

基于 BP 神经网络的乳腺癌患者存活概率的预测分析^①

李晓峰^{1,2}, 杨春山³, 王雪梅³, 丁树春⁴

¹(黑龙江外国语学院 信息科学系, 哈尔滨 150025)

²(北京理工大学 计算机科学与技术学院, 北京 100081)

³(东北农业大学成栋学院 计算机科学与技术系, 哈尔滨 150025)

⁴(黑龙江大学 电子工程学院, 哈尔滨 150080)

摘 要: 本文提出了对乳腺癌的知识进行挖掘及对其有效的网络予以探讨。在应用各种数据挖掘方法之前,利用所开发的网络找出病情发展的概率。有关结果有助于医生针对病人的病情进行合理治疗。为克服数据集的高维度问题并实现数据之间的关联性,本文采用主成分分析法来降低数据维度并找出适用的网络。运用 BP 神经网络结构进行了评估,对北京某医院的乳腺癌数据方面 BP 神经网络的性能进行了分析。结果表明主成分分析消除了网络输入之间的相关性,降低了网络的输入层数,改善从整体上提高了网络的性能。最终取得了良好的预测结果。

关键词: 乳腺癌; 数据挖掘; 主成分分析; BP 神经网络

Analysis of Predict the Probability of Breast Cancer Survival Based on BP Neural Network

LI Xiao-Feng^{1,2}, YANG Chun-Shan³, WANG Xue-Mei³, DING Shu-Chun⁴

¹(Department of Informatics and Science, Heilongjiang International University, Harbin 150025, China)

²(School of Computer Science and Technology, BIT, Beijing 100081, China)

³(Department of Computer Science and Technology, Cheng-Dong College of Northeast Agricultural University, Harbin 150025, China)

⁴(School of Electronic Engineering, Heilongjiang University, Harbin 150080, China)

Abstract: This paper presents on breast cancer knowledge mining and network to its effective to explore. Before the application of various data mining method in probability, find out the progression of the disease by using the network. The results are helpful to reasonable treatment to the doctor for the patient. For the high dimension data set and overcome the problems associated with the data, this paper uses principal component to reduce the data dimension and find out the suitable network analysis. Application of BP neural network structure is assessed, performance on the breast cancer data in a hospital in Beijing, the BP neural network is analyzed. The results show that principal component analysis to eliminate the correlation between the input of the network, reducing the input of network, improve the performance of the network as a whole. The good results have been achieved.

Key words: breast cancer; data mining; principal of component analysis; BP neural network

1 引言

当今社会上 8 个女性中约有一位在其生命中有可能患上乳腺癌。医疗科学方面的大问题主要是病情的诊断基于对病人的各种检查。检查结束后要进行最终诊断就可能不那么容易了,即便对医疗专家也是如此。过去的几十年里,人们发明了计算机化的诊断工具,旨在协助医生弄清那些令人混淆不清的诊断数据^[1]。

为了对乳腺癌进行分类,人们已经进行了各种研究(如^[2-3]所述)。对北京某医院乳腺癌数据组的病人进行癌症诊断后,他们就接受了手术;鉴于此,本文利用这些乳腺癌数据,建立 BP 神经网络对乳腺癌的病变类别的进行了预测。同时,应用主成分分析法对影响乳腺癌的病变类别的 9 个属性进行分析,对各属性之间互相关联的复杂关系进行简化,从而用较少的新属性代替原来较多的属性,并且使这些较少的新属性

① 基金项目:国家青年科学基金(61102071);教育部教师科研专项基金(CTF120771)

收稿时间:2013-09-26;收到修改稿时间:2013-10-30

尽可能多地保留原来较多的属性所反映的乳腺癌信息,并新属性数据构造 BP 神经网络的输入样本,建立基于主成分分析预测模型,提高了预测的精度。

我们从北京某医院获取了 599 关于乳腺癌数据的样品,其中完整数据是 583 份、剩余 16 份为样品。这些数据分属 9 种整数属性且每个数据值从 1—10 不等,9 种属性分别是:

