# 基于神经网络的博士生行为习惯与健康状况关联度分析

谭玉磊,徐晶,英爽
（哈尔滨工业大学研究生院）

**摘要：**博士生作为我国高等学历教育的最高层次，其生活行为习惯和心理状态等都有着独立于其他群体的特殊性，繁重学业科研任务使其体育活动时间受限，而婚恋、就业、社会等多元压力来源也促使其各类疾病的发生率高于其他群体。以某高校高年级博士生体检的数据样本为参照，以构建神经网络进行模式识别为基本方法，针对样本群体行为习惯倾向与健康状况、疾病发生种类进行关联度分析，对了解影响博士生健康的因素有一定的参考价值。

**关键词：**博士生；健康；神经网络

博士生群体是我国各高等院校推动科学研究发展的主要力量之一，近年来一些调查和研究显示该群体健康状况不容乐观，各类疾病频发，其中不乏重大恶性致死疾病。究其原因，一方面博士生需要应对较为繁重的科研任务和自身的学业需要，主观关注点集中在学习科研之上，易忽略体育锻炼等影响健康状况的因素，同时客观上也缺乏进行提升健康水平相关活动的时间；另一方面博士生由于年龄、身份的原因需要接受来自社会、婚恋、就业等多方面、多种类的压力，使得易于陷入急躁、焦虑、抑郁等已诱发各类疾病的负面情绪之中。

行为习惯与各类影响健康状况的疾病往往有着直接或间接的关联，如缺乏运动可导致脂肪肝，而附加饮酒等习惯可进一步诱发酒精肝、肝硬化等疾病。因此，针对博士生行为习惯和实际发生的疾病，进行两者的关联性分析，一方面可以揭示高致病率的行为习惯，对博士生起到警示作用，另一方面可以为高校研究生事务管理工作者提供有效参考，指导相关工作的设计和实施。

根据中国某著名工科高校若干名三年级博士生体检的行为习惯问卷和身体医学检查结果作为样本，并利用 BP 神经网络进行了模式识别和关联性判断，对上述问题进行了较为详细的分析。

## 一、样本基本情况介绍

本次选取的三年级博士生人数为 553 人，其中男性 412 人，女性 141 人，最大年龄为 35 岁，最小年龄为 22 岁，平均年龄为 27.83 岁。

身体医学检查项目包括身高、体重、血压、血常规、血生化（主要检查血脂、血糖、肝功、肾功等项目）、尿常规、彩超（消化系、泌尿系、男性前列腺、女性子宫及乳腺）、胸透、心电，女性加做妇科检查。诊断标准按照医学各类疾病诊疗指南制定的诊断标准为依据，指标异常以各类检查设备、试剂规定的参考范围为依据进行判定。

各项检查项目均在正常范畴者 81 人，占总人数的 14.65%，检查结果异常为 472 人，占总人数的 85.35%。客观上反映了博士生健康状况不容乐观。常见疾病的检查情况，总体上看，前列腺疾病、乳腺疾病、脂肪肝（未包含脂肪沉积）是检出疾病中的前三位：

表 1 样本前十种疾病顺位表

| 顺位 | 疾病名称 | 患病人数 | 患病率 |
|---|---|---|---|
| 1 | 前列腺疾病（男） | 153 | 37.32% |
| 2 | 乳腺疾病（女） | 43 | 31.39% |
| 3 | 脂肪肝 | 112 | 20.44% |
| 4 | 高尿酸 | 91 | 16.58% |
| 5 | 肥胖 | 71 | 12.93% |
| 6 | 胆囊疾病 | 68 | 12.41% |
| 7 | 泌尿系结石 | 68 | 12.39% |
| 8 | 高血压 | 58 | 10.56% |
| 9 | 高血糖 | 10 | 1.82% |
| 10 | 高血脂 | 4 | 0.73% |

