Advances in Psychology 心理学进展, 2022, 12(5), 1631-1637
Published Online May 2022 in Hans. http://www.hanspub.org/journal/ap
https://doi.org/10.12677/ap.2022.125195

Hans汉斯

基于支持向量机的大学生心理健康分析模型研究

王维虎[1]，刘艳超[2*]，杨　雷[1]，蒋　超[1]

[1]湖北工程学院，计算机与信息科学学院，湖北 孝感
[2]湖北工程学院，信息技术中心，湖北 孝感

收稿日期：2022年2月25日；录用日期：2022年5月16日；发布日期：2022年5月23日

摘　要

大学生处于校园与社会多元化的复杂环境，易出现心理健康问题，存在分析时费时费力且具有主观性等问题，本文提出基于支持向量机的大学生心理健康分析方法。首先，构建高质量大学生心理健康语料库；其次，选取有效特征；再次，结合支持向量机算法，构建分析模型；最后，实验结果证明，测试构建的模型，正确率达到88.5%。因此，本文提出的方法有效、科学。

关键词

支持向量机，大学生，心理健康

Research on College Students' Mental Health Analysis Model Based on Support Vector Machine

Weihu Wang[1], Yanchao Liu[2*], Lei Yang[1], Chao Jiang[1]

[1]The School of Computer and Information Science, Hubei Engineering University, Xiaogan Hubei
[2]The Information Technology Center, Hubei Engineering University, Xiaogan Hubei

Received: Feb. 25th, 2022; accepted: May 16h, 2022; published: May 23rd, 2022

Abstract

College students are in the complex environment of campus and social diversity, prone to mental

*通讯作者。

文章引用: 王维虎, 刘艳超, 杨雷, 蒋超(2022). 基于支持向量机的大学生心理健康分析模型研究. *心理学进展, 12(5),* 1631-1637. DOI: 10.12677/ap.2022.125195

health problems. There are time-consuming and laborious analysis and subjective problems. This paper proposes a support vector machine based mental health analysis method for college students. Firstly, the necessary corpus is constructed. Secondly, effective features are selected. Thirdly, combined with the support vector machine algorithm, the model is constructed. Finally, the accuracy of the constructed model is up to 88.5%. Therefore, the method proposed in this paper is effective and scientific.

Keywords

SVM, College, Mental Health

1. 引言

心理健康分析，是指对心理的各个方面及活动过程所处的状态进行分析判断，能够为心理健康诊断和治疗提供依据；进而有利于学校制定相应的大学生心理障碍的早期预防、干预的新方法，预防心理障碍或行为问题，提高学生心理素质。大学生心理健康事关学生全面发展、社会进步和国家未来，对大学生心理健康关注至关重要。

当前针对学生心理健康分析的研究甚多且已取得较好的研究成果，然而研究主要针对理论的研究，对采用机器学习方法研究甚少。在国内，针对大学生心理健康分析判断问题，文献(吴婷，2017) (陈秋伍，魏惠梅，2020) (杨昱梅，李婧，2015)提出使用聚类算法对大学生心理健康数据进行分析；针对聚类分析中的 C-均值缺点问题；文献(胡秀云，2016)提出基于信息熵属性加权的 FCM 算法解决了模糊 C-均值聚类算法对初始聚类过分依赖，局部收敛，对数据集要求很高等缺点；在国外，Myers B 等(Myers, Bantjes, & Lochner, 2021)采用多种建模方法分析了南非大学一年级学生中虐待的患病率以及虐待暴露的类型、数量和模式与 12 个月常见精神障碍(CMDs)之间关系；针对大学生心理求助率低问题，Canby N K 等人(Canby, Cameron, & Calhoun, 2015)运用聚类方法分析出校园资源认识的缺乏和心理疾病的耻辱感，倡导组织与认可的学生校园分会合作，促进心理健康活动。针对心理健康分析判断正确率低、分析复杂等问题，王维虎等(王维虎等，2021)提出贝叶斯算法融合 9 种有效特征，构建心理健康分析模型，能较好地对心理健康进行判断分析，正确率达到 84.136%，但是存在将心理健康数据进行等级化处理，忽略了隐藏数据中的特征和关联性等问题。

