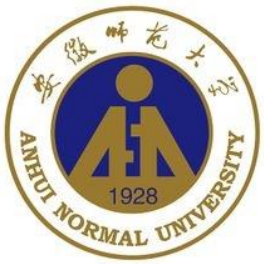
数学建模协会

队号：12



安徽师范大学

hui Normal University

2018 年数学建模暑期培训

参赛试题：B

指导老师：_____

	组　长	组　员	组　员
姓　　名	蒋舒悦	叶舟	陈苏婉
学　　号	16111201024	16111204099	16111201006
年　　级	2016 级	2016 级	2016 级
学　　院	数学与统计学院	计算机与信息学院	数学与统计学院
专　　业	数学与应用数学	计算机科学与技术	数学与应用数学
联系方式	17775484819	13093628805	17754836418

2018 年 8 月 30 日

基于BP神经网络的信用评价模型

摘要

本文针对小额信贷业务申请个人欺诈和逾期风险的精准识别问题，运用了深度学习、数据处理及量化分析等理论，构建了主成分分析法、BP 神经网络等模型，综合运用 MATLAB、SPSS 和 Dev-c++ 等软件编程求解，得到信用综合预测模型.

本文的特色是对冗余指标进行筛选以及对 BP 神经网络进行改进，结合附加动量法和弹性梯度下降法对阈值和权值进行改进，并对网络学习速度进行改进，提高神经网络训练速度.

针对题目，先对数据进行归一化处理统一量纲，利用主成分分析法对冗余指标进行筛选，筛选出 18 项主成分，对预测数据进行预测后计算精度，得到在此模型下，总预测精度为 69.03%，逾期预测精度为 25%，未逾期预测精度 79.68%，此时精度较低.为提高预测精度选择 BP 神经网络对模型进行改进，首先对指标进行筛选，通过选择平均相关性小于 0.1 的 78 项指标对数据进行处理，利用 BP 神经网络对测试样本是否逾期进行预测，得到总预测最高精度为 83.29 %，但此时逾期预测精度和未逾期预测精度较低.由于信贷机构对于逾期预测精度较为重视，为提高逾期预测精度，通过增加训练次数，修改训练组样本数，将逾期预测精度提高到 86.7%，未逾期预测精度提高到 73.7%，总预测精度为 78.3% .此时模型逾期预测精度比未优化前提高了 80%，总体拟合程度达到 0.77948，比优化前提高 0.21765.

关键词： BP 神经网络，主成分分析，数据处理，弹性梯度下降法

1. 问题重述

1.1 背景知识

1）信用体系

个人信用体系是社会主义市场经济和社会和谐发展的必要条件.建立起我国较为完善的个人信用体系，进而规范并促进社会经济活动，是一项社会、经济意义深远而且关系到治理体制创新、政府职能转变、法制建设规范等方面规模浩大的工程.因此，个人信用体系建设具有极其重要的意义.

2）市场现状

从全国个人信用体系建设的实践来看，目前我国个人信用体系建设有了一定的进展，但也存在着相当多的问题，越来越重要的信用记录与信用记录的缺失之间的矛盾日益激化，建立完善的信用体系迫在眉睫.随着近年来面向个人的小额贷款业务的不断发展，防范个人信贷欺诈，降低不良率是开展相关业务的首要目标.

3）研究意义

调动社会全员的大数据建模的创新积极，帮助金融机构准确评估个人信用情况，实现对小额信贷业务申请个人欺诈和逾期风险的精准识别，进一步提升金融机构防范欺诈和降低不良率的能力.

1.1 相关数据

1） *model_sample*（见原题附件：信用贷款数据）

2）参数表（见原题附件：信用贷款数据）

3）字段解释（见原题附件：信用贷款数据）

1.2 具体问题

根据给定的训练集数据开发信用评估模型，并依据模型结果给出是否逾期的预测结果，然后依据模型预测验证集中的贷款申请人是否逾期.