对上述参检人员，以问卷调查为具体形式进行了行为习惯调查，具体情况如下表所示：

表 2 样本行为习惯调查问卷情况表

| 序号 | 问题内容 | 采样方式 |
|---|---|---|
| 1 | 婚姻状态 | 选择 |
| 2 | 是否吃早餐 | 是否 |
| 3 | 肉类摄入量 | 程度 |
| 4 | 宿舍和谐度 | 程度 |
| 5 | 睡眠时间 | 程度 |
| 6 | 是否失眠 | 是否 |
| 7 | 是否饮酒 | 是否 |
| 8 | 是否吸烟 | 是否 |
| 9 | 是否关注健康 | 是否 |
| 10 | 体育锻炼次数 | 选择 |
| 11 | 锻炼时间 | 选择 |
| 12 | 体育爱好数量 | 程度 |
| 13 | 文艺爱好数量 | 程度 |
| 14 | 是否有压力 | 是否 |
| 15 | 压力来源 | 程度 |
| 16 | 事情多否 | 程度 |
| 17 | 解压方式 | 选择 |
| 18 | 密切朋友 | 程度 |
| 19 | 是否关注健康 | 是否 |
| 20 | 与他人相处融洽 | 程度 |

根据上述情况进行了问卷的设计制作，同时为保证行为习惯数据和健康状况数据的关联度，本次调

查为实名调查，经样本个人同意，共采集行为习惯数据 284 条，并通过调查问卷准确度方法排除部分非准确数据，得到有效数据 269 条作为样本依据。

## 二、运用神经网络进行关联度分析的可行性

人工神经网络（Artificial Neural Network, ANN），常简称为神经网络，是一种模拟大脑神经突触联接的结构进行信息处理的数学模型。神经网络由大量的人工神经元以及人工神经元之间的连接构成，经由算法调整内部人工神经元之间相互连接的关系来实现信息处理的目的[1]。神经网络涉及神经科学、思维科学、人工智能和计算机科学等多个学科，具有大规模并行处理、高度冗余和非线性运算等特点，因而具有很高的运算速度、很强的适应性、很强的学习能力、容错能力和自组织能力，已成功应用于模式识别、计算机视觉、智能控制、非线性优化、机器人等多个领域，尤其在模式识别领域具有优势。

其中，BP（Back Propagation）神经网络是一种按误差反向传播算法训练的多层前馈网络，是目前应用最广泛的神经网络模型之一[2]。其模型拓扑结构包括输入层、隐含层和输出层，每层均各有足够数量的神经元。其学习过程分为两个阶段，信号的正向传播和误差的反向传播。样本输入信号在神经网络中正向传播，其网络输出与样本给定输出值之间会存在误差，然后基于该误差使用最速下降法，通过反向传播来不断调整网络的权值和阈值，使网络的误差平方和最小，从而使网络达到理想的预测精度。

BP 神经网络能够通过使用样本进行训练学习的方法自动的找到输入与输出之间的对应关系，并且该过程不需要人来控制，人们只需设置好各训练参数即可。BP 神经网络的这种特性与样本行为习惯和健康状况的关联度分析结构高度契合，如果把样本行为习惯作为输入向量，把其健康状况作为输出向量，逻辑上设定为由一定行为习惯导致一定健康状况的基本判断，并进行关联分析，则可以得到有针对性的结果，这简便又有效的特点使之成为本次数据分析的首选。

## 三、网络建立过程

### 1、数据预处理

因样本涉及的病例较多，按照单项病例进行健康状况的输出向量设计在小样本情况下易导致网络泛化能力变弱，同时过于明确的健康状况输出实际上也缺乏很强的指导意义，因此为将输出向量简单化并使其有指导意义，将各类疾病归纳到四个大类中：