针对以上存在的问题，本文提出基于支持向量机的大学生心理健康分析模型。首先，为了方便、统一地评价构建模型性能，本文语料库采用王维虎等(2021)中语料；其次，选取有效的心理健康特征；最后，本文采用具有较强泛化能力和适用小样本学习的支持向量机算法构建心理健康分析模型。通过本文方法能够较准确、快速地、客观地、批量地对心理健康调查问卷数据进行快速地分析和评判，给出心理健康结论，为决策者提供依据。

2. 研究思路

2.1. 困难与挑战

当前，针对大学生心理健康分析研究，诸多学者采用传统方法对收集的数据进行归纳现象、分析问

题、提出解决问题的对策，或通过对群体问卷统计方法分析，给出群体性分析和解决对策，研究主要偏向理论性的研究，应用型研究甚少；由于学生心理健康分析需要专业的心理专家进行资料分析，结果具有较强的主观性、差异性；心理健康语料库资源匮乏给研究带来困难，以上问题给本文的研究带来了困难和挑战。

2.2. 研究框架

基于支持向量机的大学生心理健康分析方法，如图 1 所示。

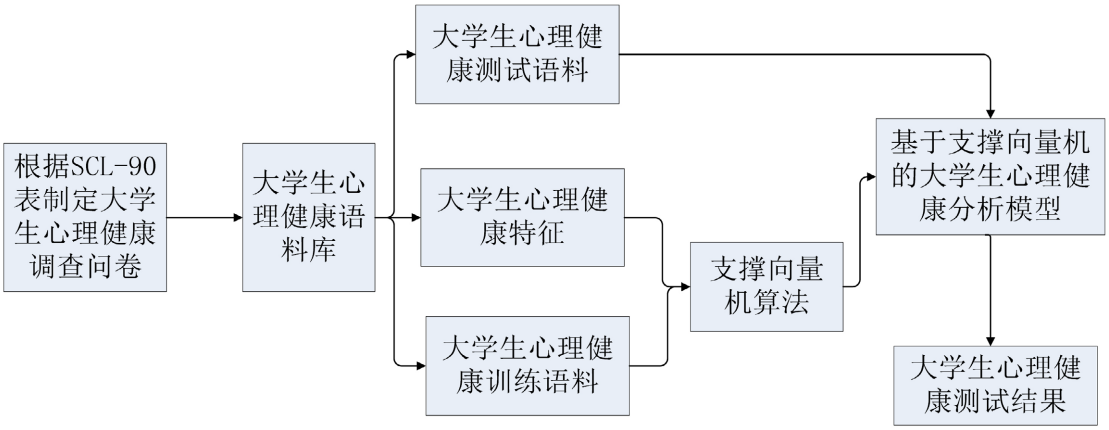

Figure 1. Research framework
图 1. 研究框架

首先，本文采用王维虎等(2021)中现有语料库，根据 SCL-90 量表制定心理健康调查问卷，邀请研究对象填写，对搜集到的调查问卷进行去噪、清洗等处理，形成学生心理健康语料库；其次，邀请大学生心理健康专家对构建的语料库进行分析并提取有效特征；再次，选取支持向量机算法融合选取有效特征，构建基于支持向量机的大学生心理健康分析模型；最后，使用大学生心理健康测试语料对模型进行测试分析和评估。

3. 研究方法

3.1. 构建语料库

语料库是科研的基础，构建语料库至关重要且语料库质量的好坏直接影响生成模型的性能。本文语料库采用王维虎等(2021)中已构建规模为 5600 条，分为训练语料和测试语料，对数据进行归一化、去噪等处理，按照格式 UTF-8 存储。

3.2. 特征选取

有效特征的选取决定模型的好坏，特征工程至关重要。将数据预处理之后的大学生 SCL-90 量表特征值，与心理测试结果目标值，做皮尔逊 Pearson 积矩相关程度分析，进一步筛选出与 CMHP (College Mental Health Problem)发生相关因素纳入预警模型。相关分析结果，如表 1 所示。

Table 1. The analysis results by Pearson
表 1. Pearson 积矩相关分析结果

特征值	躯体化	强迫症状	人际关系敏感	抑郁	焦虑	敌对	恐怖	偏执	精神病性
心理测试结果	0.7067	0.7055	0.7662	0.7827	0.7896	0.6991	0.6830	0.7244	0.7701