- ①肿块厚度
- ②细胞大小均匀程度
- ③细胞外观均匀程度
- ④边际粘附性
- ⑤单上皮细胞大小
- ⑥裸核
- ⑦布兰德染色质
- ⑧正常核仁
- ⑨有丝分裂

以上这些属性用以衡量乳腺癌九种不同程度的外观和内在染色体的变化。在乳腺癌的病变类别里有两个值,0 表示确诊为良性(非癌性),1 表示确诊为恶性(癌性)。

2 数据挖掘技术

数据挖掘是通过利用各种数据分析工具对数据模式和之间关系进行发掘的一个过程。这些数据可用来进行有效预测。其中第一步也是最简单的一步分析就是数据描述。数据挖掘可以用统计属性来概括,通过图表直观反映出来。接下来就是在各变量之间寻找可能有意义的关联。收集、挖掘和筛选正确的数据是至关重要的。但,数据描述无法单独提供一个行动方案。必须基于通过已知结果而建立的模型来构建一个预测模型,然后根据初始样品之外的结果来对模型进行检测。一个好的模型应该是一种有效的指导,协助完成所从事的工作。最后一步就是通过实际经验来验证模型。数据挖掘成功的秘诀有两个即:

①根据问题的框架提出一个确切的解决办法;正所谓“主题突出、收效显著”。

②使用正确的数据。从获取到的数据中筛选一些后,你可能需要通过一些重要途径对之进行转换和组合。

2.1 主成分分析(PCA)

有时输入向量的维度较大,但这些向量的成分却

紧密关联(冗余)。这种情况下通过一些投影法来降低输入向量的维度比较凑效。PCA 法是降低大数据集维度的有效方法。它允许对变量之间的关联性进行鉴定,从而降低数据集的维度^[4-5]。该法所达到的效果有以下三种:

(1) 通过将输入向量的成分正交以生成无关联的正交变量或通过初始关联变量乘以特征向量以生成主成分(PCs);

(2) 对所生成的正交成分(主成分)进行排序以便变差最大的成分居前;

(3) 淘汰数据集中变差最小的那些成分。输入向量应正常化,那样,它们才带有零均值和单位方差。这是一个标准的过程,我们在使用主成分分析的过程中需遵循。

2.2 主成分的简化算法

设有 n 个样本数据,每个样本数据有 p 个指标: $x_1, x_2, \dots, x_p$, 建立原始数据矩阵:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad (1)$$

则,主成分分析法处理数据的步骤简要如下:

(1) 对原始数据 X 标准化处理为 X^* , 标准化过程如下:

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{\sqrt{\text{var}(x_j)}} \quad (i=1, 2, \dots, n; j=1, 2, \dots, p) \quad (2)$$

$$\text{式(2)中} \quad \bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij} \quad (j=1, 2, \dots, p)$$

$$\text{var}(x_j) = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \quad (j=1, 2, \dots, p)$$

(2) 由标准化数据 X^* 求出相关系数矩阵 $R_{p \times p}$ 。

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix} \quad (3)$$

式(3)中, $r_{ij} (i, j=1, 2, \dots, p)$ 是原来变量 x_i^* 和 x_j^* 的相关系数,且 R 是实对称矩阵^[6]。

(3) 计算特征值、主成分贡献率和累计方差贡献率, 确定主成分个数^[7].

首先解特征方程 $|\lambda I - R| = 0$ (I 表示单位向量, R 表示相关系数矩阵) 求出特征值 $\lambda_i (i=1, 2, \dots, P)$, 因为 R 是正定矩阵, 所以其特征值 λ_i 都为正数, 并将其按大小顺序排列. 特征值是各主成分的方差, 它的大小反映了各个主成分的影响力. 主成分 Z_i 的贡献率为:

$$W_i = \lambda_i / \sum_{j=1}^p \lambda_j \quad (4)$$

贡献率越大, 说明该主成分所包含的原始变量的信息越强. 累计贡献率为:

$$\sum_{i=1}^m \lambda_i / \sum_{i=1}^p \lambda_i \quad (5)$$

根据选取主成分个数的原则, 特征值要求大于 1 或者累计贡献率达 80%-95% 的特征值 $\lambda_i (i=1, 2, \dots, m)$, 其中整数 m 即为主成分的个数.