2. 问题分析

根据题目所给出的 11017 组个案数对每组个案数的 199 项指标进行综合分析，先通过主成分分析法对冗余指标进剔除，算出每项指标所所占的权重值，再对个案进行评价，预测客户是否逾期；再构建 *BP* 神经网络模型对样本进行训练后对测试集进行预测，不断改进算法，提高逾期预测精度，未逾期预测精度和总预测精度.

3. 模型假设

①假设所有的数据真实有效，具有统计分析价值.

②假设用户是否逾期只与题目所给的 199 项指标相关.

③数据调查来源于某机构，由该机构的客户信用现象大致代表大部分全体.

④机构客户信用情况在一定时间内保持稳定不变，且不受例如金融危机之类的外界因素的影响

4. 名词解释与符号说明

4.1 名词解释

①信用贷款：指以借款人的信誉发放的贷款， 借款人不需要提供担保.其特征就是债务人无需提供抵押品或第三方担保仅凭自己的信誉就能取得贷款，并以借款人信用程度作为还款保证的.

②信用评价：以一套相关指标体系为考量基础，标示出个人或企业偿付其债务能力和意愿的过程.

③总预测精度=$\frac{\text{正确预测样本数}}{\text{总预测样本数}}\times 100\%$

④逾期预测精度=$\frac{\text{正确预测逾期样本个数}}{\text{逾期样本个数}}\times 100\%$

⑤未逾期预测精度=$\frac{\text{正确预测未逾期样本个数}}{\text{未逾期样本个数}}\times 100\%$

4.2 符号说明

序号	符号	符号说明
1	F	综合评价得分
2	D	主成分对方差的累积贡献率
3	net	净激活量
4	p	隐含层神经元个数
5	q	输出层神经元个数为
6	X_n	输入向量
7	f	激活函数

5. 模型建立与求解

5.1 数据预处理

题目给出的11017组数据中有199项指标，为通过199项指标合理预测11017组数据是否会逾期，首先对数据进行处理，筛选相关性不高的变量，剔除冗余指标[1].

1) 筛选方差为零指标与缺失值处理

题目给出的199项指标中，有过多冗余指标，首先求解每个指标的均值后求出方差，发现有一个指标具有零方差.经检验后发现对于指标m_{12}健康险标志每一组数据均为0，故删除该指标，剩余198项指标.经过观察数据发现，大量数据中具有缺失项，选择合适均值插补法将每组数据中的缺失项补充完整.

2) 归一化处理

由于题目中所提供的各个省市的198项指标的量纲并不统一，为提高预测模型的准确率，我们使用*SPSS*软件对其进行归一化处理，使得各指标量纲统一.并对操作所得的新变量重新命名为$Z_1, Z_2, \cdots Z_{198}$.所得部分结果见表1.

表1 归一化处理后部分数据

个案标记	Z_1	Z_2	Z_9	Z_{10}	Z_{13}	Z_{19}
A00002	-0.4299	0.0295	6.8787	9.4890	0.9926	0.2763
A00028	2.3322	-0.2917	-0.1454	-0.1054	0.9926	-0.2326
A00041	-0.4299	1.3146	-0.1454	-0.1054	0.9926	0.2763
A00042	2.3322	-0.6130	-0.1454	-0.1054	0.9926	-1.2505
A00044	2.3322	1.4752	6.8787	9.4890	-1.0073	-0.7415
A00046	-0.4299	0.3508	-0.1454	-0.1054	-1.0073	0.7852
A00063	2.3322	-0.2917	-0.1454	-0.1054	-1.0073	0.7852
A00064	-0.4299	-0.4524	-0.1454	-0.1054	-1.0073	-0.7415

由表 1 可以看出处理后的数据均在-1 到 1 之间，将量纲统一后，有利于对数据的集中处理和使用，可以得到更准确的结果.