**亚健康类：**脂肪肝、高血脂、高血糖、高血压等；

**男女性别类：**乳腺疾病、前列腺疾病、妇科疾病等；

**内脏器官类：**肝病、胆囊疾病、肾结石、心律失常等；

**其他类：**白化病、贫血症、骨质疏松等其他类型的疾病。

此外，将样本行为习惯的问卷调查各结果均整理为数字格式，按照常规判断针对其对疾病的贡献度进行处理，贡献度越大其数值越大。

最后将上述 269 条有效数据作为样本。其中问卷数据作为样本输入值，并且各条输入数据分别按比例归一化到[0.2,1]之间，不选择 0 作为初始值的原因是确保网络有响应；病例数据作为样本输出值，输出值为 0 或 1（0 表示无此病，1 表示患此类疾病）。另外，按照 BP 神经网络的训练和测试样本比例分配的“二八原则”将 269 条样本数据随机分为 215 条训练样本和 54 条测试样本。

### 2、建立网络

根据单一样本的输入输出向量数据数目来确定输入层、输出层神经元数目，其中行为习惯问卷调查 20 道问题，因此单一样本有 20 个输入数据，而经过疾病归类后，输出数据变为 4 个，因此将关联度判断的网络设计成输入层神经元 20 个，输出层神经元 4 个。另根据隐含层神经元推荐选取方法——最佳隐含层节点数为输入层节点数与输出层节点数之积的开平方，选取隐含层神经元 9 个。

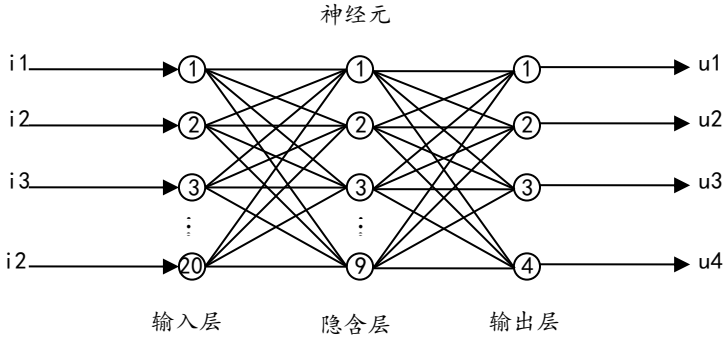


图 1 网络模型图

本文选用 MATLAB 仿真分析软件进行建模和数据处理，在运算过程中还需要确定激励函数、梯度下降算法、初始权值、阈值和学习速率等参数。

激励函数通常选取在[0,1]或[-1,1]区间内连续的单调可微分的 S 状函数 ,这类函数对权值和阈值的影响敏感，能够更加准确的控制权值和阈值的变化。本网络中输入层和隐含层的激励函数选为“tansig”，其变化区间为[-1,1]；由于输出为 0 或 1，故而隐含层和输出层间的激励函数选取变化区间为[0,1]的“logsig"函数。

梯度下降算法决定了权值和阈值的变化方式，合适的梯度下降算法能够使权值和阈值的变化迅速又准确。本网络采用的 traingdx 算法（变学习率动量梯度下降算法） 能够根据权值和阈值的影响大小选择合适的学习速率控制权值和阈值的变化大小。

神经网络模型属于非线性系统模型，初始权值和阈值的选取会影响到网络训练所需时间，甚至网络能不能收敛的问题。初始值太大会使加权后的输入值落在 S 型激活函数的饱和区域，极大地延缓收敛时间。最佳的初始权值和阈值在[-1,1]区间之内，这也是 MATLAB 默认的初始权值和阈值选取区间[3]

学习速率关系到每次权值和阈值变动的大小。学习速率太大会使系统变得不稳定，而学习速率太小又会增加训练时间。一般而言，学习速率可取在 0.01-0.8 范围之内[4]

根据以上各参数选定方式建立如下 BP 神经网络：并利用 MATLAB 软件进行仿真模拟，程序设计语言为：

```
net=newff(minmax(Input_train),[9,4],{'tansig','logsig'},'traingdx')
```

