**医保欺诈行为的主动发现**

长春理工大学

贾而穑 陈俊 叶秀玲

指导老师：周林华

## 摘 要

随着医疗保障制度的普及，医保欺诈案件呈逐年递增趋势，不仅损害了医保经营的诚信原则，也威胁到了医保基金的安全。为此，如何对医保欺诈行为实现及时自主地发现尤为重要。本文通过：①利用可信度较高的欺诈指示因子实现数据初步筛选；②基于BP神经网络和蒙特卡洛方法挖掘各类数据之间的隐匿关系，实现数据精细筛选；③对医保欺诈数据进行信息挖掘，得出三类不同医保欺诈的特征信息。最终，形成了一套完整的医保欺诈数据挖掘方案。

（1）数据初步筛选：基于病人ID号等信息将所有数据建立联系；根据病人单张处方总价、就医次数等因素，确定4个欺诈指示因子，将总体样本划分为医保欺诈数据（Y类数据）及疑似医保欺诈数据（N类数据）。

（2）数据精细筛选：根据医嘱项、买药总数量、就医次数等9个BP神经网络输入因子，利用MATLAB软件建立BP神经网络；在初步筛选的基础上，随机抽取1组数据作为BP神经网络的训练样本；利用样本进行神经网络训练与检测；进一步利用训练好的神经网络对总体进行识别；重复以上步骤n次（建立BP神经网络除外），实现蒙特卡洛随机试验；若每次试验最终确定的医保欺诈者类的个数与n次随机试验所得结果并集的个数之比都大于90%，则认为该方法可行，最终可以识别出医保欺诈者类。

（3）医保欺诈数据信息挖掘：通过统计所有医保欺诈者ID出现的频数，可以将医保欺诈数据分为重大嫌疑，次要嫌疑及嫌疑较小三类；进一步通过绘制相关信息盒图以及人员分布直方图等，实现三类ID号不同信息的特征识别，并确定了可将ID号进行分类的原因；最后，利用SPSS进行聚类分析，证实了我们的划分方法是可行的。

（4）利用Logistic回归分析对神经网络模型进行改进：对确定的9个输入因子进行二元回归分析，根据回归系数，找出其中的弱因子并将其剔除，在一定程度上可降低对BP神经网络识别的干扰，提高BP神经网络的准确度。

**关键词：**医保欺诈 数据挖掘 BP神经网络 蒙特卡洛模拟

## 一．问题重述

我国当前医疗保险欺诈呈现蔓延态势。医疗保险欺诈，是指公民、法人或者其他组织在参加医疗保险、缴纳医疗保险费、享受医疗保险待遇过程中，故意捏造事实、弄虚作假、隐瞒真实情况等造成医疗保险基金损失的行为。骗保人进行医保欺诈时通常使用的手段，一是拿着别人的医保卡配药，二是在不同的医院和医生处重复配药。下面这些情况都有可能涉嫌医保欺诈：单张处方药费特别高，一张卡在一定时间内反复多次拿药等。我们要根据题目附件中的数据，找出可能的欺诈记录。

## 问题分析

医疗保险欺诈的识别问题事实上就是索赔分类问题，即区分是欺诈索赔还是合法。通过分析索赔人信息、索赔信息、治疗情况，例如身份证号、医疗手册号、性别、购药单价、总价、医嘱类及医嘱子类，选择合适的欺诈指示因子，用统计学的方法建立识别模型。

由于题给医疗保险欺诈数据量庞大且含有较多的噪声数据，特征变量较多，而BP神经网络具有较强的处理含噪声数据的能力以及处理不严密的知识和定性特征变故的,故我们初步采用BP神经网络这种先进的人工智能技术。

首先根据附表2.1和附表2.2中病人ID，身份证及医嘱项等相关信息，初步选择a个欺诈指示因子，将总体数据进行分类，完成对总体样本的初步筛选。对初步筛选的结果赋予相应的期望，为BP神经网络提供必要的输入数据。

随机抽取初步筛选的结果，得到n组BP神经网络的训练样本。通过分析附表所给的相关信息，如：病人ID，身份证号，医嘱项，核算分类等，我们初步确定了b个模糊因子作为BP神经网络的输入因子，建立BP神经网络并对随机抽取的n组样本数据进行训练，检测。最后我们用所训练好的的n组BP神经网络对总体进行识别，输出结果。并通过设定频数阈值，将输出的结果分为重大嫌疑，次要嫌疑，嫌疑较小三类。至此精细筛选部分完成。

为了验证我们对总体识别结果的划分是可行的，我们采用SPSS进行了Ward聚类，并将聚类结果与我们划分的结果相比较，以证实我们的分类方式是合理的。

由于我们不知道b个模糊因子对医保欺诈识别的影响,故我们考虑到用Logistics回归分析对BP神经网络模型进行改进。利用Logistic回归分析对神经网输入因子进行筛选，输出每一个模糊因子对应的权重，剔除弱因子，提高BP神经网络识别的准确率。

## 模型假设

1.ID号为692316的年龄数据不详，ID号为669432的年龄为140（严重与事实不符），故剔除不予以考虑；

2.ID号为687972,31551,580015的购买频率为0，故不计入欺诈索赔的范围内；

3.若BP神经网络检测准确率大于90%，则可认为所建立的BP神经网络合理；  
4.假设数据预处理的a个欺诈指示因子相互之间没有联系；

5.所有的医疗数据均来自同一家医院。

## 符号说明

欺诈指示因子A情况下，对总体样本数据筛选出的集合

欺诈指示因子B情况下，对总体样本数据筛选出的集合  
 欺诈指示因子C情况下，对总体样本数据筛选出的集合  
 欺诈指示因子D情况下，对总体样本数据筛选出的集合

n 从总体随机抽取样本的个数

将样本进行BP神经网络训练，检测，然后对总体进行识别后输出的结果

N() 表示集合中的个数。

## 模型建立

本文以数据挖掘为主线，建立了三重数据挖掘的数学模型，以实现对总体数据的初步筛选，精细筛选和信息挖掘，具体的步骤如下（如图-1所示）：

第一步：初步筛选。确立若干个欺诈指示因子，将总体数据划分为欺诈类和疑似欺诈类。认定欺诈类是已经确定有问题的数据。

第二步：精细筛选。通过BP神经网络以及蒙特卡洛方法，实现对总体数据的更进一步挖掘；

第三步：信息挖掘。对前两步所确定的问题数据进行第三次数据挖掘，以确定所有问题数据其背后的信息特征。

下图-1为三次数据挖掘数学模型的流程图。

原始数据

第一次数据挖掘

欺诈指示因子进行初步筛选

疑似欺诈类

欺诈类

随机抽取

训练样本

第二次数据挖掘

训练BP神经网络

BP神经网络和蒙特卡洛方法进行精细筛选

信息挖掘

识别结果

识别总体

对嫌疑ID号背后的信息特征进行挖掘

第三次数据挖掘

图-1

### 5.1 初步筛选的模型建立

#### （1）欺诈因子确定

通过对附表2-1的身份证和ID号及统计买药次数等信息，我们初步确定了4个欺诈指示因子。

因子A：不满足医疗保险号与身份证的唯一识别

因子B：单张处方费用特别高

4个欺诈指示因子

因子C：同个病人的ID买药频率过高

因子D：在不同的医院和医生处重复配药

将总体样本划分为欺诈样本数据（Y类）和疑似欺诈数据（N类），分别赋予期望值0和1。

##### ① 因子A：不满足医疗保险号与身份证的唯一识别

由我国医疗保险相关政策可知，医疗保险号与身份证具有唯一识别性，即一张身份证只能有一个医疗保险号，一个医疗保险号也只能对应一张身份证，如果我们从题给表2-1病人资料中找出满足因子A的数据，那么此类数据可以归为欺诈行为。

##### ② 因子B: 处方费用特别高

通过Excel对附表2-2.费用明细表总价一栏进行数据排序，我们发现有些账单费用特别高,而通过附表2-5.核算分类表，我们发现账单费用均用来购买药品，故排除因手术等高收费项目导致的处方费用特别高。通过统计概率密度，对概率密度进行有限元积分，求出概率并且设定概率阈值,并找出对应的处方费用。则超过该处方费用的的数据归为医疗欺诈数据。

##### ③ 因子C：同个病人的ID买药频率过高

通过分析附表2-2.费用明细表，同一个病人ID号在一个月的时间频繁买药，最高的次数达到128次，远远超过正常次数。因此，我们认为频率过高的数据有可能是医疗欺诈。同样的,我们通过统计概率密度，对概率密度进行有限元积分，求出概率并且设定概率阈值,找出相对应的频数，则超过该处方费用的数据归为医疗欺诈数据。

##### ④ 因子D：在不同的医院和医生处重复配药

通过分析表2-2费用明细表，我们发现存在一个病人ID买相同数量的同种药物的情况，极有可能是在不同的医院或者不同的医生处开的药，但由于此类数据较少，我们选取频率较高的前两组数据归为医疗欺诈数据。

#### （2）对样本进行分类

按照以上四个欺诈指示因子，我们将总体样本选出 ，，四个集合。Y类样本数据应全部包含4个欺诈指示因子所确定的集合，故需要对四个集合进行并运算:

所求结果为总体样本Y类样本数据。样本中剩下的数据归为N类样本数据。

### 5.2 精细筛选的模型建立

根据医嘱项、买药总数量、就医次数等9个BP神经网络输入因子，利用MATLAB软件建立BP神经网络；在初步筛选的基础上，随机抽取1组数据作为BP神经网络的训练样本；利用样本进行神经网络训练与检测；进一步利用训练好的神经网络对总体进行识别；重复以上步骤n次（建立BP神经网络除外），实现蒙特卡洛随机试验；若每次试验最终确定的医保欺诈者类的个数与n次随机试验所得结果并集的个数之比都大于90%，则认为该方法可行，最终可以识别医保欺诈者类。具体步骤如下：

第一步，基于蒙特卡洛思想，从初步筛选阶段所划分的两类ID号中分别随机抽取n组样本，赋予期望0和1后组合成n组BP神经网络的训练样本；

第二步，确定模糊因子作为BP神经网络的输入因子，并确定隐含层和输出层，建立BP神经网络；

第三步，对样本(i=1,2……n)进行训练和检测,并对输出的准确率进行分析，检测神经网络的参数及模糊因子设置是否合理；

第四步，若模糊因子设置合理，则用训练好的BP神经网络对总体进行识别；

第五步，对样本重复上述第二步到第四步，实现蒙特卡洛模拟,并输出结果；

第六步，若90%以上，则认为欺诈指示因子可信，建立的BP神经网络具备识别医保欺诈的功能；

第七步，统计的病人ID号在n组输出数据中出现的频数，通过设定频率阈值，确定出重大嫌疑,次要嫌疑及嫌疑较小的ID号。

精细筛选的流程图如图-2所示：

#### （1）BP神经网络的结构设计

##### ① 输入层的输入因子确定

在数据预处理的部分，利用附表2.1病人资料及附表2.1费用明细表的相关数据，初步确定了4个欺诈指示因子，并将随机抽取的样本划分为Y类和N类。然而实际上，确定一个ID号是否存在欺诈行为的判断依据远不止4个，对于一些模糊的，不能确定是否和欺诈有关联的影响因素，我们并不能很好的去鉴别。因此我们根据附表，所给的性别，年龄，医嘱项，下属科室等相关信息，又设定了9个模糊因子作为BP神经网络的输入因子，并通过BP神经网络来进行判定其设定的合理性，9个神经网络输入因子如下表-1所示：

|  |  |  |
| --- | --- | --- |
| 性别 | 年龄 | 买药频率 |
| 买药总花费 | 买药总数量 | 下嘱科室 |
| 执行科室 | 病人科室 | 医嘱重复因子 |