3) 筛选冗余指标

为减少 199 项指标中的冗余指标，判断 11017 组个案是否会逾期，首先将逾期个案与未逾期个案相互区别，分别求解均值，规定未逾期个案均值指标数为 $Z_1^0, Z_2^0, \cdots, Z_{198}^0$，逾期个案均值指标数为 $Z_1^1, Z_2^1, \cdots, Z_{198}^1$，令对应指标平均相关性为 $A_i\left(i=1,2,\cdots,198\right)$，故

$$A_i=\left|Z_i^0-Z_i^1\right| \quad \left(i=1,2,\cdots,198\right)$$

剔除 $A_i<0.1$ 的指标，此时留存 78 项指标，设为 $x_1, x_2, ..., x_{78}$.

5.2 模型的建立

5.2.1 主成分分析法

1) 因子分析

对留存的 78 项指标中使用 SPSS 进行因子分析，得到成分比较矩阵和 18 项主成分，此时成分 1 至成分 18 对方差的累计贡献率（详见附录 1）达到 82.586%，并且我们得到了各个属性之间的相关关系，碎石图见图 1.

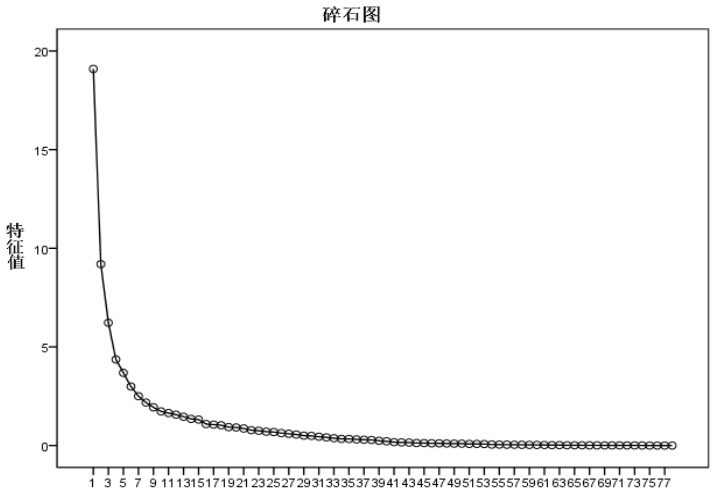


图 1 各成分碎石图

从图 1 中可以直观看出得到的 18 个主成分的特征值大于 1，其他成分的特征值均小于 1，由此我们可以用成分 1 至成分 18 来代替题目所给出的数据[2].

2) 权重确定

令 F 为综合评价得分，$F_1, F_2 \cdots F_{18}$ 分别为各数据中成分 1、成分 2 在评价体系中所得的分数.设 $C_1, C_1 \cdots C_{18}$ 分别为成分 1、成分 2 的初始特征值，$D_1, D_1, \cdots, D_{18}$ 表示成分 1 至成分 18 方差的百分比，D 为主成分对方差的累积贡献率，则

$$F_i=\sum_{i=1}^{n}\eta_i Z_i,\ i=1,2,\cdots,18.$$

$$F=\sum_{j=1}^{18}\frac{D_j}{D}F_j,\quad j=1,2,\cdots,18.$$

其中α_{st}为成分 1 至成分 18 中各项的比例系数，由于两个主成分中每个指标所对应的系数为成分矩阵中的数据除以主成分相对应的特征值开平方根，故

$$\alpha_{st}=\frac{A_{st}}{\sqrt{D_s}},\quad s=1,2,\cdots,18;t=1,2,\cdots,n.$$

此时，得到预测结果，此时误差分析见表 2.

表 2 主成分分析预测结果

逾期预测精度	未逾期预测精度	总预测精度
0.250	0.797	0.69

从表 2 中可知此时主成分分析模型逾期预测精度只有精度不高，为提高模型精度，采用神经网络对模型进行改进[3].

5.2.2 神经网络的建立

人工神经网络是具有适应性的简单单元组成的广泛并行互连的网络，是一种模仿动物神经网络行为特征，进行分布式并行信息处理的算法数学模型[4].人工神经网络是一个多输入单输出的非线性动态过程，其神经元模型见图 2.

