

基于大数据与深度学习的高职学生心理健康问题预测研究

杨娟,徐大宏

(1.湖南工业职业技术学院,湖南 长沙 410000;2.湖南师范大学数学与计算机科学学院,湖南 长沙 410000)

摘要:在高职心理健康教育具体工作中都会产生大量多样化和碎片化数据,本文通过数据预处理和深度学习技术运用,利用BP算法对影响心理问题的多种类型的历史数据,包括学生心理普查数据、个人基本信息及社会支持统计数据等进行表征学习,不仅能充分考虑到数据之间的异质性,还能探究到它们之间潜在的关联性,最终提升心理问题风险预测的作用,实现心理问题预测工作精准化和深入化。

关键词:深度学习;心理问题预测;BP算法;高职学生

中图分类号:TP391 文献标识码:A

文章编号:1009-3044(2020)29-0001-03

开放科学(资源服务)标识码(OSID):



我国各级教育部门高度重视高职学生心理预警工作,然而在实际工作中普遍缺乏信息化技术应用意识,增加工作的被动性。此外高职学生的心理健康水平受到内外诸多因素的影响,各因素间又有复杂的关联,是一个非线性问题。

近几年随着大数据时代、云计算的到来,深度学习在实际运用中都有极大地突破,它不仅具备较深的网络结构,是一种智能化机器学习,并且通过一系列算法,使计算机从海量历史数据中学习规律,建立分析模型,实现了从特征提取至特征分类的全过程端到端(end-end),对新的样本进行智能识别或对未来做预测。很多学者提出利用机器学习等算法对影响学生心理问题的影响因素进行分析,通过机器算法或建立分析模型,将原始数据开始每层表示(或特征)经过不断训练,逐渐转换为更高层的表示,获得最大数据价值,从而进行分类预测,从而达到心理问题预测的目的。

1 研究路线

本文将影响高职学生心理问题主要因素作为样本输入,数据主要体现在SCL-90量表中加入个人基本信息、社会支持等因子,利用BP神经网络建立心理问题预测模型,通过各影响因素与心理问题的映射关系,对输入的测试样本进行学习训练,只要不断输入影响因素数据,完成训练分析,使预测模型更贴近调查样本实际特点,达到心理问题预测的目的。

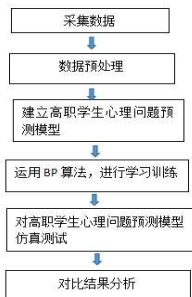


图1 研究路线图

2 数据采集及预处理

2.1 数据采集

本研究所使用的数据源于湖南工业职业技术学院、湖南科技职业技术学院、湖南工程职业技术学院具有代表性3所高职院校近5年的新生心理普查数据、学生基本情况表、社会支持问卷统计数据,共计1036名学生历史数据。在数据收集过程中,心理普查数据主要参考90项症状清单(SCL-90量表),采用五级评分法,从10个因子反映心理健康状况(F_1 - F_{10}),并有特定正常值参考范围;学生基本情况表是基于高职院校学生具有相同的基础属性特征:如年龄、教育程度、教育经历等,也有不同的属性特征,如性别、专业、生源地、家庭状况等,在本研究中主要参考了性别、年龄、专业、生源地、家庭经济收入五类数据;在外部影响因素中主要参考了物质支持满意度、精神支持满意度两类数据。

2.2 数据预处理

2.2.1 数据特征提取

本研究根据影响高职院校学生心理问题的有关的各种属性特征因素,通过对这些因素的加权调整,使调整后的结果更准确反映高职院校学生的心理健康情况。具体分为:首先收集心理问题影响内外因素的三类别大量小数据,如个人基本情况、社会支持情况等;其次依据表现特征进行分类,按影响大小进行分级,制定各特征分数的加权影响分值的范围值,以及递进和递减规则。如SCL-90量表中有10个因子,每个因子所含的项目不同,对数据进行因子均分处理,结果如下表1所示。

表1 SCL-90数据处理部分结果

F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	其他
1.45	2.11	1.91	1.87	2.09	1.23	2.23	1.09	2.09	2.15
1.79	1.23	2.34	1.12	1.86	1.83	1.96	2.10	2.11	1.34