(4) 计算主成分载荷.

主成分载荷是反映主成分 Z_i 与原变量 x_i^* 之间的相互关联程度, 在 SPSS 软件中主成分分析后的分析结果中, “成分矩阵” 反应的就是主成分载荷矩阵^[8].

(5) 计算主成分得分.

计算样品在 m 个主成分上的得分, 就可以得到各主成分下的各个样品的新数据, 即为主成分得分:

$$\begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1k} \\ F_{21} & F_{22} & \cdots & F_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ F_{n1} & F_{n2} & \cdots & F_{nk} \end{bmatrix} \quad (6)$$

依据主成分得分的数据, 就可以进行进一步的分析, 作为神经网络的训练数据.

2.3 BP 神经网络的结构

神经网络因为可以为繁多、复杂问题的建模提供有效的解决办法, 如今格外受关注. BP 神经网络是应用最为广泛的一种, 其具有自适应、自组织、自学习的能力、非局域性和非凸性的突出优点. BP 神经网络在工作过程时分学习期和工作期. 学习期由输入信息的正向传播和误差的反向传播组成^[9]. BP 神经网络的学习过程如下图所示.

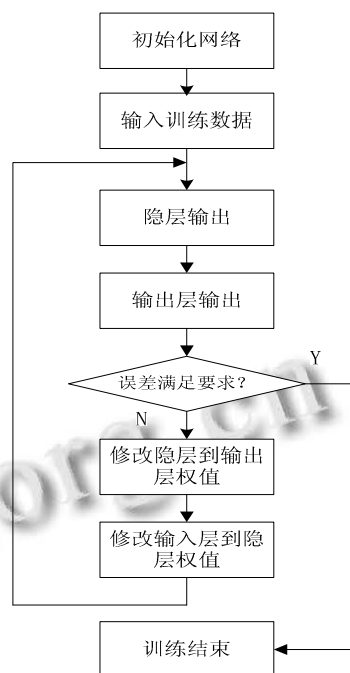


图 1 BP 神经网络的学习过程

(1) 初始化网络及学习参数. BP 神经网络的学习是典型的有师学习, 在学习前需要设置一些参数, 如初始权值、阈值矩阵、学习因子等.

(2) 前向传播过程.

对给定的训练样本输入, 计算网络的实际输出, 并与期望输出比较, 若满足误差要求, 则保存当前权值、阈值矩阵, 若不满足误差要求, 则转入误差反向传播过程.

(3) 反向传播过程.

① 计算同一层单元的误差.

② 修正权值和阈值.

③ 返回(2), 新的权值、阈值进行前向传播

BP 神经网络在工作期中只有输入信息的正向传播. 正向传播的计算按前述神经元模型工作过程进行.

3 实验与结果分析

本文的根本目的在于采取针对性算法来构建乳腺癌数据的有效网络.

① 对获取到的乳腺癌数据集进行一切调查, 获取到各种方法的初始评估结果.

② 选择适合于这一网络的合适规模. 如果规模太小, 就无法构造解决问题的理想模型. 所以, 这种网络对剩余的实际数据或任何其它不可见数据的性能

检测就不理想，更别提试验集方面的数据了。实际中，“最佳规模”的网络应是经过反复试验而来的。

3.1 BP 神经网络预测

要建立一个 BP 神经网络模型，就是要确定输入层、输出层和隐含层。本文建立的是带有一个隐层的三层 BP 神经网络模型。因为，具有一个隐层的三层 BP 神经网络，只要隐层的节点足够多，就能以任意精度逼近到有界区域上的任意连续函数。