表-1

根据BP神经网络的相关特性，训练样本与输入因子的要求。上表所确定的性别，年龄，买药频率等前8项均可在题给数据或经过相应的统计后可以获得。

对于医嘱项来说，我们经过数据处理找出一个病人ID号对应多个医嘱项，故医嘱项不能直接作为输入因子。但是，可以求得某一因子，即医嘱重复因子

其表达式为：，当越小时，表示该ID的相对重复率越大，越可能涉嫌医保欺诈。所以我们最后将输入医嘱重复因子作为输入因子。

Num：相同ID号医嘱的总数；kind：相同ID号医嘱的种数。

欺诈类

疑似欺诈类

蒙特卡洛抽样

……

样本

样本

样本

初步筛选

赋予期望值

赋予期望值

……

……

神经网络训练与检测

神经网络训练与检测

……

对总体识别

对总体识别

……

……

输出结果

**否：调整模糊因子及欺诈指示因子**

**是**

输出结果

图-2

##### ② 隐含层节点数的确定

对于医保诈骗识别的BP神经网络，我们设定其隐含层层数为3，并依据经验公式：

（M:隐含层节点个数，n：输入层神经元个数，m：输出层神经元个数，a：[0,10]之间的常数）

确定隐含层节点个数如下：

|  |  |
| --- | --- |
| **隐含层** | **节点个数** |
| Layer1 | 9 |
| Layer2 | 10 |
| Layer3 | 2 |

表-2

##### ③ 输出层的确定

根据先前数据预处理的模型建立部分，我们已经对Y类和N类赋予相应的期望：

|  |  |
| --- | --- |
| **数据类别** | **期望** |
| Y类数据 | **0** |
| N类数据 | **1** |

表-3

因此，输出层的节点个数设定为1，输出的值为0或者1。

#### （2）神经网络的建立

|  |  |
| --- | --- |
| BP神经网络 | |
| 创建函数 | newff |
| 传递函数 | Logsig |
| 训练方法 | Trainlm |
| 神经网络学习率 | 0.01 |
| 神经网络训练目标 | 0.001 |

建立的神经网络如图-3所示：

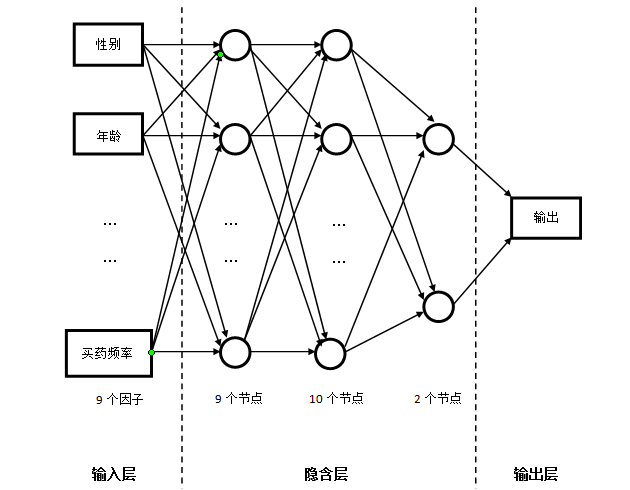


图-3

#### （3）用BP神经网络实现整体数据的精细筛选

其具体步骤为：

第一步：将训练样本加入白噪声，进行BP神经网络的训练，得到训练样本的识别结果；

第二步：将的识别结果与该训练样本的期望值进行匹配，返回识别准确率，若存在准确率小于90%的情况，则改变模糊因子及BP神经网络相关参数并对n组样本重新进行训练，重复第1,2步；

第三步：若模糊因子及BP神经网络相关参数设置合理，用训练好的BP神经网络对总体进行识别；

第四步：对训练样本重复这四步操作，实现蒙特卡洛模拟，输出结果为。

#### （4）蒙特卡洛模拟结果分析

通过上述过程，得出每组样本的输出结果。若每个样本确定的疑似医保欺诈者类的个数与所有蒙特卡洛随机试验所得结果求并集的个数之比>90%,则说明初步选取的13个因子（4个欺诈指示因子，9个模糊因子）及神经网络的参数合理，输出结果；若存在一组识别结果，使得<90%,则我们需要重新选取因子和设定BP神经网络系统的相关参数，用修改后的参数，重新对n组样本进行训练，检测，识别，直到>90%为止。

### 5.3医保欺诈类特征信息挖掘的模型建立

#### （1）医保欺诈数据的聚类分析

第一步：统计中每个病人ID号在的频数，以频数为横坐标，出现该频数的ID个数为纵坐标，转换为概率密度曲线；

第二步：通过积分，确定概率阈值；

第三步：根据第二步所确定的阈值，以及每个ID号在输出结果中出现的频数，将中的ID号划分为重大嫌疑，次要嫌疑，嫌疑较小三类，并作为识别结果。

#### （2）医保欺诈数据的特征信息挖掘

对于经过初次筛选和精细筛选后所得到的三类问题ID号，通过以下两种方式进行其背后的信息挖掘。

1. 计算三类ID号在不同项目的均值，并在三类ID之间进行比较，以发现三类ID号之间的区别；
2. 统计并绘制三类ID号在不同项目里的人员分布直方图，从而更加直观的发掘不同类ID号之间的人员分布特征。

## 六．模型求解

### 6.1 初步筛选的模型求解

#### （1）欺诈因子的确定

##### ① 因子A：不满足医疗保险号与身份证的唯一识别

对于此种情况，我们将样本导入MATLAB中，通过MATLAB程序，见附录1，遍历所有的数据，找出满足因子A的所有数据，作为，其个数为1315个。下表-4为的部分数据。

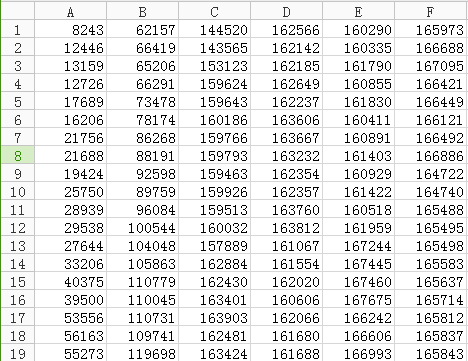


表-4

##### ② 因子B: 单张处方费用特别高

针对此种情况，我们将样本中单张处方总价按照概率密度绘制成图像，如下图-3。

通过编写MATLAB程序，对其进行有限元的积分，并通过调整积分上限来选择我们需要的阈值，具体程序见附录2。

当时，概率为0.877482；

当时，概率为0.977291；

当 时，概率为0.9964。

为了保证训练样本的准确性，概率越大越好。下图-4为概率密度曲线，经过不断选择阈值，最终我们选出了阈值概率值为0.9964，即当概率为0.0033时视为不可能事件。通过这一步我们将总价大于455的数据筛选出来，选出1382个数据作为，下表-5为部分数据。

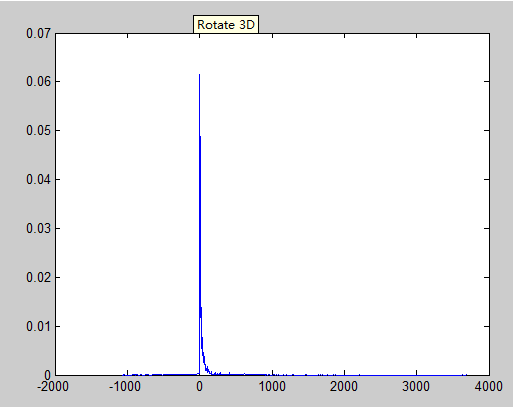


图-4

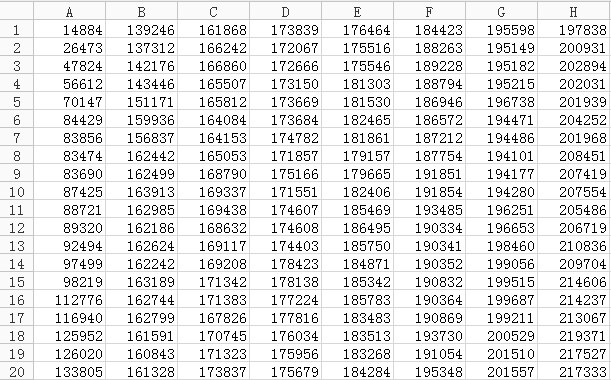


表-5

##### ③ 因子C：同个病人的ID买药频率过高

对总体数据进行分析，通过程序附录3找出病人ID号买药的频率，并描绘出概率密度曲线如下图-5，通过不断调整阈值，计算其概率。

当时，概率为0.852622；

当时，概率为0.985。

最终我们选定概率阈值为0.985，利用Excel找出的数据的集合作为，部分数据见下表-6。

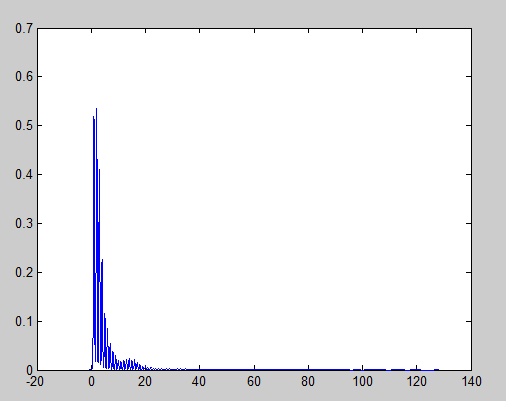


图-5

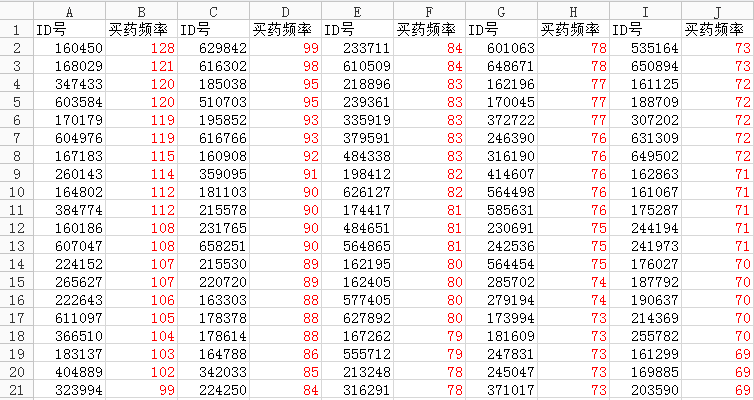


表-6

##### ④ 因子D：在不同的医院和医生处重复配药

对总体的单价，数量分析发现，有些病人ID号在短时间内买了相同数量的同种药，可能是病人重复买药。通过附录4程序我们找出可能重复买药的数据，如下表-7。我们取频率前两组的数据集合作为。

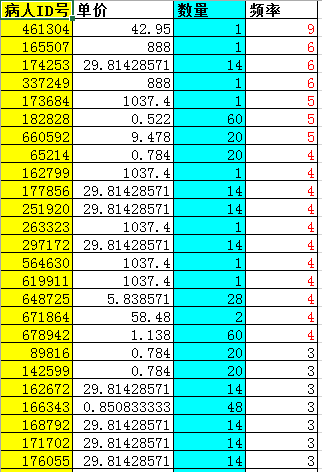


表-7

#### （3）对样本进行分类

通过附录6程序,求 ，部分数据如下表-8，自此总体样本的初步分类完成，共计2209个欺诈类ID号。MATLAB程序详见附录5.

