基于词向量的多维度正则化 SVM 社交网络抑郁倾向检测方法

王 垚 贾宝龙 杜依宁 张 晗 陈 响
(北京世相科技文化有限公司 北京 100102)

摘 要 针对目前抑郁症的诊断方式单一、诊断率低等问题,提出一种基于词向量的多维度正则化 SVM 社交网络抑郁倾向检测方法。通过人工标注获得训练数据,并请心理学硕士对数据进行验证,确保数据的可用性。在预处理阶段,统计得到常用的抑郁词,使用腾讯词向量进行文本向量化及用户向量化,在构建向量的过程中加入 TF-IDF 和抑郁词权重因子;在训练阶段,通过将情感、性别和发微博频率加入传统 SVM 的目标函数中,构建多维度正则化 SVM 模型。多组对比实验结果表明,该方法能够有效检测抑郁倾向。

关键词 抑郁倾向 微博 支持向量机 词向量

中图分类号 TP391.1 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2022.03.019

DEPRESSION DETECTION OF MULTI-DIMENSIONAL REGULARIZED SVM SOCIAL NETWORK BASED ON WORD VECTOR

Wang Yao Jia Baolong Du Yining Zhang Han Chen Xiang
(*Beijing Shixiang Technology Culture Co., Ltd., Beijing 100102, China*)

Abstract Aiming at the single diagnosis method and low diagnosis rate of current depression diagnosis, we proposes a multi-dimensional regularized SVM based on word vectors to detect depression tendency. It manually labelled the training data and asked the experts to verify the data. In the pretreatment stage, we got the dictionary of the commonly used depression words, constructed the text vectors and user vectors by Tencent word vectors, added TF-IDF and depression word weighting factor to the vectors. In the training phase, we added emotion, gender and frequency to the objective function of traditional SVM to construct a multi-dimensional regularized SVM. The experimental results show that the proposed model can predict the depression tendency of bloggers effectively.

Keywords Depression tendency Sina Weibo Support vector machine Word vector

0 引 言

微博是一种开放化的互联网社交服务,人们可以通过微博分享自己的心情、经历或故事。微博提供的评论、超级话题等功能,使人们能快速找到志同道合的朋友。海量的微博文本中蕴含着大量的情感。微博的文本内容成为抑郁倾向检测的主要数据来源之一。

国内外对于社交媒体文本内容的情感分析方法主要包括统计学方法和机器学习方法。统计学方法通过统计高频词,构建情感词典来分析文本内容的情感倾向。高一虹等[1]基于数据统计来分析抑郁症患者在现实生活中和社交媒体上的表现,发现抑郁症患者在社交媒体上发微博的频率更高,微博的文本内容中的负向情感更明显。林晔[2]对当时引起巨大轰动的“走饭”和“醒醒我们回家了”两个微博账号进行了统计分析,发现在实施自杀前,抑郁患者会反复、频繁地表达自己的抑郁、痛苦和自杀意图,纠结于生死之间。虽然基于统计的方法能够一定程度上分析出微博用户的情感,但是忽略了用户信息,并且过分依赖分词的好坏,因此不能准确地评价用户的抑郁倾向。

基于机器学习的方法是通过将微博文本、博主简

收稿日期:2020-01-08。**王垚**,工程师,主研领域:自然语言处理,语音识别。**贾宝龙**,工程师。**杜依宁**,工程师。**张晗**,高工。**陈响**,高工。