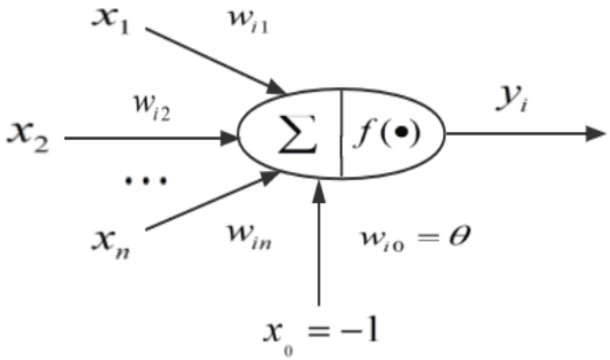


图 2 神经网络基本结构

图 2 中$x_1,x_2,\cdots,x_n$是从其他神经元传来的输入信号，w_{ij}表示从神经元 j 到神经元i的连接权值，θ表示一个阈值，令 net 为净激活量，每个神经元根据传来的输入信号，不断调整输入信号的权重贡献，判断是否达到固定激活阈值，当高于激活阈值后带入激活函数中激活输出，此时净激活量为

$$net_i=\sum_{j=1}^{n}w_{ij}x_j-\theta.$$

为将净激活量统一到量纲中，令激活函数为 f ，则

$$y_i=f\left(net_i\right).$$

1) *BP* 神经网络的建立

BP 神经网络实质是求取误差函数的最小值问题.这种算法采用非线性规划中的最速下降方法，按误差函数的负梯度方向修改权系数[5].其结构如图 3.

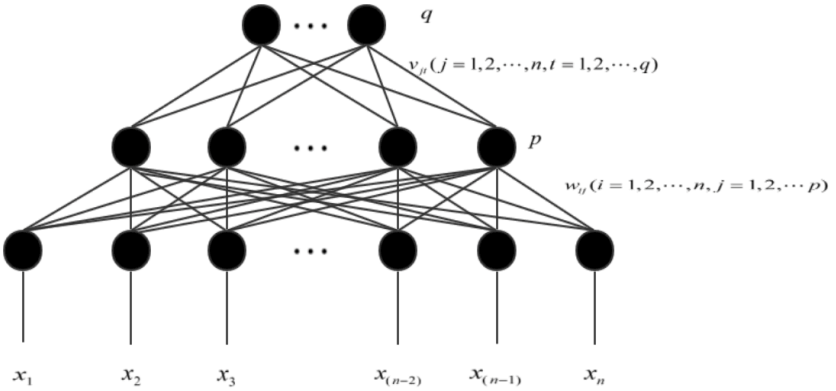


图 3 BP 神经网络结构图

设隐含层神经元个数为 p，输出层神经元个数为 q.

$$p = \sqrt{n + p} + \mu.$$

其中 μ 为 1 至 10 之间的常数.

设输入向量为 $X_n = \left(x_1, x_2, \cdots, x_n\right)$，隐含层输入向量为 $h_i = \left(h_{i_1}, h_{i_2}, \cdots, h_{i_p}\right)$，隐含层输出向量 $h_o = \left(h_{o_1}, h_{o_2}, \cdots, h_{o_p}\right)$，输出层输入向量为 $y_i = \left(y_{i_1}, y_{i_2}, \cdots, y_{i_p}\right)$，输出层输出向量为 $y_o = \left(y_{o_1}, y_{o_2}, \cdots, y_{o_q}\right)$，期望输出向量为 $d_o = \left(d_1, d_2, \cdots, d_q\right)$，输入层到隐含层的连接权值为 $w_{ij}\left(i = 1, 2, \cdots, n;\ j = 1, 2, \cdots, p\right)$，隐含层到输出层的连接权值为 $v_{jt}\left(j = 1, 2, \cdots, p;\ t = 1, 2, \cdots, q\right)$，中间层各个神经元的输出阈值 $\theta_j, j = 1, 2, \cdots, n$，输出层各个神经元的输出阈值（即偏置向量）$\gamma_t, t = 1, 2, \cdots, n.$

首先为统一量纲，对数据进行初始化，使数值在 $\left[a, b\right]$ 之间，则公式为

$$\hat{x} = \left(b - a\right)\frac{x - MAX\left(x\right)}{MAX\left(x\right) - MIN\left(x\right)} + a.$$

并对权值权值 w_{ij}, v_{jt} 阈值 θ_j 和 γ_t 赋予区间 $\left(-1,1\right)$ 内的随机数值.

