近视预测与预警

摘要

近年来，我国学生的近视问题日益严重且低龄化趋势明显，已成为重大社会公共卫生问题。本文基于统计学原理，建立视力预警的机理模型和机器学习模型，达到对视力状况进行经济高效的诊断以及进行护眼预警的目的。

问题一采用 Logistic 回归对视力的影响因素进行分析。首先我们利用网上问卷调查，采取随机抽样调查的方法抽取了二百多名学生，对可能影响视力的 6 个因素进行问卷调查。收集的数据通过 SPSS 软件使用 Logistic 回归模型对影响视力因素进行分析，利用观察法和统计法将可能影响模型性能的异常样本剔除，并对模型做出诊断修正。最后得出遗传因素、户外锻炼时间、用眼负荷强度、读写习惯是影响学生视力的四大因素。

问题二用多变量灰色模型进行预测。根据问卷调查得到的各个年龄层的视力情况，使用 matlab 程序实现算法得到时间序列，由此预测得出在四个因素的影响下不同年龄阶段的视力状况，从而建立视力的演化机理模型。

对于问题三，为建立视力预警模型，在上一问的时间序列基础上先使用线型内核向量机分类，发现预测效果不佳后改用 RBF 的多分类向量机，根据测试者提供的关于户外锻炼时间、用眼负荷、用眼习惯、父母遗传和年龄这五个因素的情况，针对不同年龄层，将视力预警的层次分为轻度近视或不近视、中度近视、重度近视，以此为依据进行预警。同时结合 C 程序设计给出针对户外锻炼时间、用眼负荷、用眼习惯这三个方面进行多角度的预警，得到预测的准确率高达 90%，面向家长和老师简单易行。

对于问题四，采用广义回归神经网络（GRNN）和概率神经网络（PNN）分类判断。首先利用前面回归保留的因素数据产生训练集和测试集。然后利用 MATLAB 自带的神经网络工具箱创建 GRNN 与 PNN。最后通过计算测试集预测类别与真实类别间的误差，同时通过对比不同因素及因素组合与近视程度的相关性，在模型精度与运算速度上作出折中选择，得到 GRNN 的判断准确率高达 93.3%，验证了模型的准确性与科学性。

对于问题五，我们准备必要的训练样本及代码以供模型的实现。通过前面的问题我们已遴选出主要因素，因此仅需对象提供相关的自身的数据导入模型代码即可实现模型。

本文利用信息时代的优势，进行网络问卷调查，收集整理大量文献资料的数据，将其与两次问卷调查的数据进行对比分析，使数据的可靠性更高，以图表形式展示直观明了。第三问建立模型由简到繁，相互印证，第四问使用两种模型进行对比得到准确率最高的模型。

关键词：Logistic 回归分析　多变量灰色预测模型　支持向量机　有导师学习神经网络的分类

一、 问题重述

近年来，我国儿童青少年的近视问题日益严重且低龄趋势明显，已成为重大社会公共卫生问题。除了先天因素外，近视高发主要是日常用眼不良习惯导致的。比如户外活动量不足、电子产品的过早大量使用，造成眼睛的近距离负荷过重。工作、阅读、打游戏所有近距离的眼负荷都是近视眼发生的主要原因。我们希望通过教育管理系统及其相关软件，硬件收集不同的数据，并借此通过数学模型分析预测学生的近视发生概率，从而及时预警，使学校或家长可以尽早进行干预，有效预防或矫正近视。

本文将建立数学模型解决下列问题：

1.分析影响视力的关键因素并建立其量化模型。给出需要获得的对应数据以及可以获取的可能途径。

2. 基于问题 1，给出眼睛视力的演化机理模型（非机器学习类模型）。

3. 基于问题 2，给出眼睛视力的预警机制模型。以供青少年父母作参考。

4. 在上述基础上设计可供学习的信息系统（数据表）， 通过查阅资料和数据以及仿真等手段进行眼睛视力的机器学习模型。5. 写一份方案来实现文中的机理模型和学习模型。

二、 问题分析

2.1 问题一的分析

问题一要求分析影响视力的因素并将其量化，我们可以采用问卷调查的方式收集数据，但是导致视力低下的因素较多，在这之前我们需要确定大致的考虑因素，因此我们可以先从深度与广度同时入手，查阅大量相关文献资料，再通过 Logistic 回归分析找出影响视力的关键因素。结合信息化时代信息传播快速方便的趋势，我们可以采用网上问卷调查的方式进行数据收集。

2.2 问题二的分析

问题二要求建立视力的演化机理模型，我们可以找到各个因素在不同年龄层对视力的影响，由于灰色预测模型能将无规律的数据整合成有规律的时间序列，因此我们可以建立此模型来进行视力演化的预测和分析。

2.3 问题三的分析

问题三要求针对视力作出多对象的预警，可以使用向量机进行分类判别，根据上一问不同年龄对视力的影响程度不同，我们可将年龄作为其中一项参数，各个因素作为另外的参数，进行机器学习，判别近视程度或将来趋势属于轻度或不近视、中度近视、重度近视的哪一类。为了让预警的角度更加全面多样，我们可以使用 C 程序设计粗略判断哪些因素可能导致被测者视力下降，并及时作出预警。

2.4 问题四的分析

不同因素对近视程度的影响呈现不同的线性或非线性关系，且其组合后对近视程度的影响更加复杂，传统方法较难分类判断。可基于提供的正确的输入/输出对（即训练样本），利用 GRNN 与 PNN 神经网络对输出与期望值进行比较，然后应用学习规则调整权值和阀值使输出接近期望值。

2.5 问题五的分析

为了实现演化机理模型和机器学习模型，我们需要收集足够的样本数据且保证其有效性，可以利用信息时代的方便快捷，将问卷内容细化，使用网上问卷进行信息采集，同时利用回归拟合方法合理地处理数据。

三、 模型假设

（1)假设所有数据真实可信且抽样样本能完全反映总体的特征。

（2）假设近视产生的过程没有意外事件发生和白内障等特殊情况导致的近视。

（3）假设用眼不卫生仅包括：躺着或行走时阅读、用眼时间隔休息少、阅读姿势不端正、看东西距离太近、无眼保健操、光线不好这六种可能。

（4）假设近视按病情水平分为 3 种情况。

（5）假设近视结果不可逆。

四、 符号说明

符号	说明
$LW_{2,1}$	第二层权值矩阵
$\|\|dist\|\|$	欧式距离函数
$IW_{1,1}$	第一层权值矩阵
b_1	隐含层的阀值
n^i	输出层神经元输出
a^i	隐含层神经元输出
P	输入矩阵
T	输出矩阵
Q	训练集样本数
C	径向基函数中心
x_0	原始时间序列
x_1	累加生成向量序列
A	发展系数

B	灰作用量
l	紧邻均值生成序列
L	Y 构造出的数据矩阵

五、 模型的建立与求解

5.1 问题一

5.1.1 资料收集与整理

影响青少年视力低下的因素是多方面的，主要包括先天遗传因素和环境因素，大量文献资料显示，环境因素包括：用眼负荷、用眼习惯、采光照明、体育锻炼、体质状况这五个方面，以下内容简要概括了这几种因素对视力的影响。

遗传因素：

遗传因素极大地影响着青少年的视力，高度近视眼基因突频率较高。王志强等的文献[1]、对一级亲属近视情况的家系调查[2]都表明：遗传作用对视力存在明显影响。

环境因素：

（1）不良用眼习惯。研究[3]表明，阅读姿势不良、眼距过近、躺着阅读等不良用眼习惯越多，近视概率越大，表 1 展示了视力不良率随着不良用眼习惯个数的增多的变化情况：