## 3、网络训练

选定训练精度在0.05，训练次数为50000次，学习速率为0.2进行网络训练。训练情况如下：

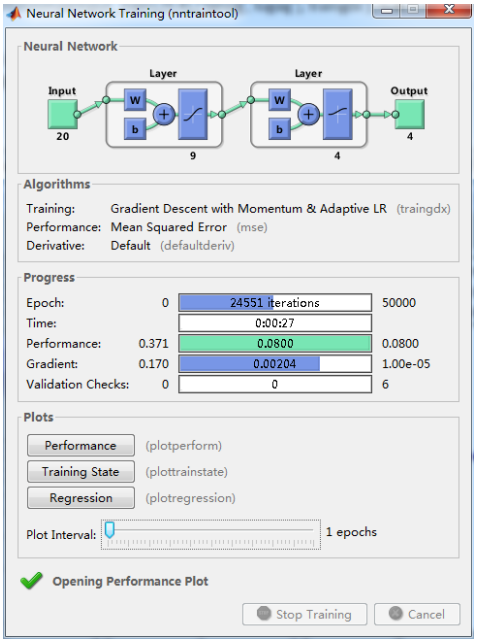


图2 神经网络训练结果（进行变图呈现）

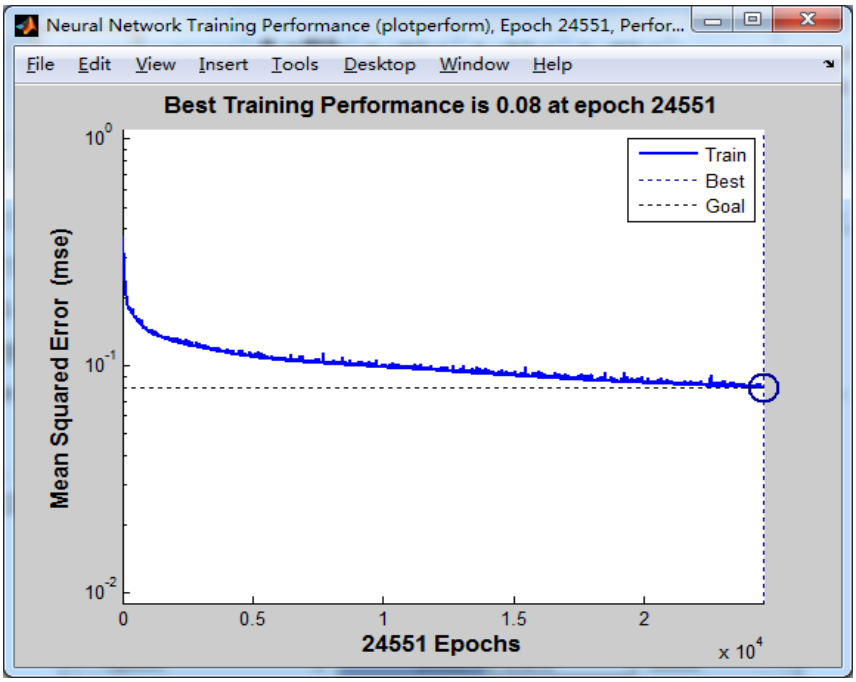


图3 神经网络精度变化过程

从上面两图可以看出，网络已满足精度需求。网络训练完毕

## 4、网络测试

将测试样本输入到网络中得到的测试样本输出与样本给定输出值的比较如下：

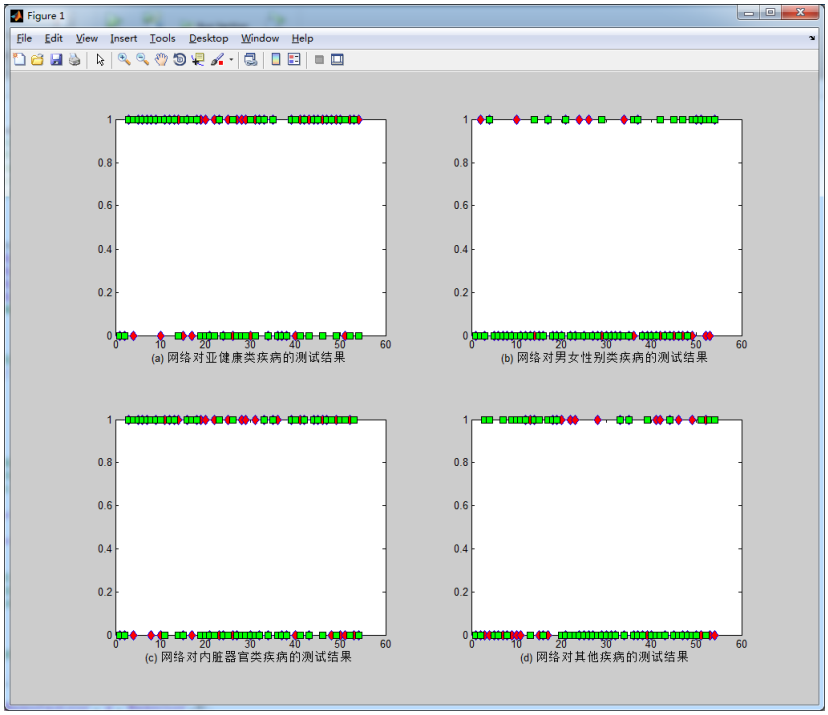


图4 网络对各类疾病的样本测试结果

其中绿色正方形代表样本给定输出值，红色菱形为测试样本输出值。从上图中可以看出，从统计学而言，测试样本输出值与样本给定输出值之间误差较小，证明经过训练后的BP神经网络对于问卷内容和病例之间的关系把握的比较好，验证了网络进行行为习惯和健康状况关联度判断的准确性。

## 四、博士生行为习惯与健康状况关联度分析

在网络测试达到标准后，为了进行单项行为习惯与健康状况的关联度判断，将数据进行如下分析处理，将单一行为习惯参数依次设为“1”，同时其他行为习惯参数为“0”，记录 4 个网络输出的数值，顺次处理全部行为习惯参数，共得到单一行为习惯和单一健康疾病全部关联度数据 80 条，如下表：

表 3 博士生行为习惯与健康疾病关联度数据

| 行为习惯 | 疾病类型 | | | |
|---|---|---|---|---|
| | 亚健康类 | 男女性别类 | 内脏器官类 | 其他 |
| 婚姻状态 | 0.7259 | 0.9853 | 0.5647 | 0.1887 |
| 是否吃早餐 | 0.9401 | 0.1017 | 0.9935 | 0.2167 |
