Advances in Applied Mathematics 应用数学进展, 2022, 11(5), 3053-3059
Published Online May 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.115325

Hans汉斯

人工神经网络在传染病研究中的应用

汤达昌

南宁师范大学数学与统计学院，广西 南宁

收稿日期：2022年4月27日；录用日期：2022年5月21日；发布日期：2022年5月31日

摘　要

人工神经网络是在人工智能基础上发展而来的重要分支，对人工智能的发展具有重要的促进作用。我们对一传染病模型给出了分析，可以进行疫情分析和识别，预测传染病的暴发和流行。

关键词

神经网络，传染病模型

Application of Artificial Neural Network in Infectious Disease Research

Sitar Thammavongsa

College of Mathematics and Statistics, Nanning Normal University, Nanning Guangxi

Received: Apr. 27th, 2022; accepted: May 21st, 2022; published: May 31st, 2022

Abstract

Artificial neural network is an important branch developed on the basis of artificial intelligence and plays an important role in promoting the development of artificial intelligence. We give an analysis of an infectious disease model, which can be used for epidemic analysis and identification, and predict the outbreak and prevalence of infectious diseases.

Keywords

Neural Network, Infectious Disease Model

文章引用: 汤达昌. 人工神经网络在传染病研究中的应用[J]. 应用数学进展, 2022, 11(5): 3053-3059.
DOI: 10.12677/aam.2022.115325

1. 引言

在信息技术的推动下，人工神经网络的应用越来越广泛。由于人工神经网络具有自适应性、并行处理能力和非线性等优点，逐渐被应用于医学和生物学领域的研究。纵观人类的发展历史，传染病的发生和流行不仅影响到个人的健康，而且对社会稳定造成严重的影响。传染病流行病学数学模型主要针对传染病的自然史和流行规律，流行病学数学模型研究较之大规模的流行病学调查具有投入少，收效快，结果准确等特点。

2. 人工神经网络介绍

人工神经网络是根据生物学中人体的神经网络的结构和运行原理而建立起来的一种计算模型，是一种具有大量连接的并行分布式处理系统。通过模拟人脑的学习、记忆、处理问题等方式，神经网络可以通过学习获取相关知识，把知识存储在连接权中，通过不断的学习对知识进行调整，并且根据已经获得的知识处理相应的问题[1]。

生物学中，神经元细胞体周围有很多树突和一个轴突。树突和细胞体与其它神经元的轴突相接触，轴突连接到其它神经元的树突或细胞体上面。神经元传递信息靠的是脉冲传递，当一个脉冲传递到一个神经元的轴突末梢后，向突触间隙释放化学物质，形成电位。当下一个神经元细胞体的周围电位差累积到一个特定的电位，也就是阈值电位时，又会产生新的脉冲传递到轴突中去，如图 1 所示。

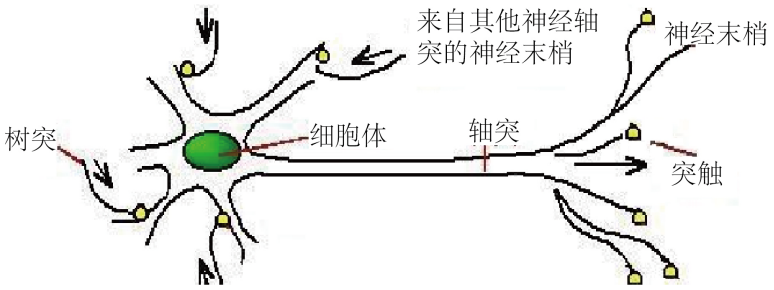

Figure 1. Biological neuron
图 1. 生物神经元

3. 人工神经网络的发展历程

(一) 萌芽时期

在 20 世纪 40 年代，生物学家 McCulloch 与数学家 Pitts 共同发表文章，第一次提出了关于神经元的模型 M-P 模型，这一理论的提出为神经网络模型的研究和开发奠定了基础，在此基础上人工神经网络研究逐渐展开。

(二) 低谷时期

在人工神经网络形成的初期，人们只是热衷于对它的研究，却对其自身的局限进行了忽视。Minskyh 和 Papert 通过多年对神经网络的研究，在 1969 年对之前所取得的研究成果提出了质疑，认为当前研究出