收稿日期:2020-05-30
基金项目:湖南省教育科学“十三五”规划2018年度课题“基于大数据与深度学习的高职学生心理健康问题预测研究”(批准号:XJK18BZY006)研究成果之一
作者简介:杨娟(1982—),女,湖南会同人,硕士研究生学历,湖南工业职业技术学院,副教授,研究方向:思想政治教育;徐大宏(1971—),男,湖南岳阳人,副教授,博士,研究方向为数字计算处理、计算机视觉。

2.01	1.89	1.56	1.67	1.34	2.13	1.65	1.67	1.23	1.66
1.32	2.56	1.23	2.09	1.03	1.44	1.23	1.23	1.78	1.76
1.68	1.09	2.34	1.02	2.78	1.61	1.67	1.56	1.22	2.31
2.21	2.45	1.57	1.89	1.56	2.01	2.22	1.20	1.41	1.21
1.23	1.56	1.23	2.62	1.45	1.89	1.23	1.84	2.01	1.77

2.2.2 数据归一化处理

样本数据进行网络训练时,样本数据呈现碎片化、多样化、维度单一等特点,必须将大量存在的小数据进行归一化处理,符合可追踪、可量化、可比较、可分析、高维度的大数据特征,并映射到[-1,1]区间处理,更加简便。设置网络输入、输出矩阵,如表 2 所示。

表 2 网络输入、输出矩阵

函数	运行设置	注释
mapminmax	[sr,ps]=mapminmax(r,rmin,rmax)	r 为归一化处理的输入矩阵
	[sc,ps]=mapminmax(c,cmin,cmax)	c 为归一化处理的输出矩阵

再将网络输入、输出数据限制在[0,1]或[-1,1]区间,公式如下:

$$\overline{Xi} = \frac{Xi - Xmin}{Xmax - Xmin}$$

X_i 表示输入或输出数据, X_{min} 表示数据中最小值, X_{max} 表示最大值。将输入输出数据设置在[-1,1],当中间值的原始数据转化为0,最大值与最小值为1与-1,公式如下:

$$Xmid = \frac{Xmax + Xmin}{2}$$
$$\overline{Xi} = \frac{Xi - Xmid}{\frac{1}{2} (Xmax - Xmin)}$$

3 构建 BP 网络预测模型

以构建 BP 神经网络模型为工具,基于内外因影响大学生心理问题因素,进行 BP 算法优化,并对模型进行仿真训练,不断达到优化状态。BP 学习网络模型指预测模型中的神经网络结构设计,本研究将影响学生心理健康问题结构化数据(训练样本数据)为输入节点,从心理预测角度设定过滤条件,即中间设置层数(隐藏层)、节点数模型结构,输出节点为心理问题预测初步预期值。

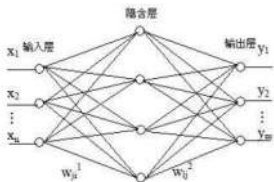


图 2 BP 网络结构图

3.1 输入层神经元数目的确定

在 BP 网络输入层神经元数目的选取时,表现样本整体特征的变量,必须满足选取的输入变量之间线性关系不强,且对输出影响大并且能检测的变量。

结合前期研究基础,从内外角度分析高职生心理问题预测指标的特征层包括三个,即心理自评、个人基本情况及社会支持。综合确定 17 个神经元数目。如: A_1 性别(xb)、 A_2 年龄(nl)、 A_3 专业(zy)、 A_4 生源地(syd)、 A_5 家庭经济收入(jtsr); B_1 物质支持程度(wzzc)、 B_2 精神支持程度(jszc); C_1 躯体化因子(qth)(量表 1、4、12、27、40、42、48、49、52、53、56、58) C_2 强迫因子(qp)(量表 3、9、10、28、38、45、46)、 C_3 人际敏感因子(rjmg)(量表 6、21、34、36、37、41、61、69、73)、 C_4 抑郁因子(yy)(量表 5、14、15、20、22、26、29、30、31、32、54、71、79)、 C_5 焦虑因子(jl)(量表 2、17、23、33、39、57、72、78、80、86)、 C_6 敌对因子(dd)(量表 11、24、63、67、74、81) C_7 恐怖因子(kb)(量表 13、25、47、50、70、75、82)、 C_8 偏执因子(pz)(量表 8、18、43、68、76、83)、 C_9 精神病性因子(jsbx)(量表 7、16、35、62、77、84、85、87、88、90)、 C_{10} 其他因子(qt)(量表 19、44、59、60、64、66、89)。