| 肉类摄入量 | 0.9426 | 0.1712 | 0.9814 | 0.2252 |
| 宿舍和谐度 | 0.9810 | 0.8266 | 0.7255 | 0.1914 |
| 睡眠时间 | 0.9891 | 0.8542 | 0.7927 | 0.1880 |
| 是否失眠 | 0.9005 | 0.9163 | 0.1135 | 0.1806 |
| 是否饮酒 | 0.9632 | 0.8673 | 0.9891 | 0.2266 |
| 是否吸烟 | 0.9793 | 0.4874 | 0.9605 | 0.2018 |
| 是否关注健康 | 0.9221 | 0.8807 | 0.2656 | 0.1752 |
| 体育锻炼次数 | 0.9864 | 0.5174 | 0.362 | 0.1856 |
| 锻炼时间 | 0.9675 | 0.1990 | 0.9543 | 0.2050 |
| 体育爱好数量 | 0.7919 | 0.5057 | 0.3834 | 0.1859 |
| 文艺爱好数量 | 0.9203 | 0.1936 | 0.8752 | 0.2077 |
| 是否有压力 | 0.7424 | 0.6685 | 0.334 | 0.1818 |
| 压力来源 | 0.8403 | 0.2562 | 0.8141 | 0.2010 |
| 事情多否 | 0.9742 | 0.6818 | 0.8697 | 0.1903 |
| 解压方式 | 0.9468 | 0.1676 | 0.8874 | 0.2113 |
| 密切朋友 | 0.7783 | 0.8680 | 0.1082 | 0.1714 |
| 是否关注健康 | 0.9843 | 0.6329 | 0.8341 | 0.1982 |
| 与他人相处融洽 | 0.6842 | 0.3996 | 0.6305 | 0.1913 |