对于隐含层输入值为 $h_{i_j} = \sum_{i=0}^{n} w_{ij}x_i - \theta_j,\ \left(i = 1, 2, \cdots, \mathrm{n};\ j = 1, 2, \cdots, \mathrm{p}\right).$

隐含层输出值为 $h_{o_j} = f\left(h_{i_j}\right),\ \left(j = 1, 2, \cdots, \mathrm{p}\right).$

输出层输入值为 $y_{i_t} = \sum_{i=0}^{n} v_{jt}h_{o_{ij}} - \gamma_t,\ \left(j = 1, 2, \cdots, \mathrm{p};\ t = 1, 2, \cdots, \mathrm{q}\right).$

输出层输出值为 $y_{o_t} = f\left(y_{i_t}\right),\ \left(t = 1, 2, \cdots, \mathrm{q}\right).$

计算期望向量与输出层输出值的差值$\hat{e}_t$，再利用权值v_{jt}与隐含层输出h_{o_j}计算一般化误差e_j，再来修正连接权值与阈值，得到公式

$$v_{jt}\left(N+1\right)=v_{jt}\left(N\right)+\alpha\hat{e}_t e_j.$$
$$\gamma_t\left(N+1\right)=v_{jt}\left(N\right)+\alpha\hat{e}_t.$$

其中$t=1,2,\cdots,\mathrm{q};\ j=1,2,\cdots,\mathrm{p};\ 0<\alpha<1$

$$w_{jt}\left(N+1\right)=w_{jt}\left(N\right)+\beta\hat{e}_t e_j.$$
$$\theta_j\left(N+1\right)=\theta_j\left(N\right)+\beta e_j.$$

其中$i=1,2,\cdots,\mathrm{n};\ j=1,2,\cdots,\mathrm{p};\ 0<\beta<1$

经过反复的学习，不断地反馈，直至全部训练样本训练完毕.

2) *BP* 神经网络的优化

经过不断优化测试，传递函数选择弹性梯度下降法对神经网络进行优化，学习函数表达式为$dX=\Delta X.*sign\left(gX\right)$.隐含层激发函数为

$$\phi=\frac{2}{1+e^{-2x}}-1$$

此时函数图像见图 4.

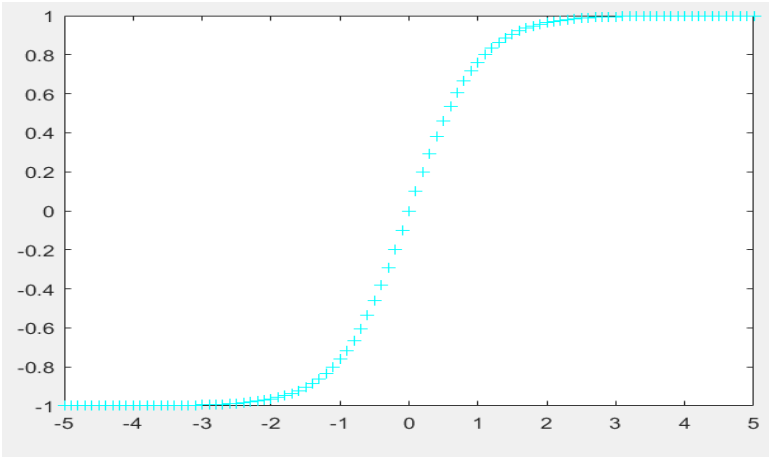


图 4 隐含层激发函数图

输出层激活函数为

$$\varphi=x$$

修改神经网络的学习速率，设为 0.3.利用附加动量法和弹性梯度下降法同时对神经网络的权值进行修正.

3) *BP* 神经网络的求解

利用 MATLAB 对模型进行求解（程序见附录 2），选择训练样本为 9537 组，测试样本为 1480 组，隐含层节点数为 10，输出层节点数为 1，此时模型误差为 0.426，训练次数为 2000 次，此时训练过程误差曲线为图 5.