表 1 ：不良用眼习惯对视力的影响

不良用眼卫生习惯个数	视力不良人数	总人数	视力不良率
0	0	2	0
1	2	19	0.1053
2	7	41	0.1707
3	38	122	0.3115
4	44	111	0.3064
5	72	141	0.5106
6	68	44	0.7234
7	67	82	0.8171
8	43	49	0.8776
9	14	15	0.9333
10	3	3	1

（2）用眼负荷大。曾秋红等[4]报告青少年的视力低下率随年龄增长（学级上升）而增高，而随着年龄增长，近距离用眼负荷也在增大，因此我们将小学三年级以下、小学三年级以上、初中、高中的用眼强度大致按照小、中、略大、大的层次划分，得到的用眼负荷与视力低下的关系如表 2 所示：

表 2 ：用眼负荷与视力低下率的关系

用眼负荷	视力低下眼数	低下率
小	149	24.19
中	216	35.07
略大	301	48.86
大	404	65.68

（3）体育锻炼。我国青少年课业繁重，严重缺乏体育锻炼时间，而研究[5]表明，适当的体育锻炼、认真做好两操，能改善眼周血液循环，解除眼睫状肌痉挛，对视力具有极好的保护作用，基于文献[6]的研究报告，本文整理出不同体育运动时间的视力不良率，结果如表 3：

表 3 ：体育活动时间对视力的影响

体育活动时间	不良人数	总人数	视力不良率
0.5 小时以下	49	58	84.48
0.5 到 1 小时	102	144	70.83
1 到 1.5 小时	67	96	69.79
1.5 以上	43	100	43

（4）采光照明。采光照明条件不良，就会导致学生眼物距过近，造成调节过度紧张，引发视力低下。由于教室照明状况和家庭大致类似且学生大部分时间都在教室学习，因此本文只考虑教室照明状况对近视的影响。文献[7]统计了在不同采光条件下的教室里学生的视力低下率，将合格与不合格教室的近视比例进行对比，得到表 4 的结论：

表 4 ：采光照明对近视的影响

采光照明	调查人数	患病人数	患病率（%）
合格	855	36	4.21
不合格	2387	295	13.17
合计	3242	331	10.21

（5）营养状况。与眼睛关系密切的营养成分的缺乏或过剩，都会直接影响视力。本文只讨论营养状况（良好或不良）对视力的影响。文献[8]表明，营养良好的近视不良率低于营养不良的，结论如表 5：

表 5 ：营养状况对近视的影响

营养状况	视力不良人数	总人数	视力不良率
营养良好	406	558	72.76
营养不良	141	187	75.4
超重	57	82	69.51
肥胖	106	139	76.26

参考上述文献资料，我们认为在之后的问题探讨中，主要应从遗传和环境因素两个大类考虑影响视力的因素，环境因素中，用眼负荷、用眼习惯、采光照明、体育锻炼、体质状况这五个因素对视力影响较大，因此我们在之后的数据收集和整理中，主要考察这几个因素的影响。

5.1.2 Logistic多元回归分析

我们在上述文献的基础上进行了问卷调查，调查了是否近视、用眼负荷、户外锻炼、读写姿势是否端正、光线是否良好、父母近视与否这七个问题，并以后六者为自变量，是否近视为因变量构建 Logistic 回归模型进行分析。在本模型中，自变量与因变量都采用虚拟变量（0、1、2）表示，如表 6 所示：

表 6 ：Logistic回归模型中的变量及赋值

变量代码	变量意义	赋值
Y	近视与否	0：不近视；1：近视
X_1	光线	0：昏暗；1：良好
X_2	读写姿势端正	0：不端正；1：端正
X_3	户外锻炼时间	0：不到一小时；1：一小时以上
X_4	营养状况	0：BIM<18；1：BIM18~25；2:BIM>25
X_5	用眼负荷	0：<8h/d；1：>8h/d
X_6	父母近视与否	0：都没近视；1：一个；2：两个

最终得到此公式：

$$\omega x+T=-0.769-0.746x_1+0.807\times 3x_2+1.187x_3+0.178x_4-1.141x_5-0.936x_6$$

$$p=\frac{e^{\omega x+T}}{1+e^{\omega x+T}}$$

并得到结果如下表所示：

表 7 ： 多分类Logistic回归分析结果汇总

因素	回归系数	标准误	Z 值	P 值	OR 值	OR 值 95%CI(L)
光线	-0.746	0.593	-1.259	0.208	0.474	0.148
读写姿势端正	0.807	0.362	2.227	0.026	2.241	1.102
锻炼时间	1.187	0.361	3.285	0.001	3.278	1.614
营养	0.178	0.303	0.586	0.558	1.195	0.659
用眼负荷	-1.141	0.367	-3.108	0.002	0.319	0.156
父母近视与否	-0.936	0.303	-3.09	0.002	0.392	0.217
截距	-0.769	0.762	-1.009	0.313	0.463	0.104

5.1.3 结论与分析

逐一分析表中各个因素的回归结果，我们可以得到以下结论：

（1）光线是否良好（X1）的回归系数值是-0.746，但是并没有呈现出显著性（z=-1.259,P=0.208>0.05），意味着光线并不会对是否近视产生极大影响。然而文献[7]中采光良好的教室学生近视率（4.21%）远远低于采光不良的教室（13.17%），经过对比我们发现，本文所做调查问卷针对群体都是教育条件较好的城市学生，如今城镇学校的教室采光普遍较为良好，所以问卷结果显示：光线良好的样本占比约89.2%，占比极大，无法体现光线优良状况是否近视情况产生影响。

(2)读写姿势（X2）的回归系数值是 0.8.7，并且呈现出 0.05 水平的显著性（z=2.227,P=0.026<0.05）,意味着读写姿势端正对视力产生正向影响，即姿势越端正，近视的概率越低。

(3)营养状况（X4）的回归系数值为 0.178，但是并没有呈现出显著性（z=0.586,P=0.558>0.05）,意味着营养并不会对近视与否产生过大影响。出现这样的现象，一方面是因为调差人群主要是城镇学生，营养条件较好，因此问卷调查中营养这部分数据无法良好地体现营养状况与近视的关系。另一方面，根据表数据，文献[8]中营养状况不同的个体近视情况差异也不显著，与调查问卷结果相近。

（4）用眼负荷（X5）的回归系数值为-1.141，并且呈现出 0.01 水平的显著性（z=-3.108,P=0.002<0.01），意味着用眼负荷对近视产生显著的负向影响，即用眼习惯不良好会导致近视的概率上升。

（5）父母是否近视（X6）的回归系数值为-0.936，并且呈现出 0.01 水平的显著性（z-3.090,P=0.002<0.01），意味着父母是否近视会对本人是否近视产生显著影响，与各类文献的结果基本相符。

基于上述分析，X1（光线）和 X4（营养状况）的 P 值分别为 0.208，0.558，没有通过检验。

为了建立更精确的模型，我们对上述模型进行修正。用后向消去法，剔除对模型影响不显著的变量 X1 和 X4，对剩下的自变量重新做回归分析，结果见下表。

表 8 ：剔除部分变量后的Logistic回归分析结果

影响因素	回归系数	标准误	Z 值	P 值	OR 值	OR 值 95%CI(L)
锻炼时间	-1.812	0.478	-3.794	0	0.168	0.064
用眼负荷	1.468	0.356	4.122	0	1.548	2.16
读写习惯	1.158	0.299	3.869	0	1.483	1.77
父母近视情况	-1.422	0.364	-3.903	0	0.34	0.118
截距	-3.636	2.228	-1.632	0.103	0.026	2.077