通过对上表的分析我们可以对影响博士生健康状况的各类行为习惯有如下分析：

第一，总体上看，行为习惯对亚健康类疾病的影响最为明显。按照疾病类型分类，各类行为习惯与 4 个疾病类型的关联度加和依次为：17.9604、11.181、13.4394、3.9237，并且各类行为习惯与亚健康类疾病关联度均值为 0.8980，可以说行为习惯与亚健康类疾病高度关联，从单项行为习惯上看，引发亚健康类疾病最大的行为习惯依次为睡眠时间、体育锻炼、健康关注度和寝室和谐程度。

第二、行为习惯对内脏器官类疾病的影响也较大，从单项行为习惯上看，引发内脏器官类疾病概率最大的几种因素依次为不吃早餐、饮酒、过度摄入肉食和吸烟，其他如睡眠、运动等因素也有一定影响。

第三、婚姻状况对男女性别类疾病影响最大。通过关联度分析和数据对比我们发现，未婚博士生的性别类疾病患病几率大大高于已婚博士生患病几率，从医学角度分析婚姻适龄人群在婚后性别类疾病患病概率会降低。其他行为习惯与男女性别类疾病的关联度呈现离散状态，经分析其原因可能是性别类疾病还与其他疾病相关联。

第四、各类行为习惯对其他类疾病的关联度较低。从医学角度来说，白化病等其他类型的疾病一般

由基因决定，与具体行为习惯并无太大直接关联，但也容易因其他疾病的发生而诱发。

上表的其他数据可进行后续深入分析，更可以作为对博士生的警示性参考。

## 五、对策与建议

博士生学业和科研任务繁重，同时其生活行为习惯和心理状态等都有着独立于其他群体的特殊性，婚恋、就业、社会等多元压力来源也促使其各类疾病的发生率高于其他群体。根据上述分析，博士生行为习惯对亚健康、内脏器官类疾病影响较为明显，为此各博士生培养单位可以有针对性开展工作，加强对博士生健康状况关注和提升的基础保证。

第一、加大宣传力度，提升博士生群体的健康意识。根据上述关联度分析，行为习惯与很多疾病的关联度较高，而行为习惯的养成源于健康意识。针对博士生的行为特点和关注焦点，有针对性开展传统媒体与新媒体相结合、线上宣传与线下活动相结合的宣传方式，全面提升博士生群体的健康关注。

第二、建立博士生参与体育锻炼的长效机制。科研任务较重是博士生群体体育锻炼较少的客观制约，各博士生培养单位应针对博士生相对集中时间，以科研基本单位、实验室、课题组等群体组织活动，提高体育健身活动的持续性。

第三、开展特色文化活动。根据博士生的多元压力，应开展纾解压力、提升健康水平的相关文化活动，如陶冶情操的文化艺术表演，以及情感类联谊、集体婚礼等活动，针对博士生心理需求进行有针对性的设计。

## 参考文献：


[1] 袁曾任.人工神经元网络及其应用[M].北京：清华大学出版社,1999.

[2] 钟淑瑛，李陶深.基于 MATLAB 的 BP-LVQ 神经网络组合分类器模型[J].计算机技术与发展，2006,16(2)：114-116.

[3] 葛哲学，孙志强.神经网络理论与 MATLAB R2007 实现[M].北京：电子工业出版社，2007.

[4] 董长虹.MATLAB 神经网络与应用[M].北京：国防工业出版社，2007.