对应每 1 项特征层各选取影响较大、关联较强的指标,并对基础层各项指标赋值,赋值为 0-10,同时,确定各项基础层系数,取值为 0-1,因此计算出每项特征层指标对应的数值。

$$Ai = \frac{\sum_{j=1}^3 a_j \partial y}{\sum_{j=1}^3 y}$$

3.2 输出层神经元数目的确定

输出层神经元数目的确定,决定于如何定标准的输出。输出节点代表系统要实现的功能目标。在本研究中,研究对象的期望输出是高职学生心理问题风险的预测。因此,输出层神经元数目选择为 3(zhengchang: 等于 1,表明处于正常范围;qingdu: 等于 1,表明有心理困惑,但处于正常范围内;yanzhong 等于 1,表明心理异常。)

3.3 隐层神经元数目的确定

在隐层节点数实际的结构实现上,选择合理的数目的隐层节点是个复杂的问题,关系到整个网络精准度。在本研究中采用黄金分割法来确定隐层节点数,即设置该 BP 网络中输入层节点数为 A,输出层节点数为 B,隐层节点数为 L,L 的取值范围为 [A,B],具体公式如下:

$$a = \frac{A + B}{2} \leq L \leq (A + B) + 10 = b$$

计算公式 $l_1 = 0.712 \times (b - a) + a$,将 l_1 作为隐层节点数计入训练模型,训练后误差 $E = (l_1)$;同样设定 l_2 为隐层节点数计算模型,训练后误差为 $E(l_2)$,对比 $E(l_1)$ 与 $E(l_2)$,假如 $E(l_1) < E(l_2)$,则删除 (a, l_2) ,重复以上步骤,即可在较短的时间内确定最佳隐层节点数 L 。

4 深度网络训练

BP(Back prapagation)算法是一类有导学习算法,主要用于模式识别、分类、函数逼近等主要方面。BP 算法基本思路是建立 BP 网络对网络权值不断调适修正,达到误差函数沿梯度方向下降,分为正向传播和误差反向传播两个部分。具体为首先输入样本数据通过输入层,经过隐含层计算每个单元的实际输出值;其次为当输出层实际输出与期望的输出值存在误差,则需不断调整权值,进行网络的学习训练过程,当误差达到期望误差,则网络学习过程结束。

4.1 训练参数设置

BP 网络中输入向量为 $X = (x_1, x_2, x_3, \dots, x_n)^T$ 隐层输出向量为 Y

$= (y_1, y_2, \dots, y_n)^T$, 输出层输出向量为 $o = (o_1, o_2, \dots, o_k)^T$, 期望输出向量为 $d = (d_1, d_2, \dots, d_n)^T$. 输入层到隐层之间的权值矩阵用 V 表示: $V = (V_1, V_2, \dots, V_n)$, 隐层到输出层之间的权值矩阵用 W 表示, W_k 为隐层第 k 个神经元对应的权向量. 学习过程具体分为将隐层和输出层各结点权值赋随机数--提供训练集样本--设置输入层节点 X_i , 计算输出值 $Y = f(\text{net}_i)$ --设置隐层节点输出 y , 隐层与输出层的权值 w_k , 计算输出值 $o_k = f(\text{net}_k)$ --计算网络输出误差--调整权值--依次取下一个样本, 核查网络总误差小于预期误差--记录权值, 结束训练.