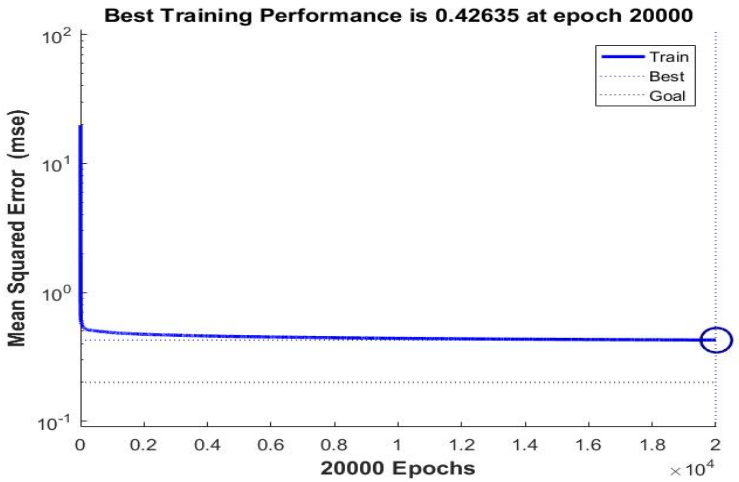


图 5 优化前训练过程误差曲线

此时预测精度见表 3.

表 3 优化前预测精度

逾期预测精度	未逾期预测精度	总预测精度
0.066	0.886	0.833

此时总预测精度较高，但逾期预测精度和未逾期预测精度较低，但考虑到银行信贷机构对于逾期精度较为重视，为提高逾期预测精度，对学习样本进行改进. 在改进选择样本后，采用训练组数据 1552 组，测试组数据 9465 组，训练次数为 2000，模型误差为 0.54354，训练过程误差曲线见图 6.

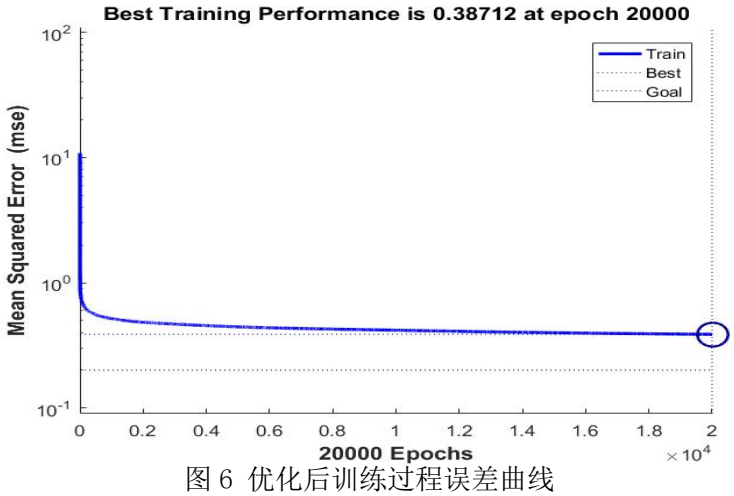


图 6 优化后训练过程误差曲线

此时预测精度见表 4.