介和博主标签等特征抽象为向量,构建分类器进行训练。施志伟等[3]通过问卷调查得到有抑郁倾向的用户,获取他们的微博文本数据,使用支持向量机模型进行有监督学习,准确率达到 82.35%。但是其训练数据单一,只考虑了微博文本的内容,没有考虑发博人的性别、情感等因素。为了考虑更多的有效信息,Peng 等[7]增加了发博人简介、发博人行为等特征,对比了传统支持向量机、朴素贝叶斯、决策树和 K-近邻等算法后,提出一种多元支持向量机模型,准确率达到了 83.5%,明显高于其他几种分类算法,但由于数据量较少,模型的泛化能力不足。Hao 等[8]提出了一种基于两种分类器的检测方法,首先训练朴素贝叶斯分类器,并生成一个抑郁患者的常用词词典,然后使用线性分类器加入更多的特征,得到了准确率较高的分类器。方振宇[9]提出了基于 Word2vec 词向量的神经网络分类模型,将用户情绪向量与微博内容向量进行拼接作为用户特征向量,准确率达到了 86.5%,但是忽略了用户的个人属性信息。为了解决上述存在的问题,本文在使用微博文本作为样本特征的基础上,将用户的情感、性别和发博频率融入到 SVM 的目标函数中,提出了一种基于词向量的多维度正则化 SVM 的社交网络抑郁倾向检测方法,并通过多组对比实验验证了该方法的有效性。

1　相关工作

1.1　抑郁症

抑郁症[11]是一种心理障碍或情感障碍,是最常见的精神疾病之一,主要表现为兴趣减退、认知功能受损和情绪紊乱。据统计,抑郁症患者的终身患病率为 13.2%[12],大约有 25% 的女性患过抑郁症,大约有 10% 的男性患过抑郁症[13]。由于基层医疗机构对抑郁症的认识不充分,仍存在着普遍的一高两低现象,即高患病率、低诊断率、低治愈率。

1.2　数据的收集

使用的数据来自新浪微博,选择 352 位有明显抑郁倾向的博主的 35 962 条微博文本作为正数据,323 位非抑郁症患者博主的 72 697 条微博文本作为负数据。筛选后得到 28 654 条微博文本的正数据,58 569 条微博文本的负数据。经过 3 位心理学系的硕士研究生进行交叉检验,仅有 10 位用户存在争议,说明数据的可信度较高。

1.3　数据的清洗

微博内容数据形式多样,包含大量“脏”数据,所以需要对数据进行清洗,通过人工观察或统计发现主要有以下形式的“脏”数据:(1) 非文本信息(图片和视频等);(2) 广告数据以及非原创数据(文本中包括投票、打榜、影响力和人气演员等);(3) 部分干扰字符(@xxx,#xxx 超话#等);(4) 长度小于 7 个字的微博文本;(5) 不规范表达方式(emoji 表情、颜文字等)。

清洗前和清洗后的数据如表 1 所示。

表 1　数据展示

数据类型	清洗前数据		清洗后数据	
	正数据	负数据	正数据	负数据
用户数	352	323	352	323
微博数	35 962	72 697	28 654	58 569

2　抑郁倾向检测方法

本文提出的抑郁倾向检测方法主要包括两部分,分别为构建用户向量、构建多维度正则化 SVM,如图 1 所示。

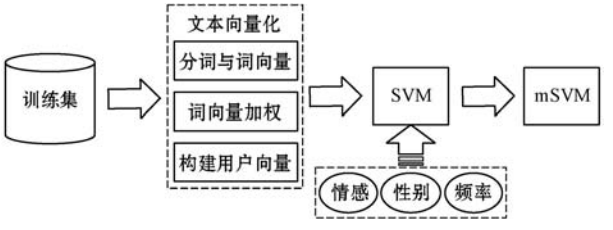

图 1　多维度正则化 SVM 检测模型

首先,微博文本的向量化主要包括:(1) 对微博文本的分词;(2) 获取每个词语的腾讯词向量;(3) 将词向量进行特征加权计算句向量;(4) 根据句向量构建用户向量。然后,进行多维度正则化 SVM 的有监督学习。