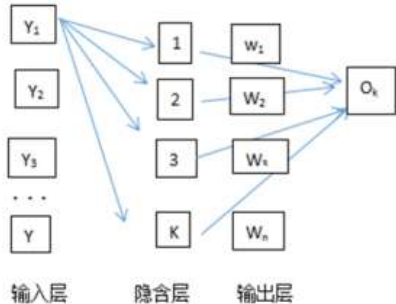


图 3 BP 算法运算过程图

4.2 具体训练过程

读入训练样本数据, 对样本数据进行删除、增补等预处理后, 对所有管理数据赋予初始值. 选取 load 函数读取样本数据, 具体如表 3 所示.

表 3 选取 load 函数读取样本数据

函数	运行设置
load	Y=load('input')
	O=load('output')

设置 100 组数据为训练样本, 20 组为预测样本, 所有数据根据其递进规则进行训练, 所得结果与实际值进行比较得到在上述训练数据情况下最匹配模型, 该模型可以随着训练标本数据量的增加以及标本数据中属数据的增加变得更加准确真实, 如表 4 所示.

表 4 数据分类

函数	运行设置	注释
divdevec	[xl,cs]=divdevec(r,c,20%)	XL 为训练数据, CS 为预测数据, 20% 为预测数据百分比

网络训练算法选取 trainlm 算法, 把学习训练网络设置为 17 个神经元, 其神经网络模型结构为 [17, 17, 1], 则神经输入第 1 层为 17 个输入数据, 对应预测结果为第 3 层, 第 2、3 层神经元选取用 tansig 函数, 如表 5 所示.

表 5 BP 网络建立设置

函数	运行设置	注释
Newff trainlm	net=newff(sr,sc,17)	隐含层为 17 个神经元
	net=trainlm(net,sr,sc)	

运用 BP 算法对网络进行学习训练, 并不断对各参数进行

调整:

```
net.trainparam.epochs=980;  
net.trainparam.ir=0.12;  
设置训练目标为 0.000000001  
则 net.trainparam.goal=0.000000001
```

在本文中所建立的神经网络不断学习训练过程中, 误差在不断减少, 在预期 900 步时, 误差率控制在 3% 能在较短时间内完成数据拟合, 基本实现对高职学生心理问题的预测研究.

4.3 仿真测试

将测试数据录入已建好的神经网络模型进行仿真测试, 通过若干组数据仿真测试可以得出, 每组预测结果误差比较小, 基本达到要求, 证明模型能对所研究样本状况进行问题的预测, 如表 6 所示.

表 6 仿真测试

函数	运行设置
Sim mapminmax	An=sim(net, sr_test) Bpsc=mapminmax(' reverse ',an,sc)

本文将影响高职学生心理问题因素通过 BP 神经网络学习训练映射到相应的心理健康状况, 经过特征选取, 不断学习训练及自我训练, 将初始权值与阈值调整, 达到问题预测并完成结果对比, 如表 7 所示.

表 7 部分样本数据的结果分析

序号	预测输出	标准值	绝对误差	相对误差
1	1.78	1	-0.132	-8.81%
2	2.21	2	-0.098	-5.34%
3	1.97	1	0.075	2.09%
4	1.76	1	0.121	3.67%
5	2.08	2	-0.097	-2.45%
6	2.12	2	0.924	4.56%
7	1.89	1	-0.216	2.09%

5 结论

结合我国现阶段高职学生心理健康特点, 将内外影响因素数据进行预处理, 建立神经网络预测模型, 并利用 BP 算法不断进行学习训练, 将神经网络进行最佳优化, 并通过仿真训练测试模型的性能, 显示误差值在可控制范围内, 在一定程度上可以对心理问题预测, 模型具有可行性.

参考文献:

[1] 杨娟. 大数据背景下高职学生心理健康影响因素调查研究[J]. 现代职业教育, 2020(7):12-13.
[2] 鲁玮. 数据挖掘技术在高职学生心理健康数据中的应用研究[D]. 合肥:安徽大学, 2019.
[3] 王丽. 基于神经网络的研究生心理健康状况预测分析[D]. 西安:西安科技大学, 2016.
[4] 耿红. 研究生心理健康问题产生的原因和对策分析[J]. 管理学家, 2013(24):497-498.
[5] 杜婷. 硕士研究生心理健康状况及其影响因素的相关研究[D]. 郑州:郑州大学, 2005.

【通联编辑:唐一东】