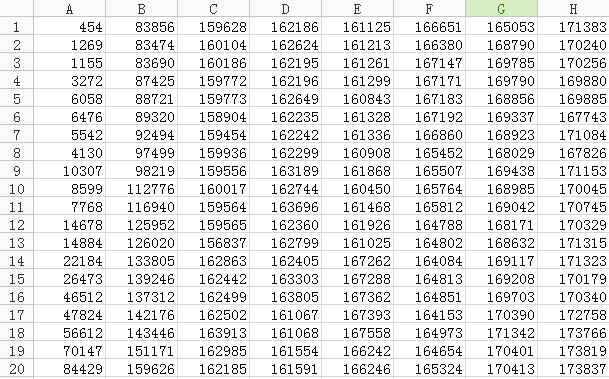


表-8

### 6.2 精细筛选模型求解

#### （1） 相应数据的匹配

在输入因子的数据统计方面，我们利用MATLAB的数据I/O编程，将每一个ID号都进行了匹配，对于不方便匹配的输入因子——医嘱项，由于每一位ID号所对应的医嘱项有很多且又不完全相同，因此，我们变相的统计出每一ID号的种类与总数量，并将每一ID号的医嘱种类与总数量的比值信息作为输入信息，统计结果如下表-9：（限于篇幅，这里只展示部分数据）

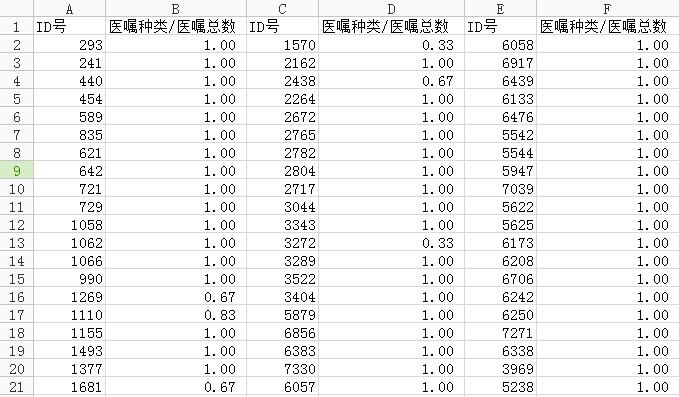


表-9

MATLAB程序见附录6程序A.

最后，ID号匹配信息如下表-10所示：（限于篇幅，这里只展示部分数据）



表-10

MATLAB程序见附录6程序B。

#### （2） 训练样本的随机抽取

从初步筛选阶段所划分的欺诈类和疑似欺诈类中分别随机抽取2000个ID号，并分别赋予期望值0和1，组合作为BP神经网络的训练样本。重复上述操作100次，共获得100组训练样本。选取结果如下表-11所示：（限于篇幅，这里只展现第52次随机抽取的部分数据结果）

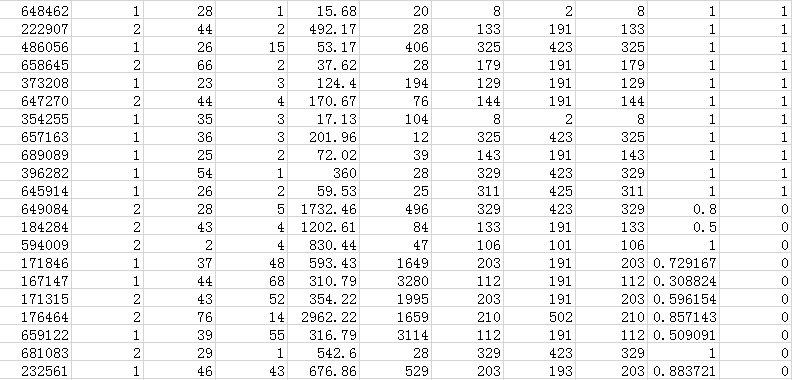


表-11

#### （3） BP神经网络建立及其测试

MATLAB程序见附录7。

对训练样本数据建立的神经网络如下图-6所示：

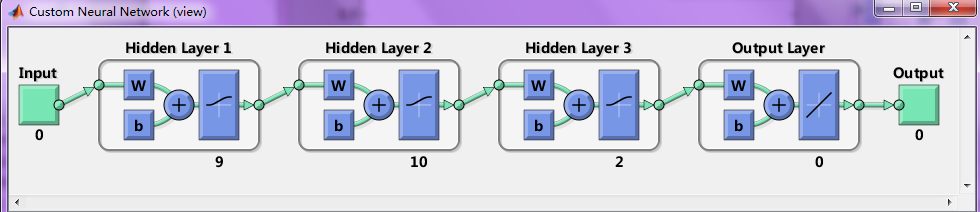


图-6

BP神经网络的训练过程：（以第52组训练样本数据为例）

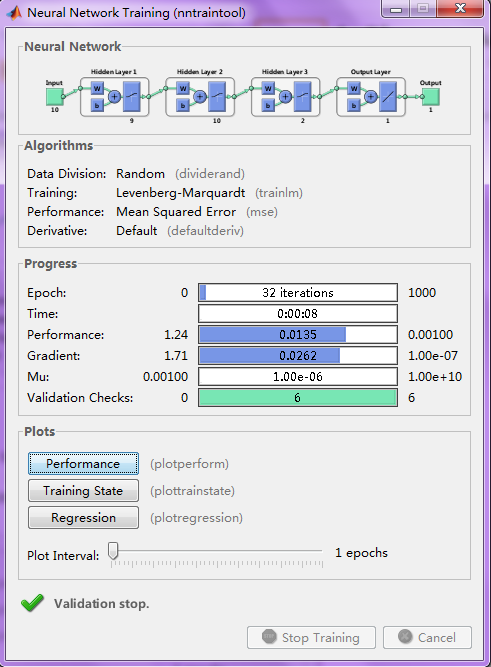


图-7

由图-7可以发现，对于第52组训练样本的训练过程，一共进行了32次迭代，用时8秒。用未加入噪声的训练样本检测BP神经网络，将输出的结果与未加入噪声的训练样本进行匹配，程序运行的结果如下：