这时表 8 的回归系数的 P 值都小于 0.005，通过了检验，因而该模型具有显著的统计学意义。

得到各个因素对视力的影响程度如图所示：

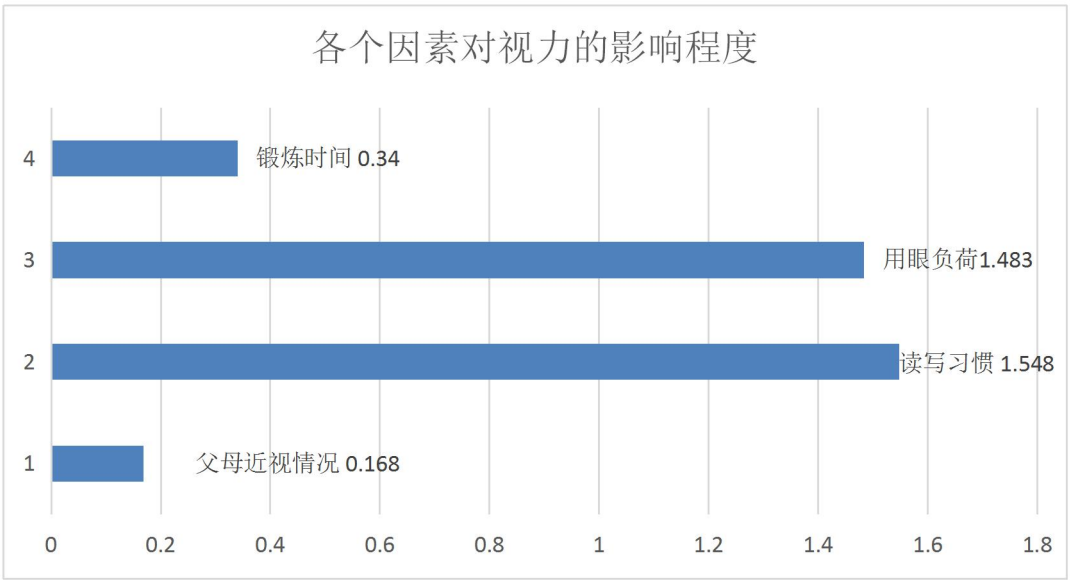

图 1 各项因素对近视的影响程度

5.1.4 数据的获得途径

考虑到如今信息化时代的趋势，我们可以使用大数据，收集更多的样本，以得到更加精确的回归函数，获得：近视与否、用眼负荷大小、用眼习惯是否良好、采光照明是否良好、体育锻炼时长、体质状况是否良好及父母是否近视这七项指标数据。

在信息化时代，以网上问卷调查的形式比传统的纸质问卷调查更加便捷、效率更高，问卷提问内容即为上述七项指标。

5.2 问题二

基于第一问的数据和分析，本文得到了遗传、年龄、用眼负荷、户外运动时间、

读写习惯这五个因素对视力有关键影响。由于有些样本先天近视（即先天就演化成了近视眼），故在此模型的样本数据中将遗传因素及其相关数据剔除。

由于该问题包含多个变量，各变量相互间也可能有相互作用，为了研究视力的随着这些因素的演化机理，我们建立多变量灰色预测模型来探究此问题。多变量灰色模型可同时综合考虑多个指标，从系统的角度对各特征参数进行统一描述，预测精度高。

5.2.1 多变量灰色模型

假定时间序列序列为 $X^{(0)}=\left\{X^{(0)}(1),X^{(0)}(2),\cdots X^{(0)}(n)\right\}$，其一次累加生成向量序列为 $X^{(1)}=\left\{X^{(1)}(1),X^{(1)}(2),\cdots X^{(1)}(n)\right\}$，其中

$$X^{(1)}(k)=\sum_{j=1}^{k}X^{0}(i),(k=1,2,\cdots,n)$$

n 为观测数据的个数，$X^{(0)}(k)=\left\{X_1^{(0)}(k),X_2^{(0)}(k),\cdots X_m^{(0)}(k)^T\right\}$ 是 m 维列向量，如果记

$$A=\begin{pmatrix}a_{11} & \cdots & a_{1m}\\ \vdots & \ddots & \vdots\\ a_{m1} & \cdots & a_{mm}\end{pmatrix}\quad B=(b_1,b_2,\cdots,b_n)^T$$

则多变量灰色模型的动态微分方程组可表示为

$$\frac{dX^{(1)}(t)}{dt}=AX^{(1)}(t)+B$$

相应的连续时间响应函数为

$$X^{(1)}(t)=e^{At}X^{(1)}(1)+A^{-1}\left(e^{At}-I\right)B$$

为了得到模型参数的估计值，需要将上述微分方程组转化为离散形式，从而得到模型参数的估计值。利用最小二乘法可以得到 $D=(A,B)^T$ 的估计值为

$$\hat{D}=\left(\hat{A},\hat{B}\right)^T=\left(L^TL\right)^{-1}L^TY$$

其中

$$L=\begin{pmatrix}\left(X_1^{(1)}(1)+X_1^{(1)}(2)\right)/2 & \cdots & \left(X_m^{(1)}(1)+X_m^{(1)}(2)\right)/2\\ \vdots & \ddots & \vdots\\ \left(X_1^{(1)}(n-1),X_1^{(1)}(n)\right)/2 & \cdots & \left(X_m^{(1)}(n-1),X_1^{(1)}(n)\right)/2\end{pmatrix},$$

$$Y=\begin{pmatrix} X_1^{(0)}(2) & \cdots & X_m^{(0)}(2) \\ \vdots & \ddots & \vdots \\ X_1^{(0)}(n) & \cdots & X_m^{(0)}(n) \end{pmatrix}$$

根据上式可得到参数 A 和 B 的辨识值 $\hat{A}$ 和 $\hat{B}$.有了参数估计就可以得到时间响应函数为

$$\hat{X}^{(1)}(k)=e^{\hat{A}(k-1)}X^{(1)}(1)+\hat{A}^{-1}\left(e^{\hat{A}(k-1)}-I\right)\hat{B} \quad \hat{X}^{(1)}(1)=\hat{X}^{(0)}(1)$$

利用（3）式还原成 原来的时间序列有

$$\hat{X}^{(0)}(k)=\hat{X}^{(1)}(k)-\hat{X}^{(1)}(k-1),k=2,3,\cdots$$

5.2.2 模型检验

对收集到的数据利用 Matlab 程序可得参数 A 和参数 B 的辨识值 $\hat{A}$，$\hat{B}$（见附录），从而有 $\hat{D}=\left(\hat{A},B\right)^T$，可计算视力水平在这些因素影响下的拟合值并预测演变情况。

下面考虑模型的检验，分别计算均方差比值和小误差概率，可得 $s<0.35$，$p=1$，根据模型等级标准，可知该模型预测和拟合精度为一级，从而可以用来预测。

5.3 问题三

5.3.1 数据的收集

由于第一次问卷调查的问题选项设置不够细致，我们又做了第二次问卷调查，这次问卷将近视程度细化为:200 度以下（轻度近视或不近视）、200-500 度（中度近视）、500 度以上（重度近视）；将户外锻炼时间细化为 0-0.5h、0.5-1h、1-1.5h、1.5-2h、2h 以上；将用眼不卫生的情况细化为：躺着或行走时阅读、用眼时间隔休息少、阅读姿势不端正、看东西距离太近、无眼保健操、光线不好这六种情况，并统计个人具有不良习惯的个数。之后的问题将结合两次问卷调查的数据进行分析。