表 4 优化后预测精度

逾期预测精度	未逾期预测精度	总预测精度
0.867	0.737	0.763

此时模型优化后逾期预测精度比优化前提高了 80%

对比优化前模型和优化后模型的梯度与学习次数，见图 7，图 8，

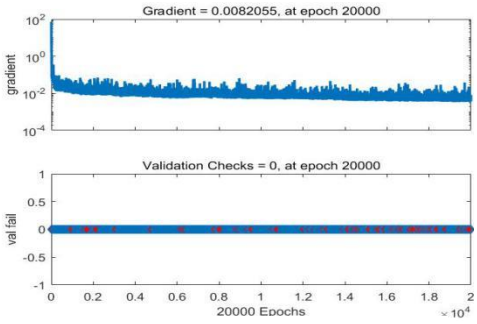


图 7 优化前模型梯度与学习次数

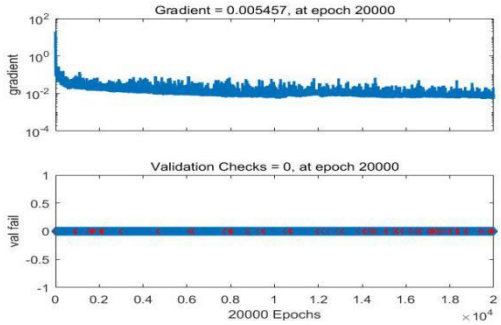


图 8 优化后模型梯度与学习次数

从图 7 和图 8 可以看出两次求解的数据梯度基本相同.对比优化前模型与优化后模型的残差正态检验图，见图 9，图 10.

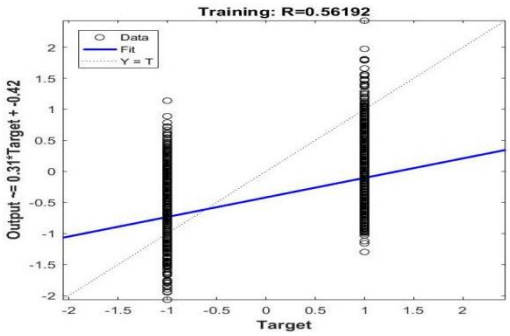


图 9 优化前模型残差正态图

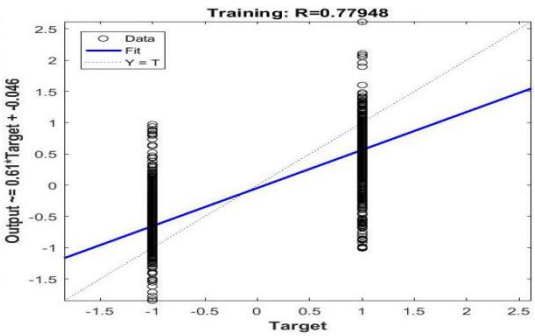


图 10 优化后模型残差正态图

对比图 9，图 10 可以看出优化前的拟合接近程度为 0.56192 此时拟合程度较差，优化后拟合接近程度为 0.77948，比优化前拟合程度提高了 0.21756，拟合较为准确，说明对于数据的处理以及测试样本选取的合理.

6. 模型优缺点分析

6.1 模型优点分析

①本文利用主成分分析法，有效减少冗余指标的个数并提取出对评价结果最有效的信息，降低模型维度，提高评价效率，同时解决了指标之间可能存在的多重共线性的问题.

②本文运用 *BP* 神经网络具有容错性、自组织与自适应能力的特点，并且其具有联想的功能，通过模拟人脑信息处理机制建立起来的网络系统，具有数学计算的能力,具有实现任何非线性映射的功能.

③ *BP* 神经网络具有处理知识的思维、学习以及记忆的功能，能通过学习带正确答案的实例集自动提取“合理的”求解规则.

6.2 模型缺点分析

①在变量降维的过程中，主成分的解释含义带有模糊性，提取的主成分个数小于原始变量个数.

② *BP* 神经网络的收敛速度较慢，可以对模型为了得到期望的输出结果，算法的需要经过长达数天的学习训练才能达到，大大地降低了工作效率；弹性梯度下降求得误差函数的极小值可能不是网络最小值，影响网络的性能.

7. 参考文献

[1]肖文兵，费奇.基于支持向量机的个人信用评估模型及最优参数选择研究.系统工程理论与实践，2006（10）：73-79.

[2]刘扬，刘江伟.特征选择方法在信用评估指标选择中的应用.数理统计与管理，2006，25（6）：667-672.

[3]方伟. 基于主成分分析和 BP 的商业银行个人信贷风险评价研究[D].华北电力大学,2012：18-20.

[4]张澜觉. 基于 BP 神经网络的 P2P 信贷个人信用评价模型研究[D].云南财经大学,2015：19-21.

[5]焦斌,叶明星.BP 神经网络隐层单元数确定方法[J].上海电机学院学报.2013(03)：114-115.

