确性，本文对群体用户所有月份的情绪分布进行幂律分布检验，均证实了推论的正确性，即群体用户的微博情绪满足“幂律分布”。

表 9　群体用户微博情绪数据的 K-S 检验

Kolmogorov-Smirnov 检验						
		VAR1	VAR2	VAR3	VAR4	VAR5
N		100	100	100	100	100
正态参数 a,b	均值	2.430 0	3.710 0	0.740 0	1.160 0	0.410 0
	标准差	2.567 30	4.399 94	1.040 78	1.323 60	0.697 69
最极端差别	绝对值	0.307	0.259	0.311	0.368	0.402
	正	0.307	0.245	0.311	0.368	0.402
	负	−0.289	−0.259	−0.239	−0.212	−0.278
Kolmogrov-Smirnov Z		3.065	2.590	3.114	3.681	4.016
渐近显著性(双侧)		0.000	0.000	0.000	0.000	0.000

注：a. 检验分布为正态分布。b. 根据数据计算得到。

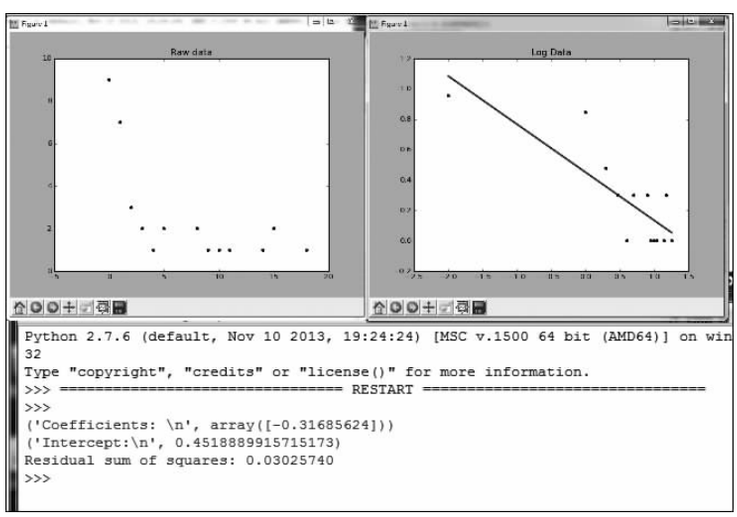

图 9　幂律分布检验

4　总结

本文结合多元高斯模型与联合概率密度，对用户微博异常情绪进行检测，从用户和时间两个角度统计并分析，将联合概率密度值作为用户异常情绪的判断指标，量化了异常检测。结合多元高斯分布和幂律分布对单个用户和群体用户的情绪进行建模。实验结果表明，按照用户和月份进行的异常检测准确率分别为 83.49% 和 87.84%。实验还通过正态分布检验，发现单个用户的“中性、开心、伤心”情绪满足正态分布，而“惊讶、生气”情绪由于具有爆发性，不服从正态分布。通过对单个用户和群体用户的微博数据进行分析，发现群体的情绪总体服从幂律分布，而单个用户的情绪则不服从，与市场规律吻合。本文的不足之处在于当前的实验数据比较稀疏，当数据足够多时，可以对用户一个月、一周甚至一天的情绪进行建模，并检测异常情绪。本文提出了一个比较完整的基于社交媒体的用户情绪建模和异常情绪检测模型，该模型对于预防大众心理问题以及检测公共安全有一定意义，进一步还可以帮助企业根据用户对产品的评价和情感倾向，来做出正

确的决策。

参考文献

[1] 微博数据中心：新浪微博. 2016 微博用户发展报告[R/OL]. [2017-2-8]. http://www.useit.com.cn/thread-14392-1-1.html

[2] 孙晓,何家劲,任福继. 基于多特征融合的混合神经网络模型讽刺语用判别[J]. 中文信息学报, 2016, 30(6): 215-223.

[3] Yang F, Liu Y, Yu X, et al. Automatic detection of rumor on Sina Weibo[C]//Proceedings of the ACM, 2012: 1-7.

[4] 何跃,邓唯茹,张丹. 中文微博的情况识别与分类研究[J]. 情报杂志,2014,2: 136-139.

[5] Lin H, Jia J, Qiu J, et al. H. Detecting Stress Based on Social Interactions in Social Networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2017,(99): 1.

[6] Zhang J,Zhu B,Liang L,et al. Recognition and classification of emotions in the Chinese microblog based on emotional factor[J]. Beijing Daxue XuebaoZiran Kexue Ban/acta Scientiarum Naturalium Universitatis Pekinensis, 2014,50(1): 79-84.

[7] Wang Z, Joo V, Tong C, et al. Anomaly Detection through Enhanced Sentiment Analysis on Social Media Data[C]//Proceedings of the IEEE International Conference on Cloud Computing Technology and Science. IEEE, 2014: 917-922.