从表 1 中发现，在 SCL_90 量表中的 9 个有效特征值(躯体化、强迫症状、人际关系敏感、抑郁、焦虑、敌对、恐怖、偏执、精神病性)与心理测试结果目标值之间存在着较强相关性。同时，根据 SCL-90 量表要求和专家分析，最终选取的特征包含躯体化、强迫症状、人际关系敏感、抑郁、焦虑、敌对、恐怖、偏执、精神病性，将其作为 CMHP 预警模型的因素，与王维虎等(2021)一致。

3.3. 算法选取

支持向量机(SVM)适用小样本学习方法，简化了分类和回归等问题；计算的复杂性取决于支持向量的数目，而不是样本空间的维数，避免了“维数灾难”；少数支持向量决定了最终结果，对异常值不敏感，有利于抓住关键样本、“剔除”大量冗余样本，且具有较好的“鲁棒性”；表示为凸优化问题，利用已知的有效算法发现目标函数的全局最小值，具有较强泛化能力。

结合大学生心理健康数据对象的高维度、小样本等特点，本文选取支持向量机算法能很好解决数据分类问题。该算法广泛应用于自然语言处理、图像识别(王亮申，欧宗瑛，朱玉才，2005)、文本分类(赵婧，邵雄凯，刘建舟，2019)、人脸检测、验证和识别、语音识别、图像处理等领域且已取得较好成果。

4. 实验与分析

4.1. 评价指标

为了评估本文提出的基于支持向量机的大学生心理健康分析的方法的性能，本文实验评价标准采用正确率(正确预测个数与预测总数的比值)作为衡量标准，如公式(1)所示。

$$\text{正确率} = \frac{\text{预测正确个数}}{\text{预测总个数}} \tag{1}$$

式中正确率数值在 0 和 1 之间，越接近 1，表示本文构建的模型分析预测越好，模型性能越优和方法越有效。

4.2. 实验工具

本本文采用 Python、Numpy、Pandas 第三方库，进行数据获取、数据处理、特征工程、数据相关性分析，利用 Scikit_Learn 第三方工具库中 SVM 算法建立模型。

4.3. 实验数据

本文语料库采用王维虎等(2021)中已构建规模为 5600 条，分为训练语料和测试语料，对数据进行特征标准化处理。在建模时，因为数据之间数量级差别较大或量纲不同对模型的预测准确性产生影响，保证输出数据中数值较小的数据不会被吞食，提高收敛速度，缩短训练的时间，故在建模之前对数据进行特征标准化处理，结果是将原始数据的数值映射到[0, 1]区间。

离差标准化公式，如公式(2)所示。

$$X^* = \frac{X - \min}{\max - \min}$$

其中，X^*表示标准化后的输出值，max 表示样本数据的最大值，min 表示样本数据的最小值，max − min 表示为极差。

4.4. 实验结果

4.4.1. 实验一：参数调优

采用 SVM 模块中的 SVC 算法，主要参数包括 C、kernel、gamma，其中，C 是惩罚系数，即对误差

的宽容度；kernel 为核函数，本模型选择两种核函数——径向基 RBF 核与线性 Linear 核；gamma 是选择核函数后，自带的一个参数，代表支持向量机的个数，隐含地决定了数据映射到新的特征空间后的分布。选用准确率、精确率与召回率作为评估 SVM 模型的性能指标。

研究利用离差标准化后的 70%数据作为训练集，30%数据作为测试集，来选择模型最佳参数。首先固定 gamma = 0.5，调整 C 值，分别观察 kernel = “rbf”与 kernel = “linear”时的拟合结果，如表 2、表 3 所示。

Table 2. The model accuracy under different parameters C (gamma = 0.5, kernel = “linear”)
表 2. 不同 C 下的模型精度(gamma = 0.5, kernel = “linear”)

C 值	0.001	0.01	0.1	1	2	3	4	5	10	100
准确率	0.8338	0.8338	0.8338	0.8318	0.8318	0.8328	0.8447	0.8338	0.8338	0.8318
精确率	0.8338	0.8338	0.8338	0.8938	0.8938	0.8938	0.8948	0.8938	0.8938	0.8928
召回率	1	1	0.9903	0.9903	0.9902	0.9901	0.9902	0.9902	0.9902	0.9902

Table 3. The model accuracy under different parameters C (gamma = 0.5, kernel = “rbf”)
表 3. 不同 C 下的模型精度(gamma = 0.5, kernel = “rbf”)