5.3.2 支持向量机的分类模型

1.模型的建立与求解

由第二问我们可以看出，在数据较为缺乏，且影响因素繁多的情况下，非机器学习无法很完备地预测、预警近视，因此我们先使用支持向量机进行分类。

根据前几问的分析，我们可以看出影响视力的因素主要为：户外锻炼时间、用眼负荷、用眼习惯、父母遗传和年龄，因此我们将这五者作为五项指标，预测在前四项指标下，不同年龄的青少年的近视情况。

用 $i=1,2,\cdots,150$ 分别表示 150 份有效的问卷调查样本，第 i 个变量的第 j 个指标的取值为 a_{ij}。y_i 表示第一类（近视度数为 200 度以下），y_j 表示第二类（近视度数

200 度以上)。将所有数据进行标准化处理得到标准化指标变量。记 $\tilde{x}=[\tilde{x}_1,\cdots,\tilde{x}_5]^T$ 。

记标准化后的 150 个已分类样本点数据行向量为 $b_i=[\tilde{a}_{i1},\cdots,\tilde{a}_{i5}],i=1,\cdots,150$ 。利用线性内核函数的支持向量机模型进行分类，求得 60 个支持向量，并得到线性分类函数为

$$c(\tilde{x})=\sum_i \beta_i K(b_i,\tilde{x}) + b$$

式中：$\beta_i=\alpha_i y_i$，$\tilde{x}=[\tilde{x}_1,\cdots,\tilde{x}_5]$，$K(b_i,\tilde{x})=(b_i \bullet \tilde{x})$。

2.结果分析与模型检验

当 $c(\tilde{x})\geq 0$，$\tilde{x}$ 属于第一类（近视度数在 200 度以下）；当 $c(\tilde{x})<0$ 时，$\tilde{x}$ 属于第二类（近视度数在 200 度以上)。用判别函数进行判别，得到需要预警的样本属于哪一类别。

将所有已知样本点回代分类函数，得到误判率为 16.67%，正确率较低，说明线性内核函数的支持向量机模型不能很好地将近视情况进行分类。

而且线型内核函数的支持向量机只能粗略地将分为两种类别，不够精确。

5.3.3 RBF 核函数的多分类向量机

由于前一种向量机不够精确，本文再使用 RBF 核函数的多分类向量机建立视力预警模型，大体上可以分为以下几个步骤：

图 2 SVM 模型的建立步骤

我们使用 matlab 程序对模型进行实现，源代码见附录，得到更加优化的分类，分成不近视或低度近视（200 度以下），中度近视（200~500 度），重度近视（500 度以上），并用其中 30 个样本来检验，得到 90%的正确率，即 30 个测试样本中，只有 3 个样本预测出现了偏差，正确率较高，预测效果较好。

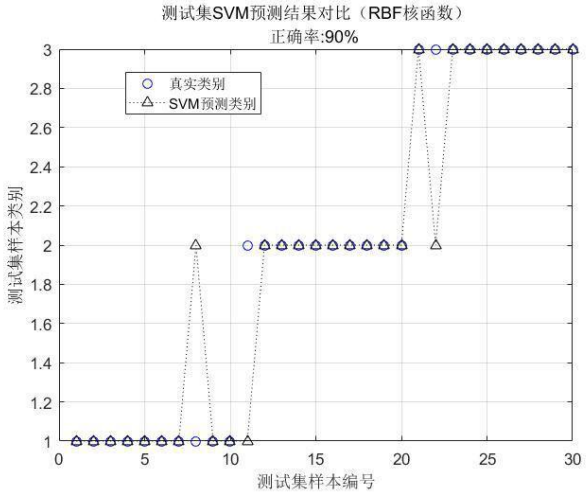

图 3 测试集 SVM 预测结果对比（RBF 核函数）

5.3.3 使用 C 语言进行多角度的预警

根据收集到的数据得到对应年龄的四个指标的平均值，进行测试时，输入四个问题对应的答案：

1. 你平时户外锻炼的时间大约多久？选项：0-0.5h、0.5-1h、1-1.5h、1.5-2h、2h。

2. 你一天的用眼负荷是多少（视近物与用眼持续时间长）？选项：0～2h、2～4h、4～6h、6～8h、8h 以上。

3. 你的父母近视吗？选项：都不近视、其中一个近视、双方都近视。

4. 你的读写习惯符合下列几条（躺着或行走时阅读、用眼时间隔休息少、阅读姿势不端正、看东西距离太近、无眼保健操、光线不好）？选项：0、1、2、3、4、5、6。

若输入值低于平均值，则说明该因素出现了问题，提出预警，提醒青少年尽快改正，以防视力进一步恶化。再结合向量机对近视程度的预警，若预警近视，则提醒青少年应该前往正规机构检测眼睛状况。