的神经网络只合适处理比较简单的线性问题，对于非线性问题以及多层网络问题却无法解决[2]。

(三) 复兴时期

美国的物理学家 Hopfield 在 1982 年提出了新的神经网络模型，并通过实验证明在满足一定的条件时，神经网络是能够达到稳定的状态的。

(四) 稳步发展时期

到 20 世纪 90 年代时，国内对于神经网络领域的研究得到了进一步的完善和发展，而且能够利用神经网络对非线性的系统控制问题进行解决，研究成果显著。随着各类人工神经网络的相关刊物的创建和相关学术会议的召开，我国人工神经网络的研究和应用条件逐步改善，得到了国际的关注。

4. 人工神经网络在传染病中应用分析

(一) 应用疾病的筛查和诊断

El-Solh 应用广义递归神经网络构建活动性肺结核诊断模型。共收集了 700 多例患者的人口统计学资料、临床症状、结核病暴露史、HIV 状态、结核菌素试验结果和临床诊断记录等资料。模型纳入了 21 个输入变量，分为 3 个隐含层，1 个输出层，模型输出提供一个活动性肺结核的似然估计结果。将模型的诊断结果与临床医生的诊断结果相比较，广义回归神经网络的灵敏度高，但特异度稍低，ROC 曲线显示广义回归神经网络的诊断结果好于临床医生的诊断，模型验证的诊断结果的 c-Index 为 92.3% (85.8%~99.1%)，而医生的诊断结果的为 71.6% (64.5%~78.9%)。该模型用于活动性肺结核能提供较精确的诊断结果。将人工神经网络应用于此类疾病的筛查和诊断，可减少成本，提高效率[3]。但不同人群不同目的需建立不同的模型进行诊断，如宋烨等 faze 建立了涂阴肺结核的诊断模型；Sham 等 37 将人工神经网络应用于 MRSA 的诊断，并与 Logistic 回归的结果相比较[4]。

在经典 ANN 模型中，简单单元，即 M-P 神经元模型。我们知道感知机和 Logistic 回归都是线性分类模型，它们的不同点在于分类函数的选取是不一样的。

我们令：

$z = w^{\mathrm{T}} x$ 。

感知机的分类决策函数：

$f(x) = g(z) = \mathrm{sign}(z)$

其中 $\mathrm{sign}(\cdot)$ 为阶跃函数：

$\mathrm{sign}(z) = 1 \quad \text{if} \quad z \geq 0 \quad \text{else } 1$

Logistic 回归的分类决策函数则是 Sigmoid 函数：

$$f(x) = g(x) = \frac{1}{1+\mathrm{e}^{-z}}$$

它表示的是将样本分类成正例和负例的几率比。也是一个阶跃函数的替代函数。传染病资料来源于某市疾病预防控制中心，共收集 2015~2020 年的传染病发病率。人口资料、气象资料由该市公安、气象部门提供，内容包括各年相应的人口数、出生率、平均气温、平均降水量等。采用 2015~2020 年的传染病发病率数据，按照 BPNN 原理将数据进行归一化处理后进行分析。其标准化后的变量分别为 x_1，x_2，x_3，x_4 和 Y。

(二) 人工神经网络模型构建

给定 n 个输入变量：$x_1, x_2, \cdots, x_n$ 以及相对应的权值变量 $w_1, w_2, \cdots, w_n$，一个传递函数 $f(\cdot)$，激发阈值变量 θ，输出变量为 y，有如下神经元模型(图 2)：

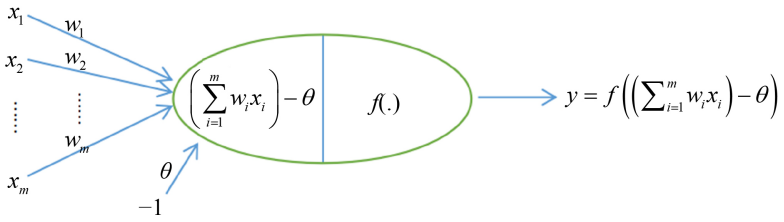

Figure 2. Neuron Model
图 2. 神经元模型

其中 x_1~x_m 这 m 个变量是与此神经元连接的上一层神经元的输出，或者为网络的原始输入变量。在实际操作中，可以将−1 看作此神经元的第 $m+1$ 个输入，把激发阈值变量 θ 作为相应的权值变量。神经元模型的传递函数 $f(\cdot)$ 一般采用 sigmoid 函数，给出表达式如下：

$$f(x)=\frac{1}{1+e^{-x}}$$

此表达式为 sigmoid 函数的单极形式，另也有双极形式的 sigmoid 函数。当神经元的加权输入和 $\sum_{i=1}^{m} w_i x_i$ 大于激发阈值 θ 时，神经元处于激发态，网络的输出 $f\left(\left(\sum_{i=1}^{m} w_i x_i\right)-\theta\right)$ 为正，否则为抑制态，输出为负。