8. 附录

1) 附录 1：解释的总方差

解释的总方差

	初始特征值			提取平方和载入			旋转平方和载入		
成份	合计	方差的 %	累积 %	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	19.091	24.476	24.476	19.091	24.476	24.476	9.850	12.628	12.628
2	9.196	11.790	36.266	9.196	11.790	36.266	8.785	11.263	23.891
3	6.225	7.981	44.247	6.225	7.981	44.247	6.236	7.995	31.886
4	4.363	5.594	49.841	4.363	5.594	49.841	5.415	6.942	38.828
5	3.685	4.724	54.565	3.685	4.724	54.565	4.427	5.676	44.504
6	2.992	3.837	58.402	2.992	3.837	58.402	3.955	5.070	49.574
7	2.501	3.207	61.608	2.501	3.207	61.608	3.871	4.963	54.537
8	2.179	2.794	64.402	2.179	2.794	64.402	3.470	4.448	58.985
9	1.939	2.486	66.888	1.939	2.486	66.888	3.072	3.938	62.923
10	1.728	2.216	69.104	1.728	2.216	69.104	3.033	3.889	66.811
11	1.648	2.113	71.216	1.648	2.113	71.216	1.875	2.404	69.215
12	1.564	2.005	73.222	1.564	2.005	73.222	1.815	2.327	71.543
13	1.462	1.875	75.096	1.462	1.875	75.096	1.623	2.081	73.624
14	1.352	1.733	76.829	1.352	1.733	76.829	1.576	2.020	75.644
15	1.314	1.685	78.514	1.314	1.685	78.514	1.570	2.013	77.657
16	1.088	1.394	79.909	1.088	1.394	79.909	1.483	1.901	79.558
17	1.060	1.359	81.268	1.060	1.359	81.268	1.199	1.537	81.095
18	1.028	1.318	82.586	1.028	1.318	82.586	1.163	1.491	82.586
19	.934	1.197	83.784						
20	.918	1.177	84.960						
21	.863	1.107	86.067						
22	.780	1.000	87.067						
23	.750	.962	88.029						
24	.704	.903	88.932						
25	.684	.878	89.810						
26	.630	.808	90.618						

2) 附录 2：神经网络程序

```
P=xlsread('055.xlsx');
T=xlsread('033.xlsx');
[Q,ps]=mapminmax(P);
[R,pt]=mapminmax(T);
net=newff(minmax(P),[10,1],{'tansig','purelin'},'trainrp');
```

```
net.trainParam.epochs = 20000;
net.trainParam.goal=0.2;
net.trainParam.lr =0.3;
[net,tr]=train(net,Q,R);
P2=xlsread('066.xlsx');
Q2 = mapminmax('apply',P2,ps);
M=sim(net,Q2);
MappedData = mapminmax(M, 0, 1);
xlswrite('answer.xlsx',MappedData);
```

```
#include <fstream>
#include <iostream>
using namespace std;

int main()
{
    int m=1480,n=2960,q=0,w=0;
    double a[m],b[m];
    double k=0,x=0,y=0;
    ifstream in("1.txt");
    //ifstream in("2.txt");
    for(int i=0;i<m;i++)
    {
        in>>a[i];
        if(a[i]==0)
        q++;
        if(a[i]==1)
        w++;
    }
```

```
    for(int i=m;i<n;i++)
    {
        in>>b[i-m];
        if(b[i-m]<0.5)
            b[i-m]=0;
        else
            b[i-m]=1;
    }
    for(int i=0;i<m;i++)
    {
        if(b[i]==a[i])
        k++;
        if(b[i]==0&&a[i]==0)
        x++;
        if(b[i]==1&&a[i]==1)
        y++;
    }
    ofstream out("1.txt");
    //ofstream out("2.txt");
    cout<<x<<" "<<q<<endl;
    cout<<y<<" "<<w<<endl;
    cout<<k/m<<endl;
    cout<<x/q<<endl;
    cout<<y/w<<endl;
    cout<<"ok";
    in.close();
    out.close();
    cin.get();
}
```