C 值	0.001	0.01	0.1	1	2	3	4	5	10	100
准确率	0.8536	0.8918	0.8928	0.8928	0.8928	0.8928	0.8938	0.8928	0.8928	0.8523
精确率	0.8473	0.8921	0.8933	0.8921	0.8941	0.8940	0.8951	0.8938	0.8939	0.8927
召回率	1	0.9918	0.9981	0.9981	0.9981	0.9981	0.9981	0.9981	0.9981	0.9981

从表 2、表 3 中可以看到，当 C 值不同时，在核函数 kernel = “rbf”'模型时，准确率与精确率评价指标略高于核函数 kernel = “linear”模型。当 C < 4 时，准确率与精确率不断增加；当 C > 4 时，准确率与精确率不断降低，故选取 4 为最佳 C 值。

然后，再固定 C = 4，调整 gamma 值，分别观察 kernel = “rbf”与 kernel = “linear”时的拟合结果，如表 4、表 5 所示。

Table 4. The model accuracy under different parameters C (C = 4, kernel = “linear”)
表 4. 不同 gamma 下的模型精度(C = 4, kernel = “linear”)

gamma 值	0.001	0.01	0.1	0.2	0.3	0.4	0.5	1	10	100
准确率	0.8823	0.8823	0.8823	0.8823	0.8903	0.8905	0.9003	0.8913	0.8901	0.8823
精确率	0.8927	0.8927	0.8936	0.8938	0.9027	0.9031	0.9045	0.9027	0.8936	0.8927
召回率	0.9981	0.9981	0.9981	0.9981	0.9981	0.9981	0.9981	0.9981	0.9981	0.9981

Table 5. The model accuracy under different parameters C (C = 4, kernel = “rbf”)
表 5. 不同 gamma 下的模型精度(C = 4, kernel = “rbf”)

gamma 值	0.001	0.01	0.1	0.2	0.3	0.4	0.5	1	10	100
准确率	0.8918	0.8923	0.8928	0.8933	0.8938	0.8945	0.9008	0.9905	0.8967	0.8946
精确率	0.8931	0.8931	0.8931	0.8947	0.8948	0.9021	0.9043	0.9039	0.8948	0.8931
召回率	0.9981	0.9987	0.9993	0.9993	0.9993	1	1	0.9987	0.9951	0.9902

从表 4、表 5 中可以看到，当 gamma 值不同时，在核函数 kernel = “rbf”模型时，准确率与精确率评价指标略高于核函数 kernel = “linear”模型。当 gamma < 0.5 时，准确率与精确率不断增加；当 gamma > 0.5 时，准确率与精确率不断降低，故选取 0.5 为最佳 gamma 值。

经过上述分析，确定了模型的最优参数为：C = 4，gamma = 0.5，kernel = “rbf”，此时模型拟合的准确率为 0.8973，精确率为 0.8997。

4.4.2. 实验二：对比实验

比较本文提出基于支持向量机的大学生心理健康分析的方法与王维虎等(2021)中方法的性能，实验结果如图 2 所示。

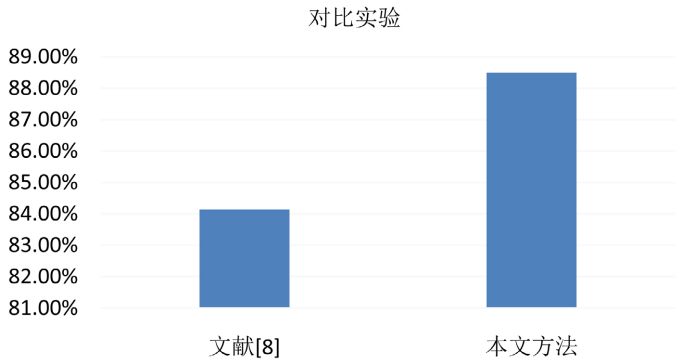

Figure 2. Comparative experiment and research
图 2. 对比实验

根据图 2 和王维虎等(2021)可知，本文提出的方法正确率达到 88.5%，王维虎等(2021)正确率达到 84.136%，正确率高出 4.364%。因此，本文提出的方法优于王维虎等(2021)中提出的模型。