当多个神经元组合起来时，人工神经网络的总体结构如下(图 3)：

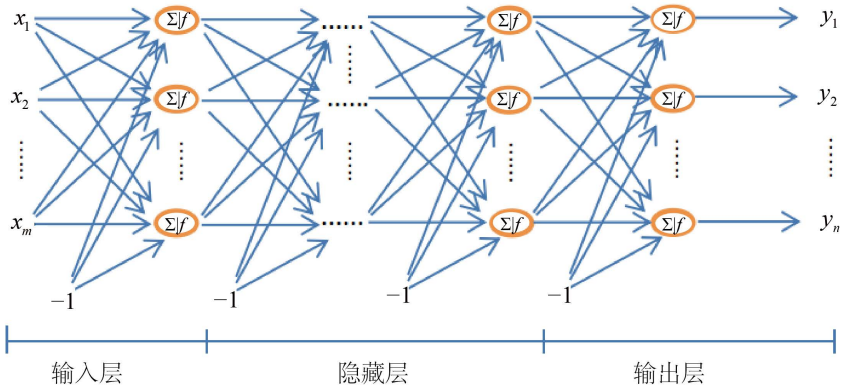

Figure 3. The overall structure of the artificial neural network
图 3. 人工神经网络的总体结构

为简洁起见，各层神经元之间的连接权值 w_{ij} 与激发阈值变量 θ_j 未在图中标出。以上是在全连接下的人工神经网络结构图。人工神经网络在本质上是由许多小的非线性函数组成的大的非线性函数，反映的是输入变量到输出变量间的复杂映射关系。映射的准确程度是由各层权值与各神经元结点的激发阈值变量共同决定的，同时也与人工神经网络的结构有关，结构变量包括隐藏层的层数与每层结点数，以及网络的连接状态是否为全连接的。

首先将人工神经网络中需要优化的变量——各层权值变量 w 与各神经元结点的激发阈值变量 θ 进行编码，表示成相应的目标函数。对于人工神经网络而言，运用进化算法优化的目标是，使网络的实际输出与理论输出之间的差值越小越好。

设网络共输入 K 个样本，每个样本的输出属性为 N 个，则网络总体误差 δ_{ANN} 可表示为

$$\delta_{ANN}=\frac{\sum_{k=1}^{K}\sqrt{\sum_{n=1}^{N}\left(\delta_{kn}\right)^2}}{2}$$

其中 δ_{kn} 表示第 k 个样本在第 n 个属性上的误差。网络优化的目标是确定一组权值 W 与阈值 Θ，使全局误差 δ 最小。通过以上分析，可以将网络的输出误差看作 W 与 Θ 的函数。如果将 Θ 对应的权值看作神经网络的额外输入连接，则可将 Θ 与 W 合并，记为 W_{exp}，表示扩展的权值向量组。由此可以得到以下的目标函数表达式：

$$\min\delta_{ANN}\left(W_{exp}\right)=\frac{\sum_{k=1}^{K}\sqrt{\sum_{m=1}^{M}\left(\delta_{kn}\left(W_{exp}\right)\right)^2}}{2}$$

接下来分析粒子的编码方式。对于一个具有 m 个输入与 n 个输出的人工神经网络，设共有 L 个隐藏层，层数编号依次为 $1,2,\cdots,L$，相应的每层结点总数为 $P1,P2,\cdots,PL$，则第 l 层第 p 个结点的编号为 lp，其中 $l\in(1,2,\cdots,L)$，$p\in(P1,P2,\cdots,Pl)$。权值与阈值的排列顺序按照输入层向输出层的方向排列，可以得到以下编码：

W_{exp}=[$w_{(11,21)}w_{(12,21)}\cdots w_{(1P1,21)}\theta_{11}$][$w_{(11,22)}w_{(12,22)}\cdots w_{(1P1,22)}\theta_{12}$]$\cdots$[$w_{(11,2P2)}w_{(12,2P2)}\cdots w_{(1P1,2P2)}\theta_{1P2}$][$w_{(21,31)}w_{(22,31)}\cdots w_{(2P2,31)}\theta_{21}$]$\cdots$[$w_{(21,3P3)}w_{(22,3P3)}\cdots w_{(2P2,3P3)}\theta_{2P3}$]$\cdots$[$w_{((l-1)1,l1)}w_{((l-1)2,l1)}\cdots w_{((l-1)P(l-1),l1)}\theta_{(l-1)1}$]$\cdots$[$w_{((L-1)1,LPL)}w_{((L-1)2,LPL)}\cdots w_{((L-1)P(L-1),LPL)}\theta_{(L-1)PL}$]