5.4 问题四

5.4.1 模型的建立

前文统计出了学生视力的四个影响因素，为了利用这四个因素进行视力的判别，我们利用广义回归神经网络（GRNN）和概率神经网络（PNN）分别建立：

（1）近视程度的判别模型，并对模型的性能进行评估。

（2）各个因素及因素组合与近视程度间的判别模型，并与（1）中所建模型的性能及运算时间进行对比，从而探求各个因素及因素组合与近视程度的相关程度。

GRNN 的结构如图 4，一般由输入层、隐含层和输出层组成。

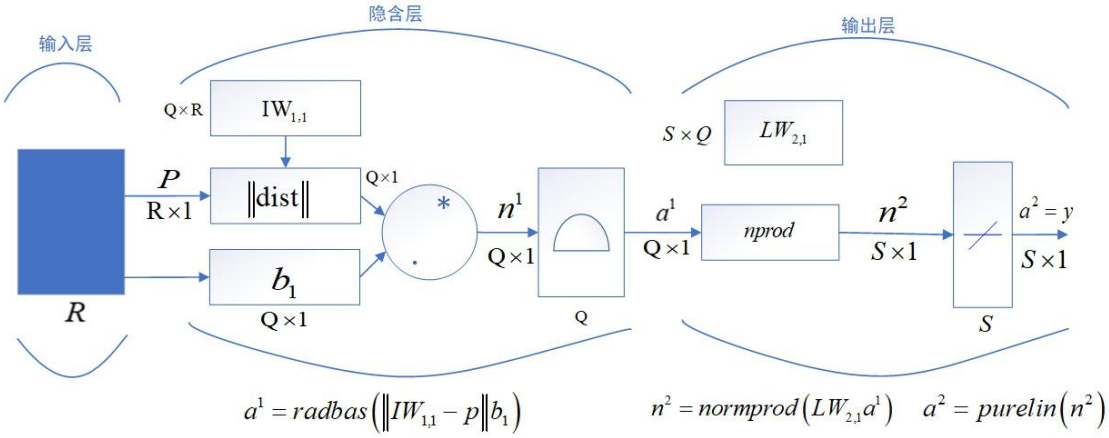

图 4 GRNN 的结构示意图

PNN 的结构如图 5，与 GRNN 类似，仅在输出层部分有细微差别。

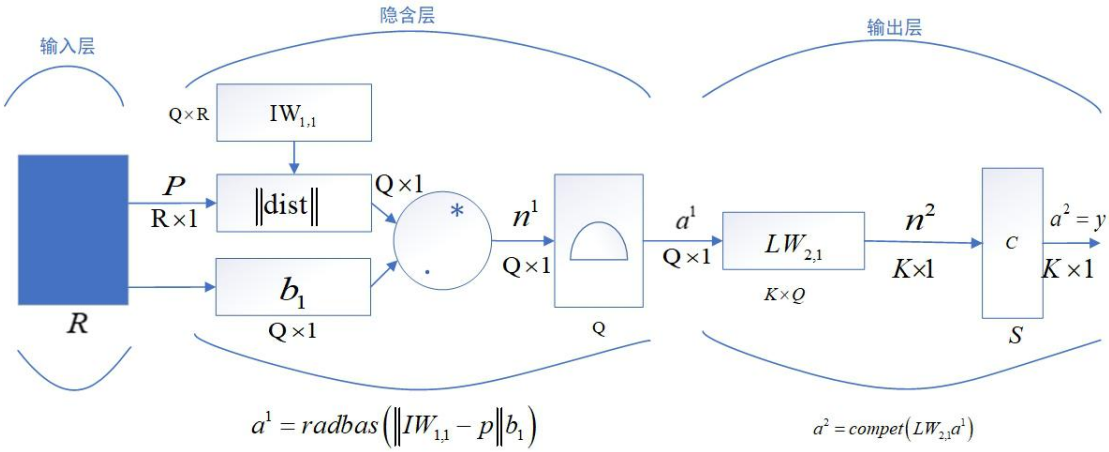

图 5 PNN 的结构示意图

（1）确定隐含层神经元径向基函数中心

为不失一般性，设训练集样本输入矩阵 P 和输出矩阵 T 分别为

$$P=\begin{pmatrix} p_{11} & \cdots & p_{1Q} \\ \vdots & \ddots & \vdots \\ p_{R1} & \cdots & p_{RQ} \end{pmatrix},\quad T=\begin{pmatrix} t_{11} & \cdots & t_{1Q} \\ \vdots & \ddots & \vdots \\ t_{S1} & \cdots & t_{SQ} \end{pmatrix}$$

其中，p_{ij} 表示第 j 个训练样本的第个输入变量；t_{ij} 表示第 j 个训练样本的第 i 个输入变量；R 为输入变量的维数；S 为输出变量的维数；Q 为训练集样本数。

隐含层的每个神经元对应一个训练样本，即 Q 个隐含层神经元对应的径向基函数中心为

$$C=P'$$

（2）确定隐含层神经元阀值

为了简便起见，Q 个隐含层神经元对应的阀值为

$$\mathrm{b}_1=\left[b_{11},b_{12},\cdots,b_{1Q}\right]'$$

其中，$b_{11}=b_{12}=\cdots=b_{1Q}=\dfrac{0.8326}{spread}$，spread 为径向基函数的扩展速度。

（3）确定隐含层与输出层间权值

当隐含层神经元的径向基函数中心及阀值确定后，隐含层神经元的输出便可以如下计算：

$$a^i=\exp(-\left\|C-p_i\right\|^2 b_1),\quad i=1,2,\cdots,Q$$

其中，$\mathrm{p}_i = \left[p_{i1}, p_{i2}, \cdots, p_{iR}\right]'$ 为第 i 个训练样本向量。并记 $a^i = \left[a_1^i, a_2^i, \cdots, a_Q^i\right]$。

GRNN 与 PNN 中隐含层与输出层间的连接权值 W 取为训练集输出矩阵，即

$$W = t$$

(4)输出层神经元输出计算

当隐含层与输出层神经元间的连接权值确定后，根据图 XX 所示，便可以计算出输出层神经元的输出，即

$$n^i = \frac{LW_{2,1}a^i}{\sum_{j=1}^{Q} a_j^i}, \quad i = 1, 2, \cdots, Q$$

$$y^i = purelin\left(n^i\right) = n^i, \quad i = 1, 2, \cdots, Q$$

5.4.2 模型的求解

对于上面建立的有导师学习的神经网络分类模型，我们利用 matlab 工具箱函数 newgrnn 与 newpnn 分别创建一个 GRNN 与 PNN，并结合使用 matlab 进行求解。在各个近视程度类别的 50 个样本中分别随机选取 40 个样本（三类共 120 个）构成训练集，剩余的 10 个样本（三类共 30 个）作为测试集。仿真训练结果如图 XX 所示：

图 6 神经网络分类模型的建立步骤

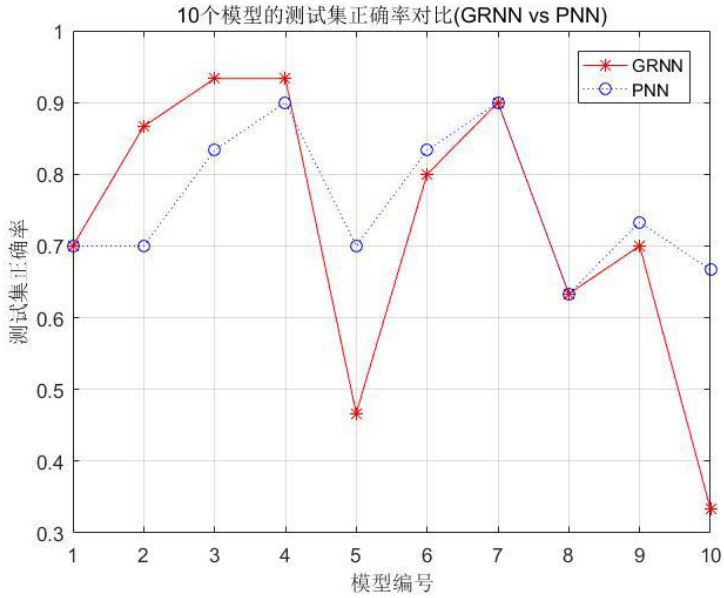

图 7　十个模型的测试集正确率对比（GRNN vs PNN）

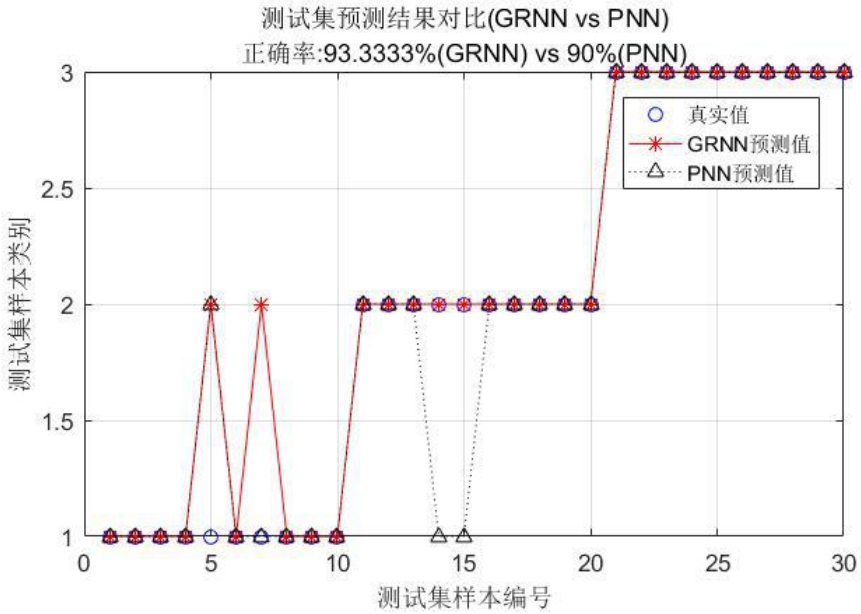

图 8 测试集预测结果对比（GRNN vs PNN）

5.4.3 模型分析

由于训练集和测试集是随机产生的，每次运行的结果也会有所不同。

从上面求解的图中不难发现：

1.GRNN 和 PNN 模型具有良好的泛化能力，预测集预测正确率分别达 93.3%和 90%。

2.如表 9 所列，利用四个因素（户外锻炼时间、用眼负荷、读写不良习惯、父母遗传）建立的模型编号分别为 1、5、8、10。由图 7 可以清晰地看出，利用四个因素单独建立的 GRNN 模型性能较好，正确率在 70%左右。与之对应的 PNN 模型结果呈现类似规律，但用眼负荷强度和父母遗传近视因素独立建立的 PNN 性能不佳。