4.4.3. 实验三：封闭和开放实验

为了验证本文方法的鲁棒性、稳定性等性能和效果，本实验进行开放与封闭的性能评估，实验结果如下图 3 所示。

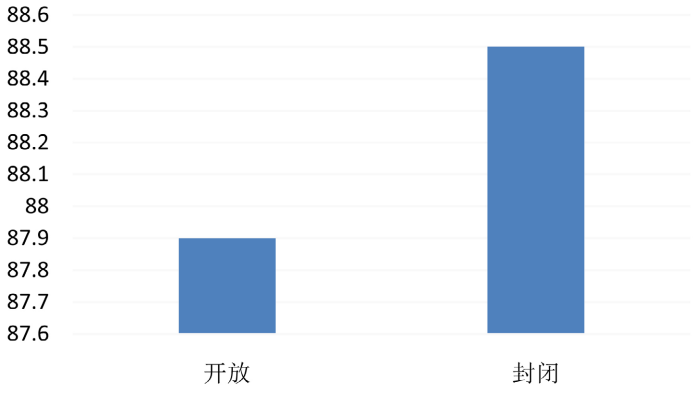

Figure 3. Open and closed test
图 3. 开放和封闭测试

由图 3 可知，在开放环境下，正确率达到 87.9%；在封闭环境下，正确率达到 88.5%，开放测试比封闭测试低 0.6%。因此，本文构建的模型具有很强的鲁棒性和稳定性等。

5. 总结

当前，社会环境多样化和复杂化，大学生群体是具有较高知识水平的和敏感的群体，大学生心理活动易受外界环境影响，心理问题也逐渐增多和凸显，及早发现和干预学生心理问题至关重要，对心理健康数据采用人工分析耗时耗力且具有主观性等问题，针对以上问题本文提出基于支持向量机的大学生心理健康分析模型。首先，构建语料库，本文在王维虎等(2021)语料库基础上扩增语料库规模；其次，选取与王维虎等(2021)相同的有效特征；再次，结合泛化能力强、适用小样本、鲁棒性优势的支持向量机算法构建模型；最后，对构建的模型进行测试。构建的模型可快速、准确地分析大学生心理健康问题，为高校学生心理健康管理者提供依据。下一步，采用深度学习方法挖掘有效的特征；由于调查问卷填写具有很强的主观性，所以需要优化调查问卷结果，提高语料库质量，进一步提高模型的性能。

基金项目

2019 年国家自然科学基金面上项目(61972136)；2018 年第二批教育部产学合作协同育人项目(201802325001)；2019 年第一批教育部产学合作协同育人项目(201901023010)；2020 年度孝感市自然科学计划项目(XGKJ2020010038)；2020 年度孝感市自然科学计划项目(XGKJ2020010064)； 2020 年湖北工程学院教学改革研究项目(2020A05)。

参考文献

陈秋伍, 魏惠梅(2020). K-means 聚类算法在分析大学生心理健康的应用. *数码世界, (4),* 144-145.

胡秀云(2016). *大学生心理健康数据模糊聚类分析研究*. 硕士学位论文, 信阳: 信阳师范学院.

王亮申, 欧宗瑛, 朱玉才, 等(2005). 基于 SVM 的图像分类. *计算机应用与软件, 22(5),* 98-99+126.

王维虎, 刘艳超, 程芳, 纪慎思(2021). 基于朴素贝叶斯算法的大学生心理健康分析研究. *心理学进展, 11(7),* 1723-1731.

吴婷(2017). *基于 K-means 聚类算法的大学生心理管理系统研究*. 硕士学位论文, 武汉: 湖北工业大学.

杨昱梅, 李婧(2015). 聚类分析算法在大学生心理健康分析中的应用研究. *中国教育学刊, (S1),* 3.

赵婧, 邵雄凯, 刘建舟, 等(2019). 文本分类中一种特征选择方法研究. *计算机应用研究, 36(8),* 2261-2265.

Canby, N. K., Cameron, I. M., Calhoun, A. T. et al. (2015). A Brief Mindfulness Intervention for Healthy College Students and Its Effects on Psychological Distress, Self-Control, Meta-Mood, and Subjective Vitality. *Mindfulness, 6,* 1071-1081. https://doi.org/10.1007/s12671-014-0356-5

Myers, B., Bantjes, J., Lochner, C. et al. (2021). Maltreatment during Childhood and Risk for Common Mental Disorders among First Year University Students in South Africa. *Social Psychiatry and Psychiatric Epidemiology, 56,* 1175-1187. https://doi.org/10.1007/s00127-020-01992-9