2.1　构建用户向量

2.1.1　分词与词向量

腾讯词向量是腾讯 AI 实验室在 2018 年开源的一个大规模、高质量的中文词向量数据集。该数据集在多个方面较现有公开数据集均有改善。在覆盖率上,该数据集包含了超过 800 万的中文词汇,覆盖了更多的短语,包含了近两年的网络用语。在准确性上,该训练算法使用的是腾讯自研的 Directional Skip-Gram(DSG)算法[14],它改进了被广泛使用的如 Word2vec 词向量模型中的词向量训练算法 Skip-Gram(SG)[15],在文本窗口中词对共现关系的基础上,加入了词对的相对位置的考量,以此提高词向量语义表示的准确性。

所以用它来作为微博内容分词后每个词的词向量[18]是合理有效的。

由于微博文本包含大量网络用语,而百度分词比较善于针对网络文本进行分词,同时也能通过构建自定义词典提高特殊词汇的分词效果,所以,首先利用百度分词 API 进行分词,然后获得对应的腾讯词向量。对于腾讯词向量库中不存在的抑郁词,则选择腾讯词向量库中与其最相近的词作为替代。对于不在抑郁词典中且腾讯词向量未收录的词语,将其赋值为 0 向量,便于之后的计算。

2.1.2 构建微博文本向量

首先使用 TF-IDF[19]进行特征加权。特征权重 W_{ij} 的计算式为:

$$W_{ij} = TF_{ij} \cdot IDF_{ij} \tag{1}$$

式中:TF_{ij}表示特征词 ω_i在文本 d_j中出现的次数,IDF_{ij}表示特征词 ω_i的逆文档频率。为了能够一定程度上增强抑郁词权重,IDF_{ij}通过大规模的微博文本数据集计算。

为了提升抑郁词对于整条微博文本的影响,赋予抑郁词相对较大的权重值,赋予非抑郁词权重 1。即加权后的词向量表示为:

$$\boldsymbol{V}_i = \boldsymbol{TV}_i \cdot W_{ij} \cdot W_d \tag{2}$$

式中:$\boldsymbol{TV}_i$表示该词的腾讯词向量,W_{ij}表示该词的 TF-IDF 值,W_d表示该词的抑郁词权重。

2.1.3 构建用户向量

根据 2.1.2 节得到的加权词向量,通过对应维度求均值的方式计算整条微博文本的向量表示,该向量表示为:

$$c_d = (x_{d1}, x_{d2}, \cdots, x_{dt}) \tag{3}$$

式中:x_{d1}表示当前文本中所有词向量第 1 维的均值。因为腾讯词向量的纬度是 200,所以由此得出的文本向量也是 200 维,进而可得到用户的矩阵表示为:

$$\boldsymbol{M}_i = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1t} \\ x_{21} & x_{22} & \cdots & x_{2t} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nt} \end{bmatrix} \tag{4}$$

式中:n 表示用户的微博总条数。最后,通过将矩阵 $\boldsymbol{M}_i$按行求均值得到用户的向量。

2.2 多维度正则化

支持向量机[20]是一种优秀的机器学习分类模型,在面对非线性以及高维度分类问题上,效果比其他二分类方法更好,因为 SVM 能够接受高维特征空间和稀疏特征向量,所以在文本分类上有很好的效果。面对微博文本分类这个非线性问题,直接利用线性化的 SVM 是无法分类的,所以将决策函数的限制条件进行一定的放松,使它对于一些异常或极端样本点有一定容错空间,SVM 模型表示为:

$$y_i \cdot [(W^{\mathrm{T}} \cdot x_i) + b] \geqslant 1 - \xi_i \quad 1 \leqslant i \leqslant N, \xi \geqslant 0 \tag{5}$$