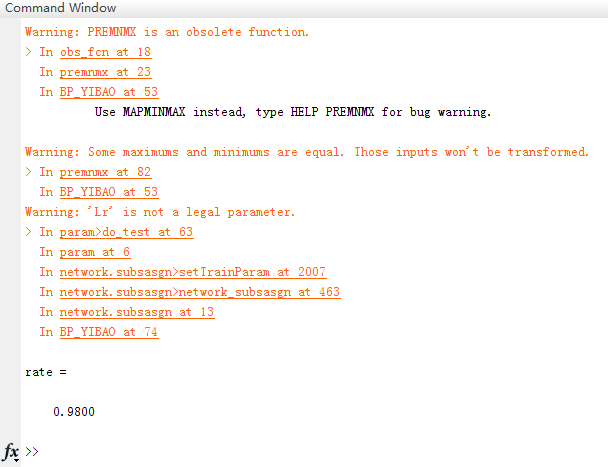


图-8

可见，对训练样本数据的识别准确率可达 98%。

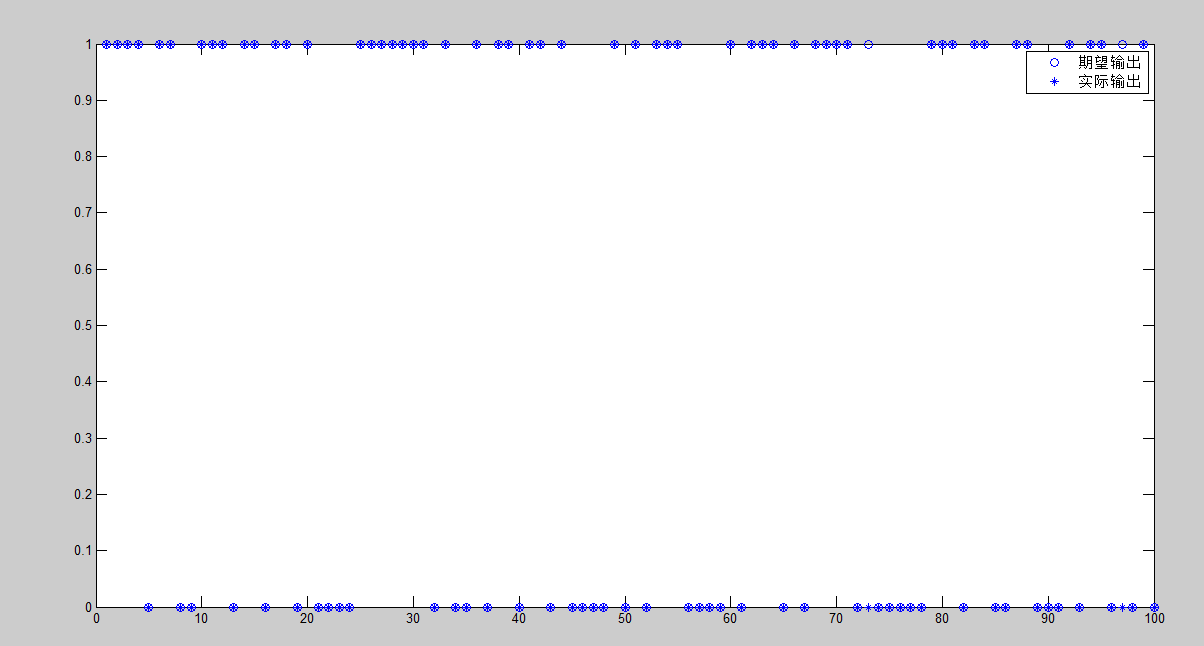


图-9

通过图-9可以更直观的发现，期望输出的‘o’与检测时的实际输出‘\*’基本吻合，而只有极少数的点出现了误判，从而说明了该神经网络是可靠的。

**误差变化曲线如下图所示（performance图）：**

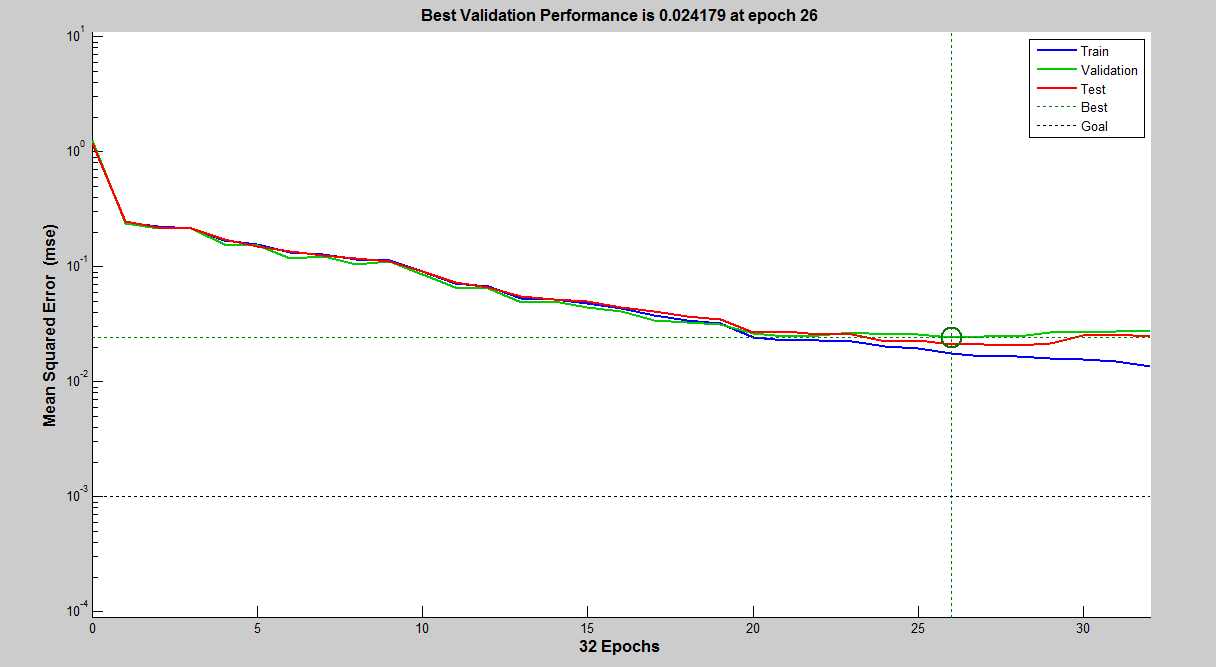


图-10

#### （4） 用BP神经网络实现总体数据的分类识别

MATLAB程序见附录8。

将随机抽取的100组训练样本数据分别进行训练，得到的检测准确率结果如下表-12：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **BP神经网络检测结果（准确率）** | | | | | | | | | |
| **97%** | **98%** | **95%** | **98%** | **98%** | **97%** | **98%** | **97%** | **98%** | **98%** |
| **99%** | **98%** | **100%** | **97%** | **96%** | **100%** | **97%** | **93%** | **100%** | **96%** |
| **98%** | **97%** | **99%** | **99%** | **95%** | **100%** | **99%** | **99%** | **97%** | **96%** |
| **99%** | **97%** | **96%** | **97%** | **99%** | **99%** | **98%** | **99%** | **97%** | **94%** |
| **98%** | **98%** | **100%** | **97%** | **97%** | **99%** | **98%** | **100%** | **98%** | **98%** |
| **98%** | **98%** | **98%** | **98%** | **99%** | **97%** | **98%** | **96%** | **99%** | **97%** |
| **98%** | **97%** | **99%** | **98%** | **98%** | **97%** | **97%** | **99%** | **97%** | **96%** |
| **98%** | **99%** | **99%** | **99%** | **96%** | **98%** | **99%** | **99%** | **99%** | **96%** |
| **98%** | **99%** | **97%** | **96%** | **96%** | **98%** | **99%** | **94%** | **99%** | **95%** |
| **99%** | **99%** | **98%** | **96%** | **96%** | **98%** | **98%** | **97%** | **97%** | **100%** |

表-12

#### （5） 蒙特卡洛模拟结果分析

由表-12分析可知，对于每一组样本数据所建立的BP神经网络，当用未加入噪声的训练数据来检测BP神经网络时，其准确率的均值几乎达到98%，而且甚至有6次其准确率达到了100%，可见：

1）对于随机产生的训练样本所建立的BP神经网络，均能达到非常高的识别率；

2）所有建立起来的BP神经网络在检测的过程中均表现出合格，因此说明，该BP神经网络的传递函数，学习函数合理。而且，先前设定的4个欺诈识别因子，9个模糊因子也均是合理的。

将所有的检测准确率表现在下图-11中：

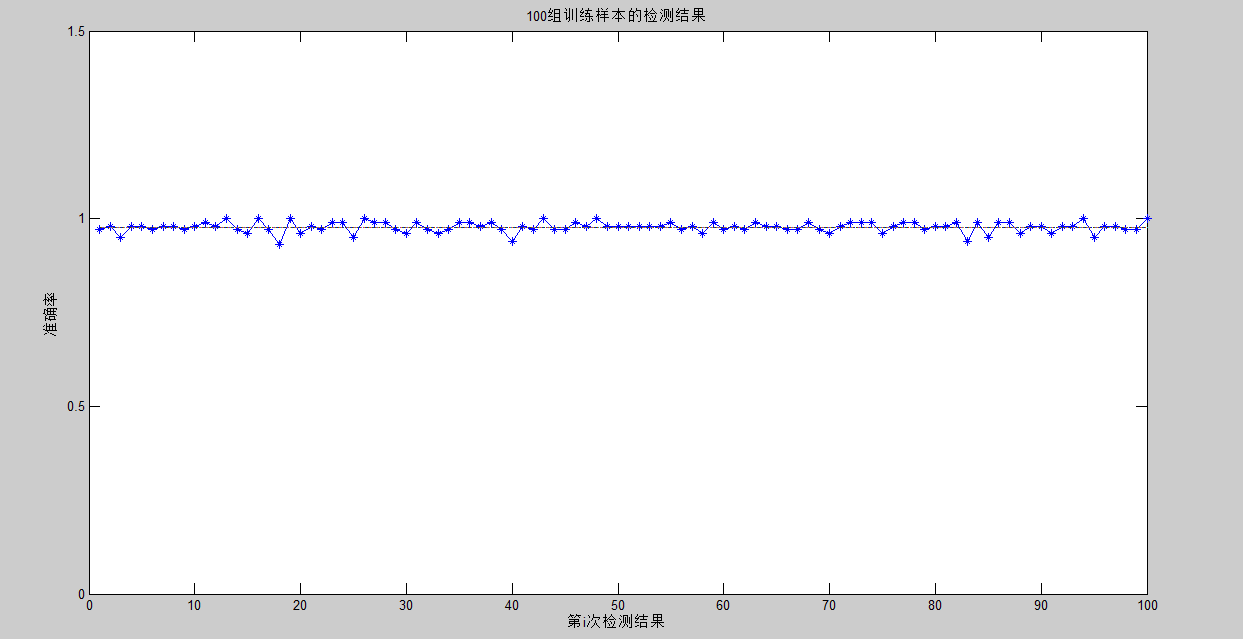


图-11

由上图可以发现，整体的检测准确率均在98%上下。

#### （6） 最终识别结果的确定

我们将初步筛选的结果与神经网络精细筛选的结果取并集运算，即：

欺诈ID号 =

一共得到5245个问题ID号作为本次医保欺诈自主识别的最终结果。在整个欺诈数据挖掘的过程中，通过欺诈指示因子对原始数据的初步筛选以及从模糊因子出发的BP神经网络精细筛选，实现了全面的自主识别。

### 6.3医保欺诈类特征信息挖掘的模型求解

#### （1）医保欺诈数据的聚类分析求解

对于不同训练样本训练出的BP神经网络，其对总体识别的结果也会有一定的不同。故我们统计同一个ID号在100组问题数据中出现的次数，统计结果如下表-13：（限于篇幅，这里只展示部分数据）



表-13

根据表-13，我们绘制其概率密度图，如图-12所示

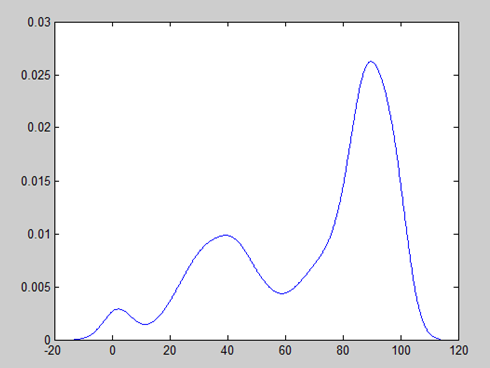


图-12

通过对其概率密度图的分析，发现ID号出现的次数呈现出三处集中点，分别为：90次、40次、3次。因此，我们根据频率的分布,对于所有的ID号进行如下的划分：

|  |  |  |
| --- | --- | --- |
| 嫌疑程度 | 阈值（出现次数） | 共计人数 |
| 重大嫌疑 | 60-100 | 2109 |
| 次要嫌疑 | 22-59 | 929 |
| 嫌疑较小 | 1-21 | 180 |

表-14

划分后的ID号数据如下表所示：

（限于篇幅，这里只展示部分数据，详细数据见附录8）

**重大嫌疑的ID号：**

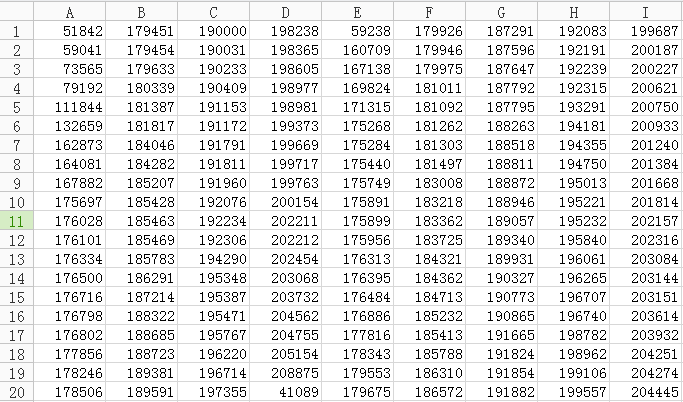


表-15

**次要嫌疑ID号：**

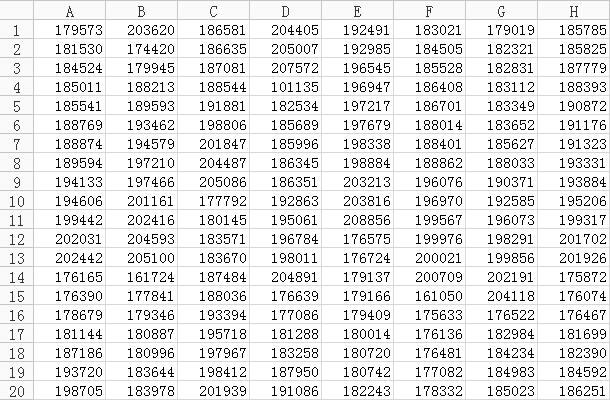


表-16

**嫌疑较小ID号：**

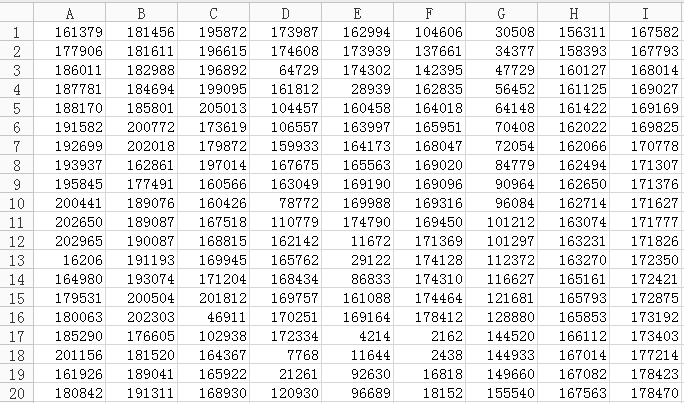


表-17

#### （2）医保欺诈数据的特征信息挖掘求解

为了探讨第52组的输出结果能出现上图三个波峰两个波谷这种特征的原因，我们从重大嫌疑，次要嫌疑和嫌疑较小三类ID号中随机抽取180个ID号，分析其信息特征，并绘制了相关的“盒图”以便于比较。