以上编码为直观起见，将 W_{exp} 用方括号进行了分段，其中每个方括号中的数组表示上一层所有结点对应下一层某一个结点的权值变量与该结点的阈值变量。将网络结构记为数组

$$S=[m,P1,P2,\cdots,PL,n]$$

表示输入层、隐藏层与输出层各自的结点数。若将 S 中的各个元素记为 $s_1,s_2,\cdots,s_{L+2}$，则将神经网络权值确定问题转换为粒子群算法的优化问题后，问题解空间的维度即 W_{exp} 的长度为：

$$D=\sum_{i=1}^{L+1}s_is_{i+1}+\sum_{i=2}^{L+2}s_i$$

式中等式右边的第一项为权值变量数，第二项为神经元结点的阈值变量数。

关于神经网络的隐藏层数与每层结点数的确定问题，本文统一采用单隐藏层，其中的结点数目参考经验公式

$$P=\left\lceil\sqrt{m+n}\right\rceil+a$$

来确定，其中 a 是一个取值介于 1~10 之间的整数常量。

关于测试数据集中训练集的与验证集的选取比例，可选用 3:2 的比例，即随机选取测试集中 60%的数据作为训练集，剩余的 40%为验证集。输入变量各个属性维度的数据分别采取最大-最小归一化方式，归一化区间为[0, 1]。

人工神经网络的评价指标主要采用均方误差 MSE 进行。定义均方误差的数学表达式如下：

$$MSE=\frac{\sum_{k=1}^{K}\sum_{n=1}^{N}\left(\delta_{kn}\right)^2}{K}$$

MSE 表示网络理论输出与实际输出之间的差距，除总体误差 MSE_{ann} 外，还有针对训练集的均方误差 MSE_{train} 与针对验证集的均方误差 MSE_{test}，分别衡量网络的拟合能力与泛化能力。此外也可用分类正确率进行衡量。

(三) 人工神经网络模型应用

将现有传染病的疫情资料和同期传染病流行影响因素资料按照原理将数据进行归一化处理后输入计算机，建立数据库。模型的建立及其应用通常需要一定数量的样本建模和对模型进行训练，并用一定数量的测试样本对模型进行检验。一般将现有疫情资料以地区或年度为单位，训练样本和测试样本比例按2:1，单纯随机方法进行分组。应用 SPSS 或 SAS 统计软件进行流行因素与传染病发病率相关分析：计算与传染病发病关系的相关系数。选取相关分析中与传染病发病率关系较为密切的因子作为自变量。

在网络的结构设计中，网络的层数确定为三层。网络输入层节点数就是系统的特征因子(自变量)个数，在本研究中为疫情影响因素变量(如人口数、出生率、平均气温、平均降水量等)；输出层节点为传染病(一个传染病流行年的患病率)；隐含层节点数设置为 4。网络的初始权值设置由神经网络软件随机产生，允许误差取 0.001~0.00001，迭代次数取 1000 次。

利用前一年标准化后的影响因素指标和前一年发病率为自变量，以当年的发病率为因变量训练网络和进行预测。经训练调整影响因素的权重，筛选影响传染病的主要流行因素，比较预测疫情和历史资料以达较高的拟合率，使网络的计算输出应变量与已知训练样本的应变量之差为最小。如果输出层没有得到期望的输出，则转入反向传播，通过修改各神经元的权值，减少误差，继续循环，直至网络误差收敛到规定的值内为止。

某市 2015~2020 年传染病发病率拟合值和实际值比较见表 1。用平均误差绝对值、MER 以及决定系数(R^2)指标检验模型的拟合程度，预测传染病发病率。结果显示，神经网络模型利用 2000~2005 年数据拟合的 MER 均较小，拟合误差率为 2.53%~13.46%，且决定系数扩接近于 1，表明模型预测效果较好。

Table 1. Incidence profile of three epidemic infectious diseases
表 1. 三种流行性传染病的发病概况

年份	乙肝阳性率(%)		丙肝阳性率(%)		梅毒阳性率(%)	
	男	女	男	女	男	女
2011	5.16	4.29	0.91	0.98	0	0
2012	5.49	4.09	0.11	0. 97	0.15	0.02
2013	4.72	3 .81	0. 78	0.62	0.99	0. 85