式中:ξ_i为松弛变量。由于任意样本都有松弛变量值与之对应,当松弛变量取任意大数时,限制条件在非线性可分数据上即可满足,但是这样无法得到最优的超平面。于是需要在原来的最大化间隔,即最小化$\frac{1}{2}||\omega||^2$函数上添加条件来限制松弛变量,使得二者平衡,得到最优超平面。如此,目标函数变为:

$$\min\left(\frac{1}{2}||\omega||^2 + C\sum_{i=1}^{N}\xi_i\right) \tag{6}$$

式中:C 用来控制 $\sum_{i=1}^{N}\xi_i$ 的影响力。

由于文本在经过前期词向量信息累加的处理后,所得到的数据的维数已经较高,所以还需要进行变换,这也是 SVM 的一个优势,它通过构造可以将已有数据 x 映射到高维空间 H 的映射函数,即 $\phi(x_i)$。因为此类映射的维度理论上是可以无限维的,无法显式求出,所以 SVM 引入核函数[21]来实现不需要知道映射向量就可以实现分类的目的。核函数形式如下:

$$k(x_i, x_j) = \varphi(x_i)^{\mathrm{T}}\varphi(x_j) \tag{7}$$

这里通过高斯核函数实现同等映射:

$$k(x, z) = \mathrm{e}^{\frac{-||x-z||^2}{2\sigma^2}} \tag{8}$$

通过将式(6)转化为对偶问题的方式,利用 KKT 条件,构造拉格朗日函数,求得最终的分类函数如下:

$$f(x) = \sum_{i=1}^{N}\alpha_i y_i k(x_i, x) + b \quad 0 < \alpha_i < C \tag{9}$$

$$b = \frac{1 - y_i\sum_{j=1}^{N}\alpha_j y_j k(x_i, x_j)}{y_i}$$

式中:α_j是拉格朗日乘子;x 表示待分类文本。

经过前期相关研究工作,发现有抑郁倾向的用户存在以下明显特征:(1) 发微博频率明显高于正常用户;(2) 有明显消极情感;(3) 女性人数明显高于男性用户,比例大致为 3∶1。因此将用户发微博频率、用户文本情感和性别特征加入到目标函数中,使 SVM 学习到的超平面更加准确,因此在原本的目标函数上增加一项由发微博频率、情感和性别组成的正则项,表示为:

$$\omega_i = \omega_e \cdot e_i + \omega_s \cdot s_i + \omega_f \cdot f_i \tag{10}$$

式中:e_i表示用户的负向情感概率;s_i表示用户的性别分数;f_i表示用户的发微博频率分数。因此,改进后的目标函数为:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i k(x, x_i) + b + W\omega_i \quad 0 < \alpha_i < C \quad (11)$$

式中:W 表示 ω_i 的影响力权重。

3　实　验

3.1　实验设计及评价标准

实验包括以下四种算法:(1) 使用腾讯词向量训练 SVM;(2) 使用腾讯词向量训练 mSVM;(3) 用 TF-IDF 加权词向量训练 mSVM;(4) 使用 TF-IDF 和抑郁词加权词向量训练 mSVM。为了便于描述,算法 1 用 SVM 表示,算法 2 用 mSVM 表示,算法 3 用 mSVM-T 表示,算法 4 用 mSVM-TW 表示。

在四种算法上进行 3 组对比实验,分别为:(1) 随着迭代次数准确率的变化趋势;(2) 随着迭代次数召回率的变化趋势;(3) 随着迭代次数 F1 值的变化趋势。准确率、召回率和 F1 值的计算公式如下:

$$Accuracy = \frac{TP + FN}{TP + FP + FN + TN} \quad (12)$$

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

$$F1 = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \quad (15)$$

式中:TP 表示真例判断为正样本;FP 表示假例判断为正样本;FN 表示假例判断为负样本;TN 表示真例判断为负样本。

3.2　实验结果与分析

为了能够更准确地反映四种算法的准确率、召回率和 F1 值随着迭代次数的变化情况,在当前迭代次数下的准确率、召回率和 F1 值均为独立训练 10 次取均值。四种算法的准确率随迭代次数的变化趋势如图 2 所示。