（MATLAB程序见附录10）

图-13展现了以下信息的统计结果：1-性别，2-年龄，3-买药频率，4-下嘱科室，5-执行科室，6-病人科室，7-医嘱类。

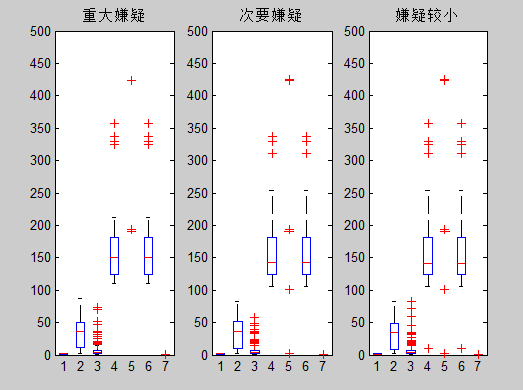


图-13

图-14展现了：1-买药总价，2-买药总数的信息统计结果：

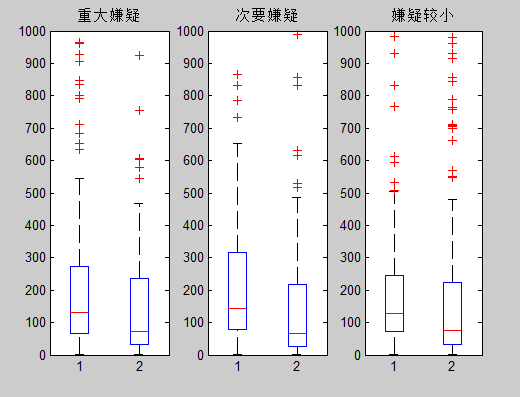


图-14

通过分析图-13，图-14，我们发现：

1. 无论是重大嫌疑，次要嫌疑还是嫌疑较小的ID号，其对应的性别，年龄，买药频率的均值均无较大差异，其均值大小如表-18所示：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 性别 | 年龄 | 买药频率 |
| 重大嫌疑 | 1.488 | 36.14 | 7.53 |
| 次要嫌疑 | 1.498 | 35.02 | 7.97 |
| 嫌疑较小 | 1.450 | 33.08 | 6.48 |

表-18

1. 嫌疑较小的ID号在买药总价，买药总数的数据信息均值与其他两类ID号存在较大的差异；

进一步的，为了了解年龄、买药频率、买药总价、买药总数以及医嘱类在三类ID号中具体的人数分布，我们绘制出了相应的人数分布直方图以便于更进一步的发掘隐含在其中的信息。（统计程序见附录11）

图-15

在年龄的人员分布图中：

1、三类ID号在0~20岁、60~80岁、80~100岁这三个年龄段的人数大致持平；

2、40~60岁年龄段中重大嫌疑的人数明显高于次要嫌疑的人数，而20~40岁年龄段恰恰相反。

图-16

在买药频率的人员分布图中：

1. 低频率阶段（0~20），三类ID号的人数基本持平；
2. 高频率阶段（>20），重大嫌疑的人数高于嫌疑较小的人数。

图-17

在买药总价的人员分布图中：

1. 在低价格区（0~80、80~160）间中，嫌疑较小的人数高于其他两类ID号的人数；
2. 在高价格区间内，重大嫌疑、次要嫌疑两类ID号的人数则高于嫌疑较小的人数。

图-18

在买药总数的人员分布图中：

1. 在低数量（0~400）区间里，三类ID号的人数基本持平；
2. 在高数量区间里，次要嫌疑的人数则明显高于其他两类ID的人数。

图-19

图-20

在性别总数直方图中：

1. 重大嫌疑和次要嫌疑的性别1与性别2人数大致相当；
2. 嫌疑较小类的性别1人数要多于性别2的人数。

图-21

图-22

图-23

在病人/下嘱科室分布直方图中：

1. 重大嫌疑的人数主要集中在：123、124、143、273这四个科室。
2. 次要嫌疑的人数主要集中在：124、203这两个科室。
3. 嫌疑较小的人数主要集中在：106、124这两个科室。

图-24

图-25

图-26

通过执行科室的人员分布直方图，我们发现三类ID号在执行科室191的人数均明显高于其他科室的人数。

综合以上所有的分析，我们得出以下结论：

1. 从各项目的均值大小来看，我们并不能很好的发现重大嫌疑和次要嫌疑的区别。但是通过买药频率和总价这两项，我们可以很好的区别出嫌疑较小类；
2. 从人员分布来看，根据年龄、买药总价和买药总数这三个人员分布直方图中的现象，我们可以发现重大嫌疑和次要嫌疑这两类ID号的区别，具体表现为：
3. 重大嫌疑里买药花费高的人数远多于次要嫌疑的人数。
4. 次要嫌疑中买药数量多的人数远高于重大嫌疑的人数。
5. 重大嫌疑人的年龄普遍高于次要嫌疑人的年龄。
6. 三类ID号的病人科室以124居多，而执行科室以191居多。

最后，为了验证上述ID号分类的可靠性，我们将第52组的识别结果及其相对应的信息导入SPSS进行ward聚类分析，聚类方法如图-27所示：

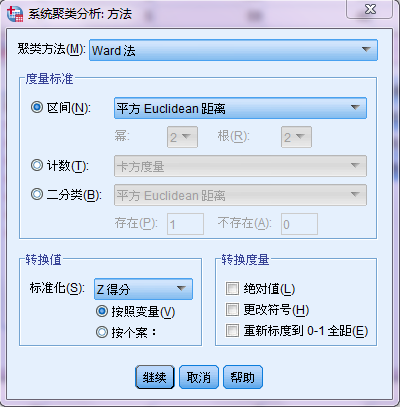


图-27

绘制出聚类树状图如图-28所示：

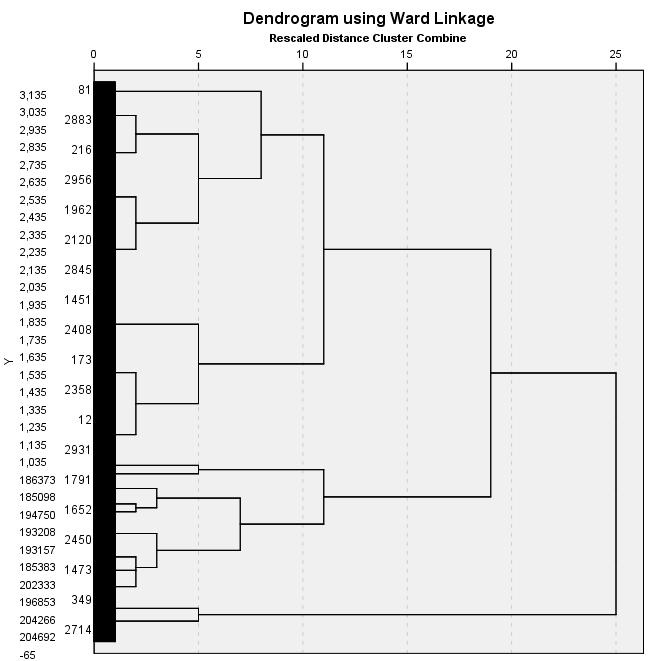


图-28

统计最后的分类结果如图-29所示（分为1，2，3类。限于篇幅，这里只展示部分结果）：

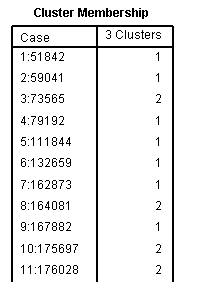


图-29

我们将SPSS聚类所划分的三类数据与我们通过设定概率密度曲线阈值所划定的三类嫌疑ID号进行比对后发现：SPSS聚类所划分的3类中：第一类ID号总数为2181，第二类总数为785，第三类总数为234。这与我们通过概率密度曲线所划定三类ID号的每类个数大致相等。所以，通过概率密度曲线确定阈值，将ID号划分为重大嫌疑，次要嫌疑还是嫌疑较小三类是合理的。

## 七．模型改进

虽然我们已经选取了9个模糊的输入因子，但是我们依旧可以通过数据的回归分析提前剔除一些几乎不影响输出结果的因子，其原理如下表-19所示：

依据模型建立中的假设：

|  |  |
| --- | --- |
| **数据类别** | **期望** |
| Y类数据 | **0** |
| N类数据 | **1** |

表-19

我们以p来表示事件发生的概率，那么事件未发生的概率为 1-p。同时，我们将p看作为自变量x的函数。在本案例中，索赔案件属于欺诈案例的概率为p，为诚信案例的概率为1-P。