表 9：10 个模型对应输入变量

模型编号 / 输入属性	1	2	3	4	5	6	7	8	9	10
户外锻炼时间	○	○	○	○						
用眼负荷强度		○	○	○	○	○	○			
读写习惯			○	○		○		○	○	
遗传问题				○			○		○	○

3.与传统的 BP 神经网络相比，GRNN 具有如下优点：

（1）网络的训练是单程训练而不需要迭代。

（2）隐含层神经元个数由训练样本自适应确定。

（3）网络个层之间的连接权重由训练样本唯一确定，避免了 BP 神经网络在迭代中的权值修改。

（4）隐含层节点的激活函数采用对输入信息具有局部激活特性的高斯函数，使得对接近于局部神经元特征的输入具有很强的吸引力。

（5）运算时间快，10 个模型的时间在 50ms 左右，如图 9 所示

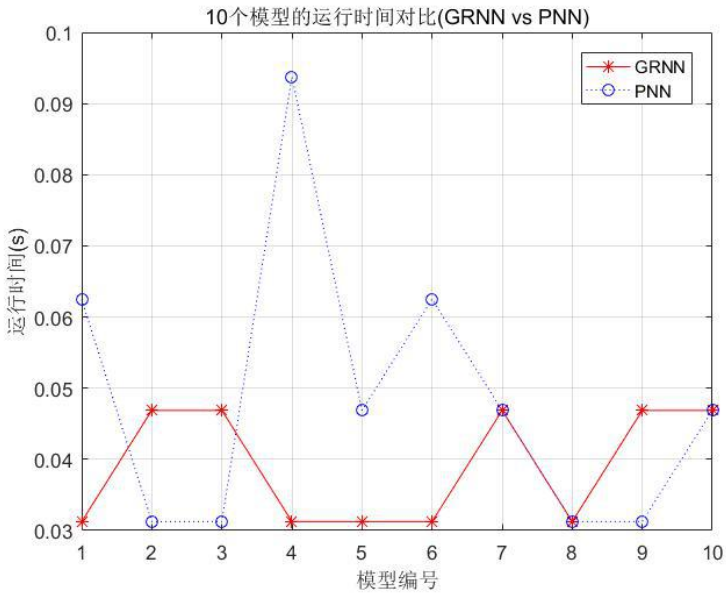

图 9 十个模型的运行时间对比（GRNN vs PNN）

5.5 问题五

5.5.1 模型的实现需满足的条件

1. 收集到足够多的数据，且保证数据的有效性与准确性

2. 用合适的方法处理数据，剔除关联性较小的因素，细化主要影响因素

5.5.2 具体实现

问题一中我们利用了 logistic 模型进行回归诊断。要想实现该模型，首先需要掌握 SPSS 进行 logistic 回归的方法，然后要根据认为的对视力有关键影响的因素收集数据，此时收集的数据在保证真实可信的情况下可尽量充实以保证模型的准确性。经分析，剔除了对模型影响不显著的变量即照明条件与身体状况，保留了其他有效影响因素。

问题二中我们使用了多变量灰色预测模型来模拟视力的演化机理，要想实现该模型，主要需要掌握代码的意义和收集到真实准确的数据。代码较简短且可移植性强，容易掌握。数据方面，其一，由于灰色模型的特点，只需收集一定量的数据并保证其真实可信即可。其二，在第一问基础上排除了一些无关因素后，需收集学龄、用眼负荷、户外运动时间、读写习惯、视力，并按时间取平均值。将其结果作为原始时间序列矩阵输入即可得到结果。

问题四中我们使用了有导师学习的神经网络分类判断模型来设计可供学习的信息系统。因此仅需收集对象的户外锻炼时间、用眼负荷强度、读写习惯与父母遗传近视这四种因素的数据。先利用已有的正确的输入/输出对（即训练样本）对模型进行仿真训练，然后在模型精度上作出选择，从 GRNN 与 PNN 中选取的判断准确率高的一者，然后对其输入对象的因素数据即可进行判断近视程度。

六、模型的评价

6.1 模型的优点

6.1.1 Logistic模型

Logistic模型是一种广义的回归分析模型，对变量的要求宽松，且不要求变量连续或服从正态分布，适用范围广。其分析结果中，P值可检验相关因素是否对近视与否产生影响，OR值可以直观地看出因素对近视的影响程度。

此外，本文收集了大量文献资料，将整理出的结果与使用Logistic回归得到的结果相互验证，使结果在真实可靠的同时具有普适性与权威性。

本文还使用了大量图表来展示结果，使用条型图清晰直观地体现各个因素对视力的影响程度。

6.1.2 灰色预测模型

不需要很多数据，能利用微分方程充分挖掘系统的本质，精度高。

将无规律的原始数据生成的到规律性较强的生成序列。

6.1.3 分类向量机模型

此模型面向家长，只需要做一份类似于问卷调查的分析，即可大致知道孩子的视力状况，并得到关于视力以及与视力相关因素的预警，多角度地提醒家长和孩子注意对眼睛的保护。

该问题我们采取的方法是先用简单的线型内核函数的SVM进行分类，但发现分类效果不佳，后改用适用度更高的RBF函数多分类SVM进行分类，并达到预警的目的。

我们还结合简单的C语言程序设计出简易评判测试者的各个影响视力因素是否出现了问题，若出现了问题则发出该因素的预警，与近视预测结合，作出多角度的预警。

6.1.4 有导师学习神经网络模型

GRNN与PNN具有良好的泛化性能，且与BP神经网络等不同，其权值和阀值由训练样本一步确定，无须迭代，计算量小速度快。

隐含层节点的激活函数采用对输入信息具有局部激活特性的高斯函数，使得对接近于局部神经元特征的输入具有很强的吸引力。

GRNN与PNN的结果可以进行对比，在模型精度与运算速度上做出折中选择。

6.2 模型的不足

（1）两次问卷调查都是采取网络问卷调查的形式，由于乡村等偏远贫穷地区网络不发达，样本数据主要来自城市学生，导致采光条件和营养状况的结果较为单一，不能很好地体现这两类因素对视力的影响。

（2）模型较依赖样本数据的有效性，而本文采取问卷调查的形式获取的数据有效性并不高，仍需剔除一些坏值。

（3）SVM模型参数大多依靠经验选取或者大范围网格搜索耗时较长，算法复杂度较大

（4）在神经网络的训练过程中，训练样本与检验样本是从总样本中随机抽取的，这样可能造成典型样本点的缺失。

（5）灰色模型仅是通过统计方法得到的模型，无法描述该现象的内部机理。

七、模型的改进与推广

（1）SVM模型可以引入遗传算法、粒子群算法等优化算法，从而自动寻找最佳的模型参数使得模型的性能达到最优，降低计算时间。

（2）只要训练样本的数据分布合理（如数据在不同年龄段的分布，不同近视程度分布等）有导师神经网络的分类就能够达到足够的精度。

八、参考文献

[1]王志强,王冬妹,欧阳镇,唐锡麟.双生子屈光状态的研究[J].中国学校卫生,1992(04):206-208+256.
[2]娄晓民,吴敏,胡全忠,张爱敏.学生近视遗传度的分析[J].中国学校卫生,1994(02):139-140.
[3]陈国民,王洁贞,薛付忠.Bayes 公式分析用眼卫生习惯与视力不良的关系[J].中国学校卫生,2001(03):264-265.
[4]曾秋红,黄颖林,周月华.湘潭市 308 名学生 12 年视力变化追踪观察[J].预防医学情报杂志,2003(03):251-253.