我们对北京某医院乳腺癌数据的样品数据进行了 BP 神经网络的试验。乳腺癌数据样品具有 9 种属性，乳腺癌病变类别预测模型的输入层神经元确定为 9 个，此模型输出层的神经元只有一个，即 0(良性)或 1(恶性)。隐层神经元的个数采用“试凑法”来确定。学习率 $\eta=0.01$ ，最大迭代次数为 500，允许精度 e^{-5} ，经试算当隐层神经元的个数为 20 时效果最佳。

对于 583 完整乳腺癌数据样品，其中 500 数据样品用于训练神经网络，83 数据样品作为测试数据，验证 BP 神经网络的性能，得到预测值和实际值的曲线如图 2 所示。理想情况下，预测值和实际值应该是完全重合的，但从图 2 中可以看到，对 83 测试数据预测时并不完全重合。

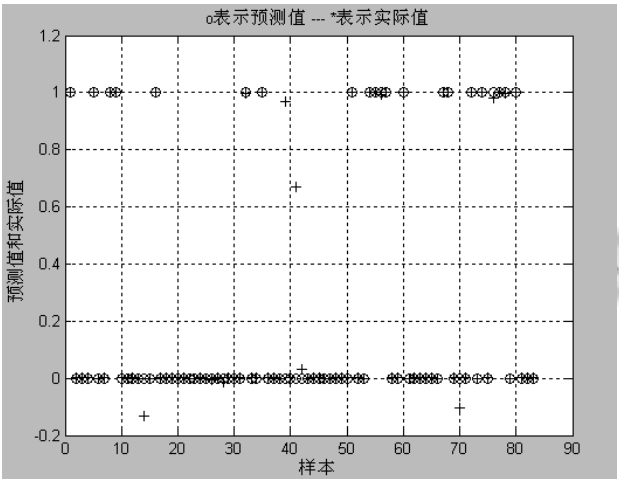


图 2 BP 神经网络预测值和实际值

BP 神经网络预测值和实际值之差，即误差曲线如图 3 所示。理想情况下误差曲线应该是零。但图 3 中最大误差为 -1，说明预测时将良性误判为恶性。

3.2 基于主成分分析的 BP 神经网络预测

3.2.1 主成分分析的必要性

主成分分析是对有较强相关性的自变量采用正交化

的方法解决自变量间的强相关关系。所以主成分分析的前提是变量 $x_1, x_2, \dots, x_p$ 之间的强相关性，如果 $x_1, x_2, \dots, x_p$ 之间正交，则它们之间就不会存在公共因子，因此作主成分分析就失去意义了。所以在进行主成分分析之前，必须先检验 $x_1, x_2, \dots, x_p$ 之间的相关性。

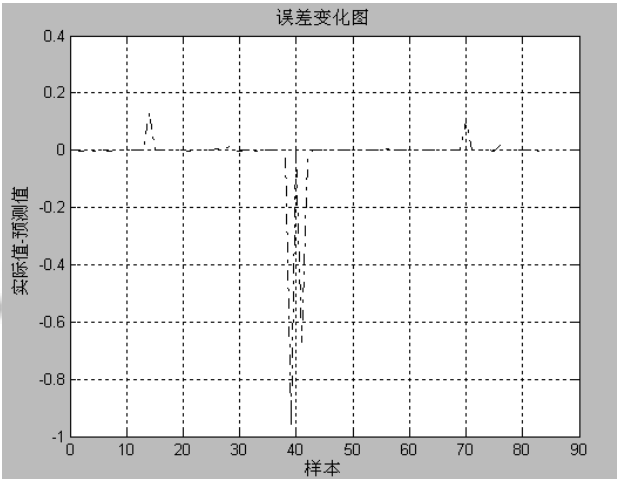


图 3 误差变化图

一般通过 KMO(Kaiser-Meyer-Olkin)检验统计量判定自变量间的相关关系。KMO 值通过 SPSS 软件计算，其值在 0~1 之间。KMO 值越接近 1，意味着变量间的相关性越强，越适合于主成分分析。反之，KMO 值越接近 0，意味着变量间的相关性越弱，则不适合作主成分分析。现在没有针对 KMO 测度的显著性检验。一般采用如表 1 的主观判断。