对于y是0-1的Bernoulli概型，有如下分布：

通过 Logistic转换，索赔事件的对数概率发生比写成Logit模型：

Logit一方面表达出它是欺诈索赔概率p的转换单位；另一方面，它作为回归的因子变量可以与自变量识别因子之间的依存关系保持传统回归模式：

其中：

。

本文采用IBM SPSS Statistics 19，对已有的输入因子进行Logistic回归分析，以便从中提取出显著因子。

SPSS导入数据如下图-30所示：（限于篇幅，只展示部分数据）

这里以第52组训练样本为例。

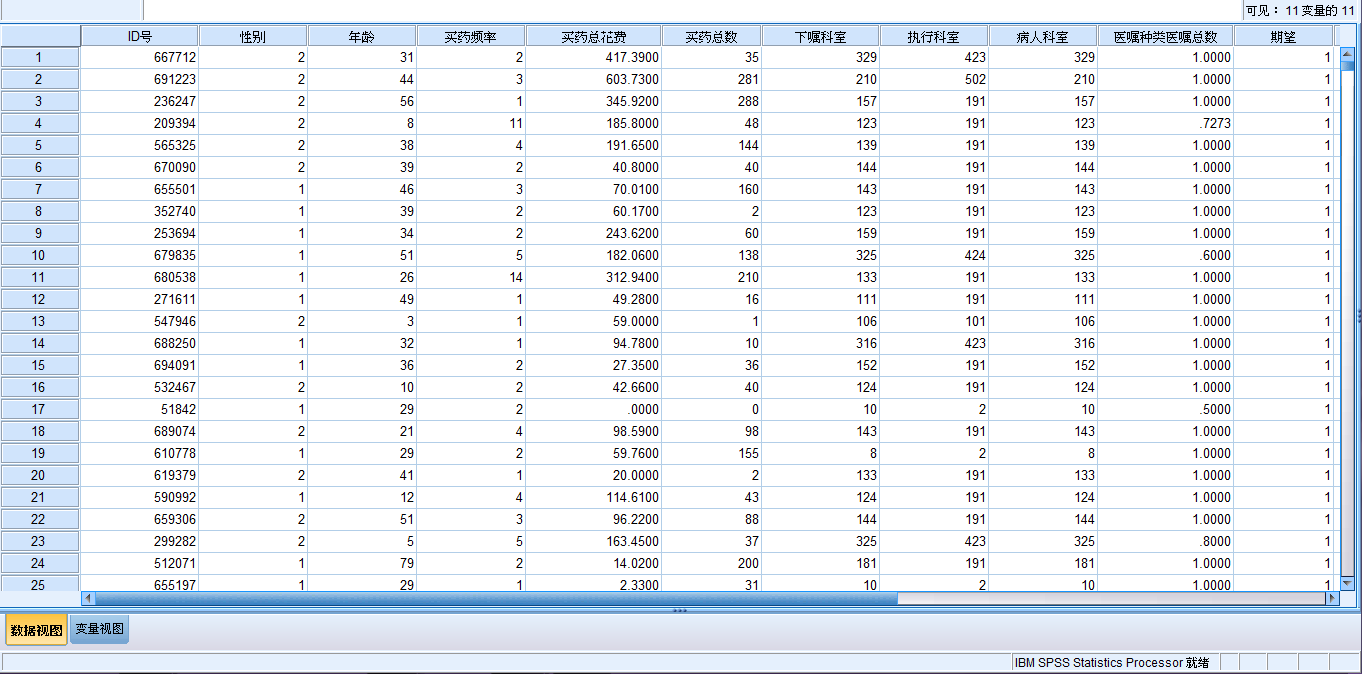


图-30

进行Logistic回归分析的结果如下：

1. 拟合结果：

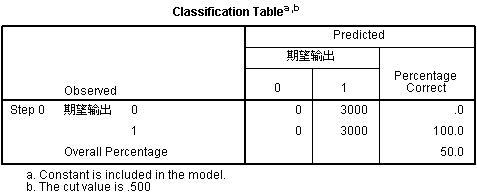


图-31

Block 0拟合的是只含有常数的无效模型，可见总预测的准确率为50%，这是不引入任何解释变量时的预测准确率。

没有引入任何变量时的变量估计值及检测值如图-32所示：

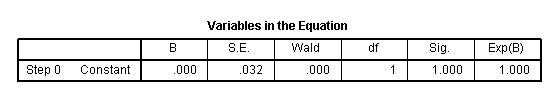


图-32

将变量分别引入回归方程后对方程的显著性影响如图-33所示：

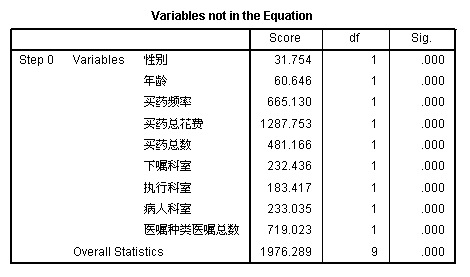


图-33

3、最终拟合的总体水平如下：

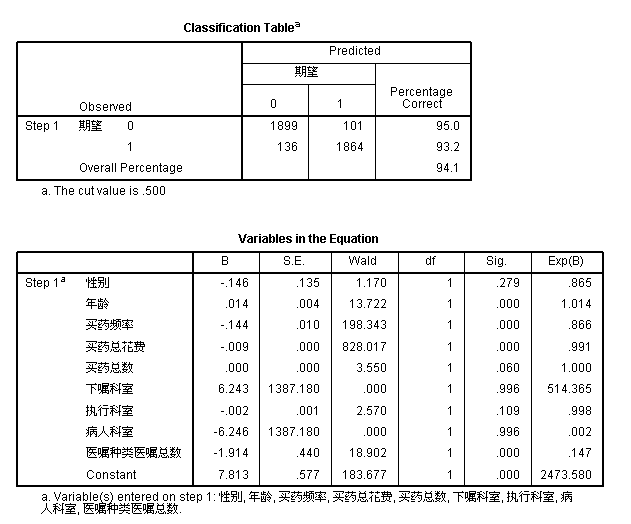


图-34

分析可得，总的预测准确率为94.1%。并且，我们可以将9个输入分为以下三类：

|  |  |
| --- | --- |
| **影响程度** | **输入因子** |
| 几乎不影响 | 下嘱科室，执行科室，病人科室，性别 |
| 有一定影响 | 医嘱类，年龄，买药总数量 |
| 主要影响 | 买药频率，买药总花费 |

表-20

## 八．模型评价

### （1）模型的优点

1.为识别涉嫌医保诈骗ID号所建立起来的BP神经网络，我们采用了LM学习法。它具有较快收敛的效果，以第52组训练过程为例，其仅需要迭代32次便可以得到理想、稳定的识别准确率，均方误差也比较小。

2.对训练数据归一化处理，这样可以使数据的大小均在[0，1]之间，方便后续数据的处理，亦方便训练过程的收敛速度。同时，训练数据的噪声加入（噪声强度：0.01）即可以有效的避免过度拟合现象的出现，又增加了整个数学模型的鲁棒性。

3.通过蒙特卡洛算法随机地对抽取100组样本，对神经网络检测，调整相关参数，使100个样本均满足要求，以验证初步拟定4个人欺诈指示因子和9个模糊因子的可行性，提高识别的准确性。

4. 通过Logistics模型，剔除弱因子对神经网络识别的影响，从而提高结果输出的精确度。

（2）模型的缺点  
1. 没有考虑到所有的影响因子；  
2. 在数据预处理中我们没有考虑四个影响因子之间的内在联系，导致对样本

的分类确定有一定误差；

3.对样本进行随机抽样后再分类，由于样本的局限性，不能完全代表总体，故

存在一定误差；

4. 在BP神经网络的训练过程占用了过多的时间，每训练一个BP神经网络都要耗费8-10秒。

## 参考文献

[1]林 源，国内外医疗保险欺诈研究现状分析[J].保险研究，2010(12）:115-122。

[2]叶明华，基于BP神经网络的健康保险欺诈识别研究[J]，保险研究，2011(3):79-86。

[3]姜启源 等，数学模型[M] 第三版.高等教育出版社，2003:271-289

[4]雷桂媛，[关于蒙特卡罗及拟蒙特卡罗方法的若干研究](http://www.cnki.net/kcms/detail/detail.aspx?filename=2003123425.nh&dbcode=CDFD&dbname=CDFD2003&v=)[D]，浙江大学 2003。

[5]陈明等，神经网络原理与实例精解[M]，北京：清华大学出版社，2013：156-195。

## 附 录

注意事项：

1. 此题程序均在MATLAB 2013a，win7 64位环境下运行；
2. 运行本题程序请严格按照文件读取及输出路径；
3. 为了方便数据之间的比较，现将部分数据做如下调整：

表2.1病人资料的顺序调整：

A、列F,G,H,I,J,K,L分别对应信息为：PAPMI\_ROWID1，PAPMI\_NAME3，PAPMI\_ID，PAPMI\_DVANUMBER，PAPMI\_SAFETYNETCARDNO，PAPMI\_NAME2，PAPMI\_DOB。

表2.2费用明细表的顺序调整：

A、 请按照workload\_papmi\_dr做升序排列

B、列C,D,E,J,K,L,M,N,O,S分别对应信息分别为：WorkLoad\_RecDep\_DR，WorkLoad\_ItemOrd\_DR，WorkLoad\_ResDep\_DR，WorkLoad\_ItemCat\_DR，WorkLoad\_PAADM\_DR，workload\_papmi\_dr，workload\_totalprice，workload\_quantity，WorkLoad\_UnitPrice，WorkLoad\_PatDep\_DR。

附录2说明：

1. 运行时请注意数据读取的路径。

B、运行时的读取文件：新建文本文档.txt已经随程序配套。文件：问题1为本程序的输出文件。

附录5说明：

1. 运行时的读取文件：ID\_danjia.xlsx已经随程序配套。

附录6说明：

A、运行时的读取文件：新的问题样本.xls已经随程序配套。

附录7 B说明：

1. 运行时的读取文件：age.txt已经随程序配套。

附录 11说明：

请注意文件读取的路径；

#### 附录1 根据医保手册号，提取其中号码不为1的相关数据

%% 实现目标：根据医保手册号，提取其中号码不为1的相关数据，并且建立新的数据表格

%% 数据的读取

data=xlsread('C:\Users\Lenovo\Desktop\2015深圳杯A题\2.1病人资料 pa\_patmas.xlsx','A1:BH58018');

[x,~]=size(data)

%% 医保手册号的提取与相关数据的剔除

yibaonum=data(:,7);

j=1;

for i=1:x % 遍历医保手册号

if(yibaonum(i)~=1)

data2(j,:)=data(i,:);

j=j+1;

end

end

%% 数据的输出

xlswrite('C:\Users\Lenovo\Desktop\医保手册号',data2,1);

%% 找出可能的医保卡号造假的，所对应的身份证号

[data1,data2]=textread('C:\Users\Lenovo\Desktop\新建文本文档.txt','%d%s');

[x,~]=size(data1);

q=1;

for i=1:x-1

if(strcmpi(data2(i,1),{''}))

continue;

end

for j=(i+1):x

h=strcmpi(data2(i,1),data2(j,1));

if h==1

pro(q,:)=[data1(i)+1,data1(j)+1];

q=q+1;

h=0;

end

end

end

xlswrite('C:\Users\Lenovo\Desktop\问题1.xls',pro,1);

%% 清屏

clc,clear

%% 可能的问题身份证号位置的读取

data1=xlsread('C:\Users\Lenovo\Desktop\问题1.xls','A1:B1826');

%% 医疗保险号的读取

data2=xlsread('C:\Users\Lenovo\Desktop\2015深圳杯A题\2.1病人资料 pa\_patmas.xlsx','G:G');

%% 相同身份证号所对应的医疗保险号的对比

[x,~]=size(data1);

j=1;

for i=1:x

a=data1(i,1);

b=data1(i,2);

if(data2(a)~=data2(b))

surepro(j,:)=data1(i,:);

j=j+1;

end

end

%% 确定问题的输出

xlswrite('C:\Users\Lenovo\Desktop\ IDtoYLID.xls',surepro,1);