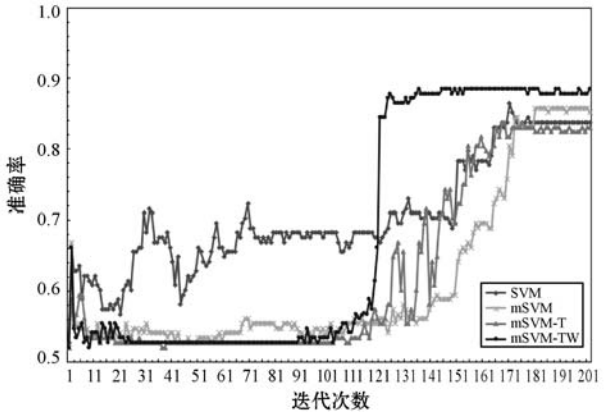

图 2　准确率变化趋势图

由图 2 可看出,mSVM-TW 在 140 次迭代后收敛,达到最优值 0.89 ±0.05。SVM 和 mSVM-T 在 170 次迭代后收敛,分别达到 0.83 ±0.05 和 0.82 ±0.05,mSVM 在 180 次迭代后收敛,达到 0.85 ±0.05。在收敛速度和最优值上,mSVM-TW 均明显优于其他三种算法,主要原因有两点,一是输入向量通过 TF-IDF 和抑郁词加权,改变了原始数据分布,使得数据的分布对于当前的任务更加清晰,因此更容易被分类;二是通过情感、性别和发博频率使得目标函数的损失变得更小,因此收敛速度更快。

召回率随迭代次数的变化趋势如图 3 所示,从收敛速度和最优值,mSVM-TW 也明显优于其他三种算法。mSVM-TW 的最优召回率达到 0.86 ±0.05。在迭代次数较低时,召回率异常偏高,甚至达到 1.0。这是由于当迭代次数较低时,分类器处于欠拟合状态,此时分类器将所有样本判断为正样本,因此召回率会异常高。随着迭代次数的增加,处于分类超平面较近的真负或假正样本逐渐增多,因此召回率逐渐下降,并趋于稳定。

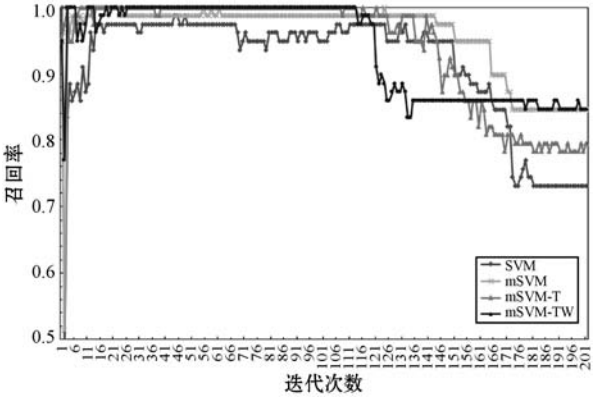

图 3　召回率变化趋势图

F1 值随迭代次数的变化趋势如图 4 所示,从收敛速度和最优值来看,mSVM-TW 也明显优于其他三种算法,最优 F1 值达到 0.89 ±0.05。

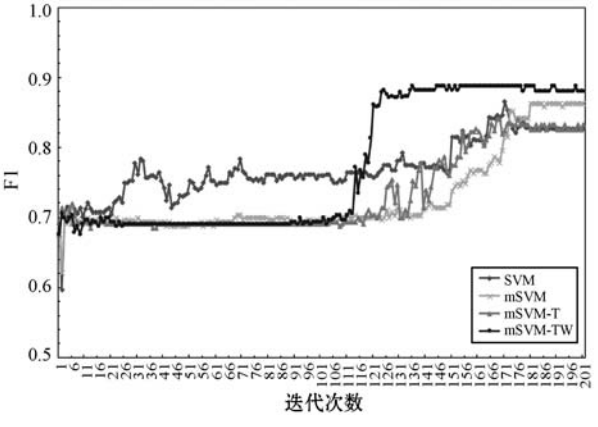

图 4　F1 值变化趋势图

综合上述实验结果,mSVM-TW 在各评价指标上均有较大提升,说明通过词向量加权和多种特征的正则化能够有效提升传统 SVM 在抑郁倾向检测任务上的分类性能。

4 结 语

本文提出的基于词向量的多维度正则化 SVM 方法,由于在传统 SVM 的损失函数中融入情感、性别和发微博频率,所以在 SVM 的监督学习过程中,能够根据用户的多种特征约束损失函数,使得学习到的分类超平面更加准确,泛化能力更强。因此,对于那些文本特征不够明显的用户也能较好地分类。