表 1 通过 KMO 检验是否适合主成分分析

KMO 值	是否适合主成分分析
0.9~1.0	非常适合
0.8~0.9	很适合
0.7~0.8	适合
0.6~0.7	不太适合
0.5~0.6	勉强
0.5 以下	不适合

使用 SPSS 软件计算出乳腺癌数据的 9 个属性间的 KMO 值为 0.935，所以对乳腺癌数据进行主成分分析是非常有必要的。

3.2.2 对数据样品主成分分析

使用 SPSS 软件，对 583 完整乳腺癌数据样品进行主成分分析可得到 9 个属性的特征根，其中特征根按从大到小排序，得到贡献率和累积贡献率，如表 2 所示。

表 2 解释的总方差

成份	特征值	贡献率 %	累积贡献率 %
1	5.899	65.550	65.550
2	0.776	8.622	74.172
3	0.539	5.992	80.163
4	0.460	5.107	85.270
5	0.380	4.225	89.496
6	0.302	3.354	92.850
7	0.294	3.271	96.121
8	0.261	2.897	99.018
9	0.088	0.982	100.000

通常主成分个数 m 的取值标准是使得累积贡献率大于 85% 即可, 为此选取前 5 个主成分(见表 1). 5 个主成分的主成分载荷矩阵见表 3, 其中每一个载荷量表示主成分与对应变量的相关系数.

表 3 主成分载荷矩阵

主成份 1	主成份 2	主成份 3	主成份 4	主成份 5
0.734	-0.124	0.636	0.073	0.050
0.925	-0.041	-0.015	-0.138	-0.090
0.917	-0.073	0.025	-0.119	-0.067
0.808	-0.046	-0.303	0.334	-0.012
0.817	0.145	-0.064	-0.290	-0.393
0.814	-0.230	0.001	0.338	-0.077
0.840	-0.201	-0.156	0.009	0.140
0.815	0.030	-0.099	-0.283	0.426
0.559	0.798	0.059	0.176	0.065

用主成分载荷矩阵中的数据除以主成分对应的特征值的平方根, 便得到 5 个主成分中每个指标所对应的系数, 进行处理后得到的 5 个主成分的表达式为:

$$F_1 = 0.3025X_1^* + 0.3812X_2^* + 0.3779X_3^* + 0.333X_4^* + 0.3367X_5^* + 0.3354X_6^* + 0.3462X_7^* + 0.3359X_8^* + 0.2304X_9^* \quad (7)$$

$$F_2 = -0.1407X_1^* - 0.0465X_2^* - 0.0829X_3^* - 0.0522X_4^* + 0.1646X_5^* - 0.2611X_6^* - 0.2281X_7^* + 0.0341X_8^* + 0.9057X_9^* \quad (8)$$

$$F_3 = 0.8427X_1^* - 0.0199X_2^* + 0.0331X_3^* - 0.4015X_4^* - 0.0848X_5^* + 0.0013X_6^* - 0.2067X_7^* - 0.1312X_8^* + 0.0782X_9^* \quad (9)$$

$$F_4 = 0.1076X_1^* - 0.2034X_2^* - 0.1754X_3^* + 0.4923X_4^* - 0.4275X_5^* + 0.4928X_6^* + 0.0133X_7^* - 0.4171X_8^* + 0.2594X_9^* \quad (10)$$

$$F_5 = 0.0811X_1^* - 0.146X_2^* - 0.1087X_3^* - 0.0195X_4^* - 0.6374X_5^* - 0.1249X_6^* + 0.2271X_7^* + 0.691X_8^* + 0.1054X_9^* \quad (11)$$

式中, $X_i^* (i=1, 2, \dots, 9)$ 为乳腺癌数据经过标准化处理之后的值. 按式(7)~式(11)进行变换, 分别对应样品数据 9 种整数属性经主成分分析变换后的 5 个主成分作为 BP 神经网络的训练数据.