#### 附录2 划定单张处方费用偏高的ID号

data2=xlsread('C:\Users\CharelCHEN\Desktop\ 2015深圳杯A题\2.2费用明细表 dhc\_workload.xlsx.xlsx','O:O');

[f1,x1]=ksdensity(data2,'npoints',10000);

plot(x1,f1)

j=1;

x1(10001)=0;

sum=0;

for i=1:3516

sum=f(j)\*(x(j+1)-x(j))+sum;

j=j+1;

end

fprintf('sum=%f',sum)

%-------------------------------------------------------------------------%

%根据单笔处方费用特别高所划定阈值（0.99464）而得到的1573账单条目，找出其所对应

% 的病人ID号

%-------------------------------------------------------------------------%

%% 清屏

clc

clear

%% 数据导入

% 总ID 号的导入

ID=xlsread('C:\Users\Lenovo\Desktop\2015深圳杯A题\2.1病人资料 pa\_patmas.xlsx','F2:F58018');

% 阈值所划定的ID号读入(里面有重复)；

ID2=xlsread('C:\Users\Lenovo\Desktop\2.2费用明细表 dhc\_workload.xlsx','N2:N1675');

%% 开始进行ID号的提取

[x,~]=size(ID);

j=1;

for i=1:x

curID=ID(i);

place=find(ID2==curID);

if ~isempty(place)

newpro(j,:)=curID;

j=j+1;

end

end

%% 数据的输出

xlswrite('C:\Users\Lenovo\Desktop\单张账单费用偏高（新）.xls',newpro,1);

#### 附录3 划定买药频率过高的ID号

data=xlsread('C:\Users\CharelCHEN\Desktop\ excel数据\6.xlsx','C:C');

[f,x]=ksdensity(data,'npoints',10000);

plot(x,f)

j=1;

x1(10001)=0;

sum1=0;

for i=1:1462

sum1=f1(j)\*(x1(j+1)-x1(j))+sum1;

j=j+1;

end

fprintf('sum1=%f',sum1)

#### 附录4 划定重复配药的ID号

% 病人ID 号的读取

ID=xlsread('C:\Users\Lenovo\Desktop\2015深圳杯A题\2.1病人资料 pa\_patmas.xlsx','F2:F58018');

% 将ID号升序排列的数据导入

ID1=xlsread('C:\Users\Lenovo\Desktop\ID\_danjia.xlsx','A2:A289800');

% ID所对应的价格，数量的导入

dprice=xlsread('C:\Users\Lenovo\Desktop\ID\_danjia.xlsx','B2:B289800');

num2=xlsread('C:\Users\Lenovo\Desktop\ID\_danjia.xlsx','C2:C289800');

%% 根据病人ID号 定位 其药价数据

[x,~]=size(ID);

n=1;

for i=1:x % 遍历病人ID 号

curID=ID(i); % 获取当前ID号

data=find(ID1==curID); % 定位，返回当前ID号所在位置

if isempty(data) % 如果没有消费记录，跳过

continue

end

curIDprice=dprice(data); % 获取当前定位价格

curIDnum=num2(data); % 获取当前定位买药数量

[x1,~]=size(data); % 获取当前ID号所对应账单的个数

if x1==1 % 如果只有一组数据，跳过

continue

end

% 判定

priceexp=curIDprice(1); % 获取一个单价比较样本

numexp=curIDnum(1); % 获取一个数量比较样本

bijiao1=(curIDprice==priceexp); % 开始比较

bijiao2=(curIDnum==numexp);

bijiao1=sum(bijiao1);

bijiao2=sum(bijiao2);

if(bijiao1==x1)&&(bijiao2==x1) % 问题数据的提取

pro(n,1)=curID;

pro(n,2)=priceexp;

pro(n,3)=numexp;

pro(n,4)=x1;

n=n+1;

end

end

%% 问题数据的输出

xlswrite('C:\Users\Lenovo\Desktop\重复配药问题数据.xlsx',pro,1);

#### 附录5 并集运算

%---------------------------------------------------------------------------------------------------------------------------%

% 将新确立的四个问题样本取并集

%----------------------------------------------------------------------------------------------------------------------------%

%% 清屏

clc

clear

%% 数据的导入

% 总ID号的数据导入

ID=xlsread('C:\Users\Lenovo\Desktop\2015深圳杯A题\2.1病人资料 pa\_patmas.xlsx','F2:F58018');

pro1=xlsread('C:\Users\Lenovo\Desktop\重复配药问题数据.xlsx','D2:D19');

pro4=xlsread('C:\Users\Lenovo\Desktop\IDtoYLID.xls','A1:B1636');

pro2=xlsread('C:\Users\Lenovo\Desktop\单张账单费用偏高（新）.xls','A:A');

pro3=xlsread('C:\Users\Lenovo\Desktop\新的问题样本.xls','D2:D821');

%% ID与身份证号唯一识别规则所确定的问题ID

% [x,~]=size(pro4);

% j=1;

% for i=1:x

% curPLACE=pro4(i);

% pro5(j,:)=ID(curPLACE-1);

% j=j+1;

% end

%% 上述问题数据求并集

[x1,~]=size(ID);

j=1;

for i=1:x1 % 开始遍历ID号

curID=ID(i);

place1=find(pro1==curID);

place2=find(pro2==curID);

place3=find(pro3==curID);

place4=find(pro4==curID);

if(~isempty(place1))||(~isempty(place2))||(~isempty(place3))||(~isempty(place4))

pro(j,:)=curID;

j=j+1;

end

end

%% 数据的输出

xlswrite('C:\Users\Lenovo\Desktop\总的问题样本（新）1.xls',pro,1);

#### 附录6 训练样本的确定

##### 程序A 医嘱类输入因子的处理

%---------------------------------------------------------------------------------------------------------------------------%

% 利用医嘱项的数据，计算每一个ID号其医嘱种类与医嘱个数之比

%---------------------------------------------------------------------------------------------------------------------------%

%% 清屏

clc

clear

%% 数据导入

% 待验证的医嘱项——医嘱子类的数据导入

yizhup=xlsread('C:\Users\Lenovo\Desktop\2.2费用明细表 dhc\_workload.xlsx','L2:L289800');

% 对应ID号的导入

ID1=xlsread('C:\Users\Lenovo\Desktop\2.2费用明细表 dhc\_workload.xlsx','N2:N289800');

% 总ID号的导入

ID=xlsread('C:\Users\Lenovo\Desktop\2015深圳杯A题\2.1病人资料 pa\_patmas.xlsx','F2:F58018');

disp('数据导入完毕');

%% 开始验证

[x,~]=size(ID);

j=1;

reality=zeros(x,1);

for i=1:x

curID=ID(i);

if (curID==687972)||(curID==31551)||(curID==580015)

continue;

end

place=find(ID1==curID);

[curnum,~]=size(place); % 医嘱的个数

ID\_yizhu=yizhup(place); % 找出当前ID所对应的所有医嘱

kind=unique(ID\_yizhu); % 医嘱种类

[kindnum,~]=size(kind);

relcof=kindnum/curnum; % 计算相对系数

reality(i,1)=relcof;

end

%% 数据的输出

data=[ID,reality];

xlswrite('C:\Users\Lenovo\Desktop\医嘱种类与医嘱总数的比值.xlsx',data,1);

##### 程序B 信息匹配，并随机抽取100组训练样本

%---------------------------------------------------------------------------------------------------%

% 信息匹配，并随机抽取100组训练样本

%---------------------------------------------------------------------------------------------------%

%% 清屏

clc

clear

%% 数据的导入

% 问题样数据导入

pro=xlsread('C:\Users\Lenovo\Desktop\总的问题样本（新）1.xls','A:A');

% 总ID号的数据导入

ID=xlsread('C:\Users\Lenovo\Desktop\2015深圳杯A题\2.1病人资料 pa\_patmas.xlsx','F2:F58018');

%%总的ID号中剔除问题样本数据

[x,~]=size(ID);

j=1;

for i=1:x

curID=ID(i);

place=find(pro==curID);

if isempty(place)

ID\_safe(j,:)=curID;

j=j+1;

end

end

%% 样本数据相关信息的匹配

%-------------------------------------------------------------------------------------------------------------------------%

% 确定的输入因子有：性别、年龄、买药频率、总消费额、下嘱科室、执行科室

% 病人科室、买药总数、医嘱类

%----------------------------------------------------------------------------------------------------------------------------%

%% 数据的导入

% 性别数据导入

sex=xlsread('C:\Users\Lenovo\Desktop\2015深圳杯A题\2.1病人资料 pa\_patmas.xlsx','O2:O58018');

% 买药频率读取

pinlv=xlsread('C:\Users\Lenovo\Desktop\所有ID号频率数据.xlsx','A1:B58017');

% 买药总价格,总数量数据导入

sumprice=xlsread('C:\Users\Lenovo\Desktop\ID对应账单的总价格,总数目.xlsx','A1:C58017');

% 年龄数据的导入

% 病人ID692316出生不祥，剔除

% 病人ID669432出生于1875-01-12，与实事严重不符！剔除！

age=textread('C:\Users\Lenovo\Desktop\age.txt');

age(:,2)=2015-age(:,2); % 转化为岁数

% 从费用明细表中读取下嘱科室、执行科室、病人科室的数据

ID2=xlsread('C:\Users\Lenovo\Desktop\2015深圳杯A题\2.2费用明细表 dhc\_workload.xlsx','L2:L289800');

xiazu=xlsread('C:\Users\Lenovo\Desktop\2015深圳杯A题\2.2费用明细表 dhc\_workload.xlsx','S2:S289800');

zhix=xlsread('C:\Users\Lenovo\Desktop\2015深圳杯A题\2.2费用明细表 dhc\_workload.xlsx','C2:C289800');

binr=xlsread('C:\Users\Lenovo\Desktop\2015深圳杯A题\2.2费用明细表 dhc\_workload.xlsx','E2:E289800');

% 医嘱种类与医嘱总数比值的数据导入

yizhu\_yzzl=xlsread('C:\Users\Lenovo\Desktop\医嘱种类与医嘱总数的比值.xlsx','A1:B58017');

disp('匹配数据已经全部成功导入');

%% 开始信息的匹配

for jishu=1:101

%% 从已经导入的数据中随机选取匹配2000个样本

fprintf('-------第 %d 次随机抽取-------\n',jishu);

[x1,~]=size(ID\_safe);

a=randperm(x1); % 产生随机数列

b=a(1:2000);b=b'; % 取转置

examp\_safe=ID\_safe(b);

% 因为统计年龄时剔除了692316和669432，判断这两个ID号在否在选定

% 的诚信数据内

% 687972 31551 580015 购买平率为0 不考虑

abandon1=692316;

abandon2=669432;

abandon3=687972;

abandon4=31551;

abandon5=580015;

p1=find(examp\_safe==abandon1);

p2=find(examp\_safe==abandon2);

p3=find(examp\_safe==abandon3);

p4=find(examp\_safe==abandon4);

p5=find(examp\_safe==abandon5);

if ~isempty(p1)

examp\_safe(p1)=ID\_safe(a(2001));

end

if ~isempty(p2)

examp\_safe(p2)=ID\_safe(a(2002));

end

if ~isempty(p3)

examp\_safe(p3)=ID\_safe(a(2003));

end

if ~isempty(p4)

examp\_safe(p4)=ID\_safe(a(2004));

end

if ~isempty(p5)

examp\_safe(p5)=ID\_safe(a(2005));

end

disp('完成诚信数据的选取');