由于微博内容的形式具有多样性,除了文本,还有图片、视频、音频等,所以只考虑微博的文本内容会丢失用户的大量有效信息。因此,下一步考虑加入用户更多的有效信息,构建多模态的抑郁倾向检测模型,进一步增强模型的性能。

参 考 文 献

[1] 高一虹,孟玲. 自杀倾向的话语表述——大学生“走饭”微博分析[J]. 外语与外语教学,2019(1):43 - 55,145 - 146.

[2] 林晔. 浅析网络背景下媒体对抑郁症患者的形象建构——基于新浪微博文本的考察[J]. 新西部,2018(17):87 - 89.

[3] 施志伟,高俊波,胡雯雯,等. 基于文本的抑郁情感倾向识别模型[J]. 计算机系统应用,2017,26(12):155 - 159.

[4] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP),Philadelphia,2002:79 - 86.

[5] Youn S J, Trinh N H, Shyu I, et al. Using online social media, Facebook, in screening for major depressive disorder among college students[J]. International Journal of Clinical and Health Psychology, 2013, 13(1): 74 - 80.

[6] 李鹏宇. 微博社交网络中的学生用户抑郁症识别方法研究[D]. 哈尔滨:哈尔滨工业大学,2014.

[7] Peng Z, Hu Q, Dang J. Multi-kernel SVM based depression recognition using social media data[J]. International Journal of Machine Learning and Cybernetics,2019,10(1):43 - 57.

[8] Hao B, Li L, Li A, et al. Predicting mental health status on social media[C]//International Conference on Cross Cultural Design. Springer, 2013: 101 - 110.

[9] 方振宇. 基于词向量方法的微博用户抑郁预测[J]. 电子技术与软件工程,2017(7):199 - 200.

[10] 李晓晶,马欣欣,李素琴. 抑郁症发病机制与药物治疗研究进展[J]. 河北医药,2006,28(2):130 - 131.

[11] Whiteford H A, Degenhardt L, Rehm J, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010 [J]. The Lancet, 2013, 382(9904): 1575 - 1586.

[12] 杜占梅. 40 例抑郁症患者自杀原因分析及护理对策经验谈[J]. 实用临床护理学电子杂志,2019,4(3):15.

[13] Chauvet-Gelinier J C, Bonin B. Stress, anxiety and depression in heart disease patients: A major challenge for cardiac rehabilitation[J]. Annals of Physical and Rehabilitation Medicine, 2017, 60(1): 6 - 12.

[14] Song Y, Shi S, Li J, et al. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings [C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018: 175 - 180.

[15] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[EB]. arXiv preprint arXiv:1301.3781, 2013.

[16] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [C]//Advances in neural information processing systems. 2013: 3111 - 3119.

[17] Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013: 746 - 751.

[18] 魏广顺,吴开超. 基于词向量模型的情感分析[J]. 计算机系统应用,2017,26(3):182 - 186.

[19] 罗燕,赵书良,李晓超,等. 基于词频统计的文本关键词提取方法[J]. 计算机应用,2016,36(3):718 - 725.

[20] Haddoud M, Mokhtari A, Lecrop T, et al. Combining supervised term-weighting metrics for SVM text classification with extended term representation[J]. Knowledge & Information Systems,2016, 49(3):909 - 931.

[21] Peng Z, Hu Q, Dang J. Multi-kernel SVM based depression recognition using social media data[J]. International Journal of Machine Learning and Cybernetics,2019,10(1):43 - 57.

[22] 陈海红. 多核 SVM 文本分类研究[J]. 软件,2015,36(5):7 - 10.