3.2.3 基于主成分分析的 BP 神经网络预测

乳腺癌数据样品具有 9 种属性, 经主成分分析变换后为 5 个主成分. 此时, 乳腺癌病变类别预测模型的输入层神经元确定为 5 个, 输出层的神经元还是只有一个, 即 0(良性)或 1(恶性). 学习率 $\eta=0.01$, 最大迭代次数为 500, 允许精度 e^{-5} , 经试算当隐含层神经元的个数为 12 时效果最佳.

对于 583 经主成分分析变换后的完整乳腺癌数据样品, 其中 500 数据样品用于训练神经网络, 83 数据样品作为测试数据, 验证 BP 神经网络的性能, 得到预测值和实际值的曲线如图 4 所示. 预测值和实际值基本完全重合, 提高了预测的精度.

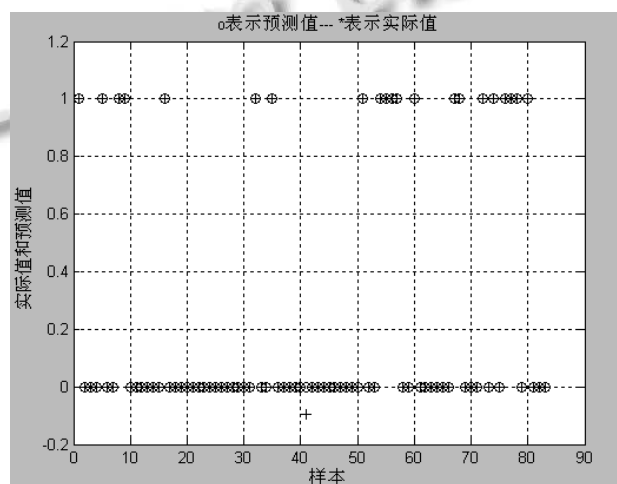


图 4 基于主成分分析的 BP 神经网络预测值和实际值

预测值和实际值之差, 即误差曲线如图 5 所示. 最大误差为仅为 0.1.

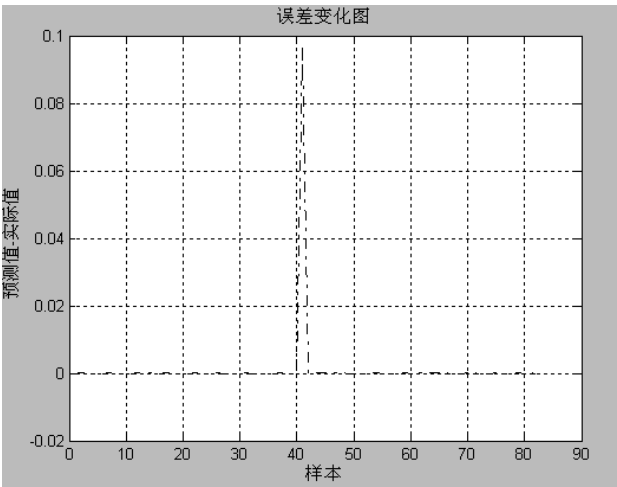


图 5 误差变化图

3.2.4 两种预测模型的对比

这里将使用 9 种属性的乳腺癌数据进行预测的模型称为直接预测模型。将使用主成分分析变换后的乳腺癌数据进行预测的模型称为基于主成分分析预测模型。

对两种预测模型，各训练和预测 50 次，得到平均的 MSE(均方误差)和预测准确率。如表 4 所示。

表 4 两种预测模型的对比

预测模型	MSE	预测准确率
直接预测模型	0.017	96.87%
基于主成分分析 预测模型	1.1187e-004	99.4%

3.2.5 不同训练算法的检验

根据表 5 所给出的算法，我们进行了有关研究，用缩略词来表示相应的研究结果。我们应用四种不同的算法来检测 BP 神经网络的性能。

表 5 性能检测算法及对应的缩略词

缩略词	算法
LM	trainlm – Levenberg - Marguardt
RP	trainrp – Resilient Backpropagation
CGB	traincgb – Conjugate Gradient with Powell Beale Restarts
OSS	trainoss – One-Step Secant