[8] 李凌云. 基于微博的事件实时监测框架与系统[D]. 北京邮电大学硕士学位论文, 2014.

[9] Yin G, Zhang Y, Dong Y, et al. A boost factor based detection method for abnormal rank of microblogging [J]. Journal of Harbin Engineering University, 2013, 34(4): 488-493.

[10] 孙晓, 叶嘉麒, 唐陈意,等. 基于多策略的新浪微博大数据抓取及应用[J]. 合肥工业大学学报自然科学版, 2014(10): 1210-1215.

[11] Chang C C, Lin C J. LIBSVM: A library for support vector machines[J]. Acm Transactions on Intelligent Systems & Technology, 2007, 2: 27.

[12] Yuan S F, Wang S T. Multi-classification method applied to face recognition based on mixed Gaussian distribution[J]. Application Research of Computers, 2013, 30(9): 2868-2871.

[13] Diehl P U, Neil D, Binas J, et al. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing[C]//Proceedings of the International Joint Conference on Neural Networks. 2015: 1-8.

[14] Liang J, Du R. Model-based Fault Detection and Diagnosis of HVAC systems using Support Vector Machine method[J]. International Journal of Refrigeration, 2007, 30(6): 1104-1114.

[15] Idé T, Lozano A C, Abe N, et al. Proximity-Based Anomaly Detection using Sparse Structure Learning [J]. SDM, 2009: 97-108.

[16] Ma S H, Wang J K, Liu Z G, et al. Density-Based Distributed Elliptical Anomaly Detection in Wireless Sensor Networks[J]. Applied Mechanics & Materials, 2012, 249-250: 226-230.

[17] Andrew Ng: Machine Learning. Week9, Anomaly Dection [EB/OL]. [2017-4-21]. https://www.coursera.org/learn/machine-learning

[18] Dang X, Serfling R. Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties[J]. Journal of Statistical Planning & Inference, 2010, 140(1): 198-213.

[19] Huang G, Song S, Gupta J N, et al. Semi-supervised and unsupervised extreme learning machines[J]. Cybernetics. 2014, 44(12): 2405.

[20] Chen K, Lei J. Network Cross-Validation for Determining the Number of Communities in Network Data [J]. Journal of the American Statistical Association, 2014, 178(5): 410.

[21] Yu H, Yang J, Han J, et al. Making SVMs Scalable to Large Data Sets using Hierarchical Cluster Indexing[J]. Data Mining and Knowledge Discovery, 2005, 11(3): 295-321.

[22] Hu X, Tang J, Zhang Y, et al. Social spammer detection in microblogging[C]//Proceedings of the International Joint Conference on Artificial Intelligence. 2013: 2633-2639.

[23] Eghbali H J. K-S Test for Detecting Changes from Landsat Imagery Data[J]. IEEE Transactions on Systems Man & Cybernetics, 1979, 9(1): 17-23.

[24] AaronClauset A, Shalizi C R, Newman M E J. Power-Law Distributions in Empirical Data[J]. Siam Review, 2007, 51(4): 661-703.

孙晓(1980－)，博士，副教授，主要研究领域为自然语言处理，智能人机会话及相关机器学习算法的研究与开发。
E-mail：sunx@hfut.edu.cn

任福继(1959—)，博士，教授，主要研究领域为自然语言处理，人工智能，语言理解与交流，情感计算等。
E-mail：ren2fuji@gmail.com

张陈(1993—)，硕士，主要研究领域为自然语言处理，微博情感分析，异常检测.神经网络等。
E-mail：1725685823@qq.com

(上接第 86 页)

[12] Lü L, Zhou T. Link prediction in complex networks: A survey [J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150-1170.

[13] Lao N, Cohen W W. Relational retrieval using a combination of path constrained random walks[J]. Machine Learning, 2010, 81(1): 53-67.

[14] Vrande či ć D, Krötzsch M. Wikidata: A free collaborative knowledgebase [J]. Communications of the ACM, 2014, 57(10): 78-85.

[15] Miller G A. WordNet: A lexical database for English [J]. Communications of the ACM, 1995, 38(11): 39-41.

[16] Bollacker K, Cook P, Tufts, P. Freebase: A shared database of structured general human knowledge [C]//Proceedings of the 21st AAAI Conference on Artificial Intelligence, 2007(7): 1962-1963.

张宁豫(1989—)，博士，主要研究领域为语义挖掘。
E-mail：zhangningyu@zju.edu.cn

陈矫彦(1988—)，博士，主要研究领域为语义挖掘。
E-mail：jiaoyanchen@zju.edu.cn

陈曦(1990—)，博士，主要研究领域为语义挖掘。
E-mail：xichen@zju.edu.cn