世界大约有三分之一的人口有乙肝病毒感染史，中国是拥有乙型肝炎病毒携带者最多的国家，约占世界总人口的 10%。我国也是丙肝的中高度流行区，而丙肝感染者约一半以上会发展为慢性肝炎。梅毒是由梅毒螺旋体引起的慢性的系统性性传播疾病，近年来发病率逐渐升高，对人们的生活质量造成巨大的影响。近年来，人工神经网络不断地应用于医学研究领域，取得良好效果。人工神经网络模型(ANN)及灰色理论在乙型肝炎(乙肝)，丙型肝炎(丙肝)及梅毒三种流行性疾病诊断中的应用，方法对 2011~2013 年 XX 第一附属医院三种流行性疾病进行统计。

搜集资料对 2011~2013 年在 XX 大学第一附院就诊的门诊病人的传染病检测结果(乙肝、丙肝、梅毒)进行统计，乙肝的检测人数共 95,254 例，男 32,667 例，女 62,587 例，丙肝的检测人数共 97,838 例，男 33,837 例，女 64,001 例，梅毒的检测人数共 57,050 例，男 15,035 例，女 42,015 例。

构建 BP 神经网络对三种流行病进行预测：基于 2011-01~2013-12 的每个月份时间为输入层，各个月份的发病率为输出层构建基于观察的 HP 神经网络预测模型，选 2011-01~2012-12 每个月份为训练样本的输入层参数，再选择 2013-0l~2013-12 每个月份为预测样本的输入层参数，各个月份的发病率为输出层，

将输入和输出层归一化。

构建灰色理论模型对三种流行病进行预测：通过上述方式可以计算 2013-01-12 乙肝、丙肝及梅毒三种流行病的 G(l, l)模型(表 1)。

三种流行性传染病的发病概况和趋势概况 2011~2013 年男性乙型肝炎病毒表面抗原检测阳性率分别为 5.16%，5.49%，4.72%，女性阳性率为 4.29%，4.09%，3.84%，男性丙型肝炎病毒抗体检测阳性率分别为 0.91%，1.11%，0.78%，女性阳性率为 0.98%，0.97%，0.62%，男性梅毒螺旋体抗体检测阳性率分别为 0%，0.15%，0.99%，女性阳性率为 0%，0.02%，0.85%。

值得警惕的是梅毒螺旋体抗体阳性率的上升，这种性病在我国 20 世纪末曾几乎消失，然而 2013 年的男性梅毒螺旋体抗体阳性率达 0.99%，女性达 0.85%，已经超过此年度丙肝抗体的阳性率，其阳性患者主要集中在青壮年人群，这与外来文化的冲击以及人们的理念密不可分。梅毒主要通过性接触途径传播，我们应该重视并采取相应的手段措施，并密切监控疾病的发展流行状况，而此时通过模型对疾病的发病率做出预测就显的极为重要。

在疾病的预防控制工作中，如果能简单并准确预测出流行病的发病趋势，不仅能够提供直观的参考数据，也为流行病的防控提供了巨大帮助，国家可以及时地采取相关措施，最大限度地控制病情发展。

5. 结语

人工神经网络虽然是一种处理非线性问题的好方法，但在传染病研究的应用还处于探索阶段，一些问题有待解决，如变量的筛选和假设检验方法；权重系数的假设检验，计算权重系数的可信区间，含隐含层时权重系数的流行病学意义；输入变量的选择；人工神经网络的类型和结构的选取等问题都还需要进一步研究。人工神经网络最有用的特性之一是应用时对分析问题的概率模型不要求通过演绎作出假设，并具有逼近任意连续函数和非线性映射的能力，进行高维非线性的精确映射。因此。人工神经网络在传染病研究中的应用不但可以进行疫情分析和识别，预测传染病的暴发和流行情况、评价防治效果，也可对某种疾病的流行情况进行模拟，对患者进行筛查和诊断，对预期的经济损失进行评估等，有较高的应用和推广价值。

参考文献

[1] 蒋思瑶. 基于统计学习理论的传染病预警方法研究比较[D]: [硕士学位论文]. 大连: 辽宁师范大学, 2016.

[2] 侯瑞生, 陈文玲, 赵盛, 薛云红, 张蕾, 李丽, 王凯娟. 人工神经网络在传染病疫情分析与预测中的应用[J]. 旅行医学科学, 2018(2): 31-33.

[3] 王峰. 基于灰色神经网络模型的全国病毒性肝炎发病率预测[D]: [硕士学位论文]. 太原: 山西医科大学, 2012.

[4] 刘艳, 郑彬, 邬文亮, 胡建利, 朱叶飞, 刘文东. RBF 神经网络在甲肝流行趋势预测中的应用研究[J]. 江苏预防医学, 2019(1): 7-10.