[5]潘斌,吴梦奎.宁波市区初中学生近视现状及相关因素分析[J].中国公共卫生管理,2003(02):162-163.
[6]陈维格.体育锻炼与学生视力的关系[J].中国学校卫生,1994(03):182.
[7]宋俊生.教室采光照明对学生视力的影响[J].中国学校卫生,1996(05):355.
[8]徐蓓燕.学生视力不良与营养状况的关系[J].中国校医,2001(05):390-391.
[9]王丰效.修正 GM(1,1)模型在销售量预测中的应用[J].渭南师范学院学报,2003(05):10-11+58.
[10]LIU Sifeng, DENG Julong. GM (1,1)coding for exponential series[J]. The Journal of Grey System, 1999,2,147- 152.
[11]胡斌,曾学贵.不等时距灰色预测模型[J].北方交通大学学报,1998(01):38-42.
[12]王钟羡,吴春笃,史雪荣.非等间距序列的灰色模型[J].数学的实践与认识,2003(10):16-20.
[13]FENGXIAO WANG Department of Mathematics and Computer Science, Shaanxi University of Technology, Hanzhong, shaanxi, China. Improvement on Unequal Interval Grey Forecast Model[A]. 中国运筹学会.Proceedings of the Second Conference on Fuzzy Information & Engineering[C].中国运筹学会:中国运筹学会,2005:6.
[14]翟军,盛建明,冯英浚.MGM(1,n)灰色模型及应用[J].系统工程理论与实践,1997(05):110-114.
[15]王丰效.GM(1,1)组合预测模型及其应用[J].统计与决策,2006(21):142-143.
[16]郁磊,史峰,王辉,胡斐.MATLAB 智能算法 30 个案例分析(第 2 版)[M].北京:北京航空航天大学出版社,2015.
[17]司守奎,孙兆亮.数学建模算法与应用(第 2 版)[M].北京:国防工业出版社,2015.
[18]黄恒振,邓春亮,王朋炎,尹长明.基于 Logistic 回归模型的大学生视力影响分析[J].广西科学院学报,2010,26(01):13-16.

九、附录

附录 1 第二题灰色预测模型代码

```
clc,clear
%输入待预测时刻 k 及时间序列 X0
k=;
X0=[ ];
%对时间序列 X0 累加生成序列 X1
[n,m]=size(X0);
for j=1:m c=0;
for i=1:n
c=X0(i,j)+c;
X1(i,j)=c;
end
```

```
end
%计算数据矩阵 L
for j=1:m
for i=1:n-1
l(i,j)=(X1(i,j)+X1(i+1,j))/2;
end
end
L=[l ones(n-1,1)];
%计算 Y 及参数估计值
for j=1:m
Y(1:n-1,j)=X0(2:n,j);
a(:,j)=inv(L'*L)*L'*Y(1:n-1,j)
end
a=a';
A=a(1:end,1:end-1);
B=a(1:end,end);
%计算模型的拟合值或预测
S=X1(1,1:end);
if k==1 Z=S'
elseif k>1 Z=exp(A*(k-1))*S'+inv(A)*(exp(A*(k-1))-eye(size(exp(A*(k-1)))))
*B-(exp(A*(k-2))*S'+inv(A)*(exp(A*(k-2))-eye(size(exp(A*(k-2))))) *B)
else disp(' 输入错误! k 不得小于 1')
end
```

附录 2 第三题线型内核函数支持向量机

```
a0=load('fenlei.txt');
a=a0';b0=a(:,[1:150]);dd0=a(:,[151:end]);
[b,ps]=mapstd(b0);
dd=mapstd('apply',dd0,ps);
group=[ones(50,1);2*ones(100,1)];
s=svmtrain(b',group)
sv_index=s.SupportVectorIndices
beta=s.Alpha
bb=s.Bias
mean_and_std_trans=s.ScaleData
check=svmclassify(s,b')
err_rate=1-sum(group==check)/length(group)
solution=svmclassify(s,dd')
```

附录 3 第二题 RBF 核函数的多分类支持向量机

```
% 随机产生训练集和测试集
n = randperm(size(shili,1));
```

```
% 训练集——130 个样本
train_shili = shili(n(1:130),:);
train_label = label(n(1:130),:);
% 测试集——20 个样本
test_shili = shili(n(131:end),:);
test_label = label(n(131:end),:);

%% 数据归一化
[Train_shili,PS] = mapminmax(train_shili');
Train_shili = Train_shili';
Test_shili = mapminmax('apply',test_shili',PS);
Test_shili = Test_shili';

%% SVM 创建/训练(RBF 核函数)

% 寻找最佳 c/g 参数——交叉验证方法
[c,g] = meshgrid(-10:0.2:10,-10:0.2:10);
[m,n] = size(c);
cg = zeros(m,n);
eps = 10^(-4);
v = 5;
bestc = 1;
bestg = 0.1;
bestacc = 0;
for i = 1:m
    for j = 1:n
        cmd = ['-v ',num2str(v),' -t 2',' -c ',num2str(2^c(i,j)),' -g ',num2str(2^g(i,j))];
        cg(i,j) = libsvmtrain(train_label,Train_shili,cmd);
        if cg(i,j) > bestacc
            bestacc = cg(i,j);
            bestc = 2^c(i,j);
            bestg = 2^g(i,j);
        end
        if abs( cg(i,j)-bestacc )<=eps && bestc > 2^c(i,j)
            bestacc = cg(i,j);
            bestc = 2^c(i,j);
            bestg = 2^g(i,j);
        end
    end
end
cmd = [' -t 2',' -c ',num2str(bestc),' -g ',num2str(bestg)];
% 创建/训练 SVM 模型
model = svmtrain(train_label,Train_shili,cmd);
```

```
%% SVM 仿真测试
[predict_label_1,accuracy_1] = svmpredict(train_label,Train_shili,model);
[predict_label_2,accuracy_2] = svmpredict(test_label,Test_shili,model);
result_1 = [train_label predict_label_1];
result_2 = [test_label predict_label_2];

%% 绘图
figure
plot(1:length(test_label),test_label,'r-*')
hold on
plot(1:length(test_label),predict_label_2,'b:o')
grid on
legend('真实类别','预测类别')
xlabel('测试集样本编号')
ylabel('测试集样本类别')
string = {'测试集 SVM 预测结果对比(RBF 核函数)';
          ['accuracy = ' num2str(accuracy_2(1)) '%']};
title(string)
```

附录 4　第二题中变量值

发展系数 A=

0.0654	-0.1801	0.0912	0.1605	-0.0401
0.0759	-2.3533	-1.7369	3.8304	-1.6882
-0.2217	-1.6514	-1.6907	5.3865	-2.9872
0.7582	-0.1636	-0.6744	1.7319	-3.3476
-0.4942	1.6054	1.4603	-5.6958	5.532