表 6 给出的是上述四种不同的算法的试验检测结果。表 1 中的每个入口分别代表了 50 次不同的试验，每次试验都用到了不同随机性的初始权值。每个案例

中，我们都要对该网络进行反复试验直至均方误差小于 0.01 为止。其中有几次运行没能成功兼容一些算法，所以，我们只利用前 75%的运行数据进行统计。

表 6 各种算法对 BP 神经网络性能的检测结果

算法	MSE	预测准确率
LM	1.1187e-004	99.4%
RP	1.7011 e-004	99.07%
CGB	0.0013	97.85%
OSS	0.0265	96.87%

4 结论

通过这所医院获取到乳腺癌方面的实际诊断数据后，对这方面的不同分类的神经网络的应用进行了一些深入研究。通过应用 BP 神经网络，构建了一个诊断系统，由于 BP 神经网络模型具有局部逼近的特征和较强的非线性映射能力，因此它能够较好地模拟具有较强非线性变化特点的乳腺癌预测问题。并将主成分分析法与之结合，用于乳腺癌的病变类别的诊断。

采用主成分分析法 583 对乳腺癌样本数据进行处理，提取 5 个主成分作为神经网络的输入。构造了两种 BP 神经网络，分别采用没有经过任何处理的数据和经过主成分分析预处理后的主成分数据进行网络的训练和识别。结果表明经过主成分分析预处理之后的 BP 神经网络，由于消除了网络输入之间的相关性，改善了程序执行效率，从整体上提高了网络的性能。识别精度明显优于另外一种网络。

综上所述，BP 神经网络将来会不断应用众多医学领域。为了适应现代医学的发展，在乳腺癌临床医疗中，不能依靠经验决策来预测存活率，而要采用科学的决策来预测存活率。在冗繁医疗临床工作中，实现 BP 神经网络诊断系统研发与应用，对临床医学诊断具有重要价值。由于 BP 神经网络仅是临床医师最终诊断的一种辅助手段，不能取代医学临床诊断，因而有待在未来医学研究和实践中不断探索与完善。在 BP 神经网络应用研究中，临床医者应根据乳腺癌患者的病因需求支持研发出更精准 ANN 算法，并通过提供准确数据样本，使基于 BP 神经网络技术在医疗临床中真正作为一种高精度，高效率辅助治疗手段，为现代医疗服务。神经网络必将得到临床医者认同并为临床工作带来诸多便利。

参考文献

- 1 袁前飞,蔡从中,肖汉光等.基于支持向量机的乳腺癌预后状态预测和疗效评估.北京生物医学工程,2007,26(4):372-376.
- 2 杨冬风,杨冬秀.粗糙神经智能疑似乳癌图像分类方法研究.计算机工程与应用,2010,46(12):188-191.
- 3 王枚,王国宏.基于主成分分析的目标确认方法及其在车标定位识别中的应用.南京理工大学学报(自然科学版),2009,33(1):42-46.
- 4 张洪波,何怡刚,周炎涛等.主成分分析法与概率神经网络在模拟电路故障诊断中的应用.计算机测量与控制,2008,16(12):1789-1792.
- 5 Agarwal S, Jacobs Jr. DR, Vaidya D, et al. Metabolic syndrome derived from principal component analysis and incident cardiovascular events: The multi ethnic study of atherosclerosis(MESA) and health, aging, and body composition (Health ABC). Cardiology Research and Practice, 2012, Article 919425.
- 6 龙训建,钱鞠,梁川.基于主成分分析的 BP 神经网络及其在需水预测中的应用.成都理工大学学报(自然科学版),2010,37(2):206-210.
- 7 张文峰,史忠科.主成分分析法在城市道路交通安全评价中的应用.计算机工程与应用,2007,43(36):246-248.
- 8 陈建宏,刘浪,周智勇等.基于主成分分析与神经网络的采矿方法优选.中南大学学报(自然科学版),2010,41(5):1967-1972.
- 9 程玉桂,黎明,林明玉.基于遗传算法和 BP 神经网络的城区中长期电力负荷预测与分析.计算机应用,2010,30(1):224-226.