灰作用量 B=

6.5196
0.2612
2.2639
6.2292
1.5793

附录 5 第四题有导师学习神经网络的分类代码

```
%% 清空环境变量
clear all
clc

%% 训练集/测试集产生
```

```
% 导入数据
load X.mat
% 随机产生训练集和测试集
P_train = [];
T_train = [];
P_test = [];
T_test = [];
for i = 1:3
    temp_input = features((i-1)*50+1:i*50,:);
    temp_output = classes((i-1)*50+1:i*50,:);
    n = randperm(50);
    % 训练集——120 个样本
    P_train = [P_train temp_input(n(1:40),:)'];
    T_train = [T_train temp_output(n(1:40),:)'];
    % 测试集——30 个样本
    P_test = [P_test temp_input(n(41:50),:)'];
    T_test = [T_test temp_output(n(41:50),:)'];
end

%% 模型建立
result_grnn = [];
result_pnn = [];
time_grnn = [];
time_pnn = [];
for i = 1:4
    for j = i:4
        p_train = P_train(i:j,:);
        p_test = P_test(i:j,:);
      %% GRNN 创建及仿真测试
        t = cputime;
        % 创建网络
        net_grnn = newgrnn(p_train,T_train);
        % 仿真测试
        t_sim_grnn = sim(net_grnn,p_test);
        T_sim_grnn = round(t_sim_grnn);
        t = cputime - t;
        time_grnn = [time_grnn t];
        result_grnn = [result_grnn T_sim_grnn'];
      %% PNN 创建及仿真测试
        t = cputime;
        Tc_train = ind2vec(T_train);
        % 创建网络
        net_pnn = newpnn(p_train,Tc_train);
        % 仿真测试
```

```
        Tc_test = ind2vec(T_test);
        t_sim_pnn = sim(net_pnn,p_test);
        T_sim_pnn = vec2ind(t_sim_pnn);
        t = cputime - t;
        time_pnn = [time_pnn t];
        result_pnn = [result_pnn T_sim_pnn'];
    end
end

%% 性能评价

% 正确率 accuracy
accuracy_grnn = [];
accuracy_pnn = [];
time = [];
for i = 1:10
    accuracy_1 = length(find(result_grnn(:,i) == T_test'))/length(T_test);
    accuracy_2 = length(find(result_pnn(:,i) == T_test'))/length(T_test);
    accuracy_grnn = [accuracy_grnn accuracy_1];
    accuracy_pnn = [accuracy_pnn accuracy_2];
end
% 结果对比
result = [T_test' result_grnn result_pnn]
accuracy = [accuracy_grnn;accuracy_pnn]
time = [time_grnn;time_pnn]
%% 绘图
figure(1)
plot(1:30,T_test,'bo',1:30,result_grnn(:,4),'r-*',1:30,result_pnn(:,4),'k:^')
grid on
xlabel('测试集样本编号')
ylabel('测试集样本类别')
string = {'测试集预测结果对比(GRNN vs PNN)';['正确率:' num2str(accuracy_grnn(4)*100)
'%(GRNN) vs ' num2str(accuracy_pnn(4)*100) '%(PNN)']};
title(string)
legend('真实值','GRNN 预测值','PNN 预测值')
figure(2)
plot(1:10,accuracy(1,:),'r-*',1:10,accuracy(2,:),'b:o')
grid on
xlabel('模型编号')
ylabel('测试集正确率')
title('10 个模型的测试集正确率对比(GRNN vs PNN)')
legend('GRNN','PNN')
figure(3)
plot(1:10,time(1,:),'r-*',1:10,time(2,:),'b:o')
```

```
grid on
xlabel('模型编号')
ylabel('运行时间(s)')
title('10 个模型的运行时间对比(GRNN vs PNN)')
legend('GRNN','PNN')
```

附录 6 神经网络训练样本数据库

你的视力程度是多少	你平时户外锻炼的时间大约多久	你一天的用眼负荷是多少（视近物与用眼持续时间长）	你的读写习惯符合下列几条	你的父母近视吗
1	3	3	2	1
1	4	3	4	3
1	3	2	3	3
1	5	3	3	3
1	4	3	4	1
1	3	4	2	3
1	3	3	2	3
1	3	3	4	1
1	5	2	2	3
1	3	2	2	3
1	4	1	4	3
1	4	1	4	3
1	3	2	3	3
1	2	2	4	1
1	3	2	4	3
1	4	3	1	1
1	4	4	2	3
1	3	2	3	3
1	4	1	1	1
1	3	3	3	3
1	4	3	2	3
1	3	1	4	3
1	4	1	1	3
1	3	4	4	3
1	3	2	4	3
1	3	1	1	3

1	4	3	1	3
1	3	4	3	3
1	3	4	2	3
1	2	4	3	3
1	3	2	4	3
1	4	4	4	1
1	3	2	3	3
1	2	2	3	3
1	2	3	2	3
1	3	4	3	1
1	4	4	3	3
1	2	1	3	3
1	3	4	4	3
1	2	3	2	1
1	3	1	4	3
1	3	1	4	3
1	5	3	3	3
1	3	3	2	1
1	5	2	2	3
1	4	2	4	3
1	4	3	3	3
1	3	3	2	3
1	3	4	1	3
1	5	2	2	3
2	2	4	3	3
2	2	5	2	1
2	2	3	4	1
2	2	2	3	1
2	2	3	4	3
2	2	3	3	1
2	3	4	4	1
2	2	5	4	3
2	2	3	3	1
2	3	4	4	1
2	2	3	3	1
2	2	5	4	1
2	3	3	4	3
2	3	4	3	1
2	3	5	3	3
2	2	3	4	1
2	3	5	2	1
2	2	3	4	1
2	2	3	4	1
2	3	4	4	1

2	2	3	4	3
2	2	3	4	1
2	1	3	4	3
2	2	3	4	2
2	3	4	2	1
2	2	3	3	3
2	3	4	3	2
2	3	5	4	1
2	2	4	5	1
2	4	4	4	1
2	3	3	3	1
2	3	3	5	1
2	2	3	4	3
2	3	4	3	1
2	2	4	5	3
2	3	3	3	1
2	2	3	5	1
2	2	4	4	1
2	2	3	5	1
2	3	3	4	1
2	2	5	3	1
2	3	5	5	1
2	2	4	5	1
2	2	4	4	1
2	2	4	5	1
2	3	4	4	1
2	2	4	5	1
2	2	4	4	1
2	2	3	3	1
2	2	5	3	3
3	2	5	4	1
3	2	5	4	1
3	1	5	5	1
3	1	5	3	1
3	2	4	3	2
3	3	5	6	2
3	1	5	6	1
3	2	4	3	1
3	1	5	6	3
3	2	4	5	2
3	3	5	4	1
3	2	4	4	1
3	1	5	4	3
3	2	3	4	1

3	2	5	5	2
3	1	5	5	1
3	1	6	4	3
3	2	5	5	1
3	2	5	5	1
3	1	5	4	2
3	2	4	5	2
3	2	5	3	1
3	2	4	5	2
3	1	3	4	1
3	2	5	6	2
3	1	3	3	1
3	2	5	6	2
3	1	5	6	1
3	2	3	6	2
3	2	5	6	1
3	1	5	5	2
3	1	5	5	1
3	1	5	5	1
3	2	4	6	2
3	1	5	4	3
3	2	3	6	2
3	1	5	6	2
3	1	5	6	2
3	2	4	4	2
3	1	5	6	1
3	2	4	5	2
3	1	5	4	3
3	2	5	5	1
3	2	5	5	2
3	2	4	5	2
3	2	5	6	1
3	2	3	6	2
3	1	4	6	1
3	1	5	6	2
3	1	5	6	1