[x2,~]=size(pro);

c=randperm(x2); d=c(1:2000);d=d';

pro=pro(d);

example=[examp\_safe;pro]; % 样本数据已经选定

if (size(example)~=4000)

error('警告！样本数据没有正常选取');

end

disp('样本数据正常选取！');

%% 开始匹配

[x,~]=size(example);

pinlv\_exm=zeros(x,1); % 矩阵初始化

sex\_exm=zeros(x,1);

price\_exm=zeros(x,1);

exm\_age=zeros(x,1);

num\_exm=zeros(x,1);

exm\_xiazu=zeros(x,1);

exm\_zhix=zeros(x,1);

exm\_binr=zeros(x,1);

exm\_yizhu=zeros(x,1);

j=1;

for i=1:x % 开始遍历，匹配相应的数据

curID=example(i);

place=find(ID==curID);

place1=find(pinlv(:,1)==curID);

place2=find(sumprice(:,1)==curID);

place3=find(age(:,1)==curID);

place4=find(ID2==curID);

place5=find(yizhu\_yzzl(:,1)==curID);

pinlv\_exm(j,:)=pinlv(place1,2); % 定位匹配

sex\_exm(j,:)=sex(place);

price\_exm(j,:)=sumprice(place2,2);

exm\_age(j,:)=age(place3,2);

num\_exm(j,:)=sumprice(place2,3);

exm\_xiazu(j,:)=xiazu(place4(1));

exm\_zhix(j,:)=zhix(place4(1));

exm\_binr(j,:)=binr(place4(1));

exm\_yizhu(j,:)=yizhu\_yzzl(place5,2);

j=j+1;

end

%% 数据的输出

output1=zeros(2000,1); output2=ones(2000,1);

output=[output2;output1]; % 定义输出：诚信索赔为1，欺诈索赔为0

yangben=[example,sex\_exm,exm\_age,pinlv\_exm,price\_exm,num\_exm,exm\_xiazu,exm\_zhix,exm\_binr,exm\_yizhu,output];

filename=strcat('C:\Users\Lenovo\Desktop\100组数据\神经网络最终样本数据',num2str(jishu),'.xlsx');

xlswrite(filename,yangben,1);

end

#### 附录7 BP神经网络的医保诈骗识别

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% %

% 基于BP神经网络的医保诈骗识别 %

% (采用加入噪声法) %

% %

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%% 清屏

clc

clear all

close all

%% 相关常量数据约定

train\_num=6000; % 训练样本

noisevar=0.001; % 噪声强度

learntate=0.01; % 神经网络学习率

train\_goal=0.001; % 神经网络训练目标

%% 相关数据导入与合并

% ID号的数据导入

ID=xlsread('C:\Users\Lenovo\Desktop\100组数据\神经网络最终样本数据2.xlsx','A1:A4000');

% 训练数据的导入

train\_data=xlsread('C:\Users\Lenovo\Desktop\100组数据\神经网络最终样本数据2.xlsx','B1:J4000');

% 期望数据的导入

except\_data=xlsread('C:\Users\Lenovo\Desktop\100组数据\神经网络最终样本数据2.xlsx','K1:K4000');

%% 相关数据的初步处理

train\_data=train\_data'; % 确定输入层有9个节点，共计4000个样本数据

except\_data=except\_data';

%% 阈值的加入

[x,y]=size(train\_data);

% 训练数据插入阈值

train\_b=ones(1,y);

train\_data(x+1,:)=train\_b;

%% 数据归一化

% 训练数据归一化

% mm=mean(train\_data,2);

% % 均值平移

% for i=1:9

% train\_data\_s(i,:)=train\_data(1,:)-mm(i);

% end

% % 方差标准化

% for i=1:9

% ml(i)=std(train\_data\_s(i,:));

% train\_data\_s(i,:)=train\_data\_s(i,:)/ml(i);

% end

[traind\_data,traindmin,traindmax]=premnmx(train\_data);

% 用于神经网络检测检测数据的确定

test\_data=traind\_data;

%% 输入数据加入噪声

train\_import\_s=traind\_data+randn(10,4000)\*noisevar;

%% BP神经网络的创建（基于MATLAB BP神经网络工具箱）

fx=[];

net=newff(fx,[],[9,10,2],{'logsig','logsig','logsig'},'trainlm');

%net.trainFcn='purelin'; % BP神经网络学习方式的选择

view(net);

%% 神经网络的训练

% 训练输入样本 train

% 训练输出样本 train\_outport

net.trainParam.epochs = 1000; % 训练次数

net.trainParam.goal =train\_goal; % 训练目标

net.trainParam.Lr=learntate; % 学习率

[net,tr]=train(net,train\_import\_s,except\_data);

%% BP神经网络的测试阶段

test\_r\_outport=sim(net,test\_data); % 获取识别数据

%% 预测，绘图准备阶段

% 随机抽取100点用于绘图

a=randperm(4000);

a=a(1:100);

i=1:100;

plot(i,except\_data(a),'o');

hold on

for j=1:4000

if test\_r\_outport(j)<0.5

test\_r\_outport(j)=0;

elseif test\_r\_outport(j)>0.5

test\_r\_outport(j)=1;

end

end

plot(i,test\_r\_outport(a),'\*');

legend('期望输出','实际输出');

rate=sum(test\_r\_outport(a)==except\_data(a))/100

#### 附录8 对100组数据分别建立神经网络，并对总体进行识别

%------------------------------------------------------------------------------------------------------------------------%

% 对于已经随机抽取的100组训练样本，分别建立

% 相应的神经网络，并对全部数据进行识别

% 得到最终识别的结果

%-------------------------------------------------------------------------------------------------------------------------%

%% 清屏

clc

clear

%% 相关常量数据约定

train\_num=4000; % 训练样本

noisevar=0.001; % 噪声强度

learntate=0.01; % 神经网络学习率

train\_goal=0.001; % 神经网络训练目标

rate=zeros(1,100); % 准确率矩阵初始化

%% 开始循环读取数据，进行神经网络的训练

for i=1:100

%% 数据的读取

fprintf('------------ 正在对第%d组数据进行网络训练------------\n',i);

filename=strcat('C:\Users\Lenovo\Desktop\100组数据\神经网络最终样本数据',num2str(i+1),'.xlsx');

% 训练数据的导入

train\_data=xlsread(filename,'B1:J4000');

% 期望数据的导入

except\_data=xlsread(filename,'K1:K4000');

disp('数据读取完毕！');

%% 相关数据的初步处理

train\_data=train\_data'; % 确定输入层有9个节点，共计4000个样本数据

except\_data=except\_data';

%% 阈值的加入

[x,y]=size(train\_data); % 训练数据插入阈值

train\_b=ones(1,y);

train\_data(x+1,:)=train\_b;

%% 数据归一化

[traind\_data,traindmin,traindmax]=premnmx(train\_data);

% 用于神经网络检测检测数据的确定

test\_data=traind\_data;

%% 输入数据加入噪声

train\_import\_s=traind\_data+randn(10,4000)\*noisevar;

%% BP神经网络的创建（基于MATLAB BP神经网络工具箱）

fx=[];

net=newff(fx,[],[9,10,2],{'logsig','logsig','logsig'},'trainlm');

% BP神经网络学习方式的选择

disp('神经网络成功建立！开始进行训练');

%% 神经网络的训练

net.trainParam.epochs = 1000; % 训练次数

net.trainParam.goal =train\_goal; % 训练目标

net.trainParam.Lr=learntate; % 学习率

[net,tr]=train(net,train\_import\_s,except\_data);

%% BP神经网络的测试阶段

test\_r\_outport=sim(net,test\_data); % 获取识别数据

%% 神经网络识别准确性的存储

% 随机抽取100组数据

a=randperm(4000);

a=a(1:100);

for j=1:4000

if test\_r\_outport(j)<0.5

test\_r\_outport(j)=0;

elseif test\_r\_outport(j)>0.5

test\_r\_outport(j)=1;

end

end

rate(i)=sum(test\_r\_outport(a)==except\_data(a))/100;

%% 开始对所有数据进行识别

disp('开始识别总体数据！');

% 需要识别的数据导入

% ID号的数据导入

ID=xlsread('C:\Users\Lenovo\Desktop\2015深圳杯A题\神经网络全部样本数据.xlsx','A2:A58013');

% 训练数据的导入

shibie\_data=xlsread('C:\Users\Lenovo\Desktop\大数据\神经网络最终样本数据.xlsx','B2:J58013');

shibie\_data=shibie\_data';

% 识别数据归一化

[shibie\_data,shibiemin,shibiemax]=premnmx(shibie\_data);

[~,y]=size(shibie\_data);

b=ones(1,y);

shibie\_data(10,:)=b;

shibie=sim(net,shibie\_data);

[~,y]=size(shibie);

for j=1:y

if shibie(j)<0.5

shibie(j)=0;

elseif shibie(j)>0.5

shibie(j)=1;

end

end

place=find(shibie==0);

output=ID(place);

filename1=strcat('C:\Users\Lenovo\Desktop\100组问题数据\问题数据',num2str(i+1),'.xlsx');

xlswrite(filename1,output,1);

end

#### 附录9 重大嫌疑、次要嫌疑、嫌疑较小的ID号

**重大嫌疑的ID号**

51842 179926 176635 197687 200647 190385 180353 194210 196891 197959

59041 179946 176673 197826 201168 190910 181091 194583 197107 197995

73565 179975 176729 197968 201852 191473 181307 194679 197347 198172

79192 181011 177114 198643 202256 191525 181509 194736 197533 198718

111844 181092 177355 198654 202992 192566 181609 195045 197863 199142

132659 181262 177826 200063 203043 192950 181681 195640 198354 199381

162873 181303 177833 200149 203320 193418 182185 195895 199125 200102

164081 181497 178314 200502 203560 193592 182328 196528 199779 200184

167882 183008 178802 200753 203975 193959 182465 196643 199948 200489

175697 183218 179030 200877 204464 194301 183111 196720 200032 200935

176028 183362 179061 200957 204703 194465 183264 196853 201507 202333

176101 183725 179079 201478 204861 194468 186600 196999 202336 203153

176334 184321 179315 201663 205042 194846 187356 197180 202787 203317

176500 184362 179342 201846 101390 195812 189491 198355 202820 203438

176716 184713 179540 201898 166449 196080 189546 198610 202976 203487

176798 185232 179600 202221 174413 196222 189758 198896 203207 203636

176802 185413 179632 202284 175874 196284 189798 199054 204145 204209

177856 185788 179971 202884 175876 196640 189879 199240 204420 205319

178246 186310 180502 203219 176122 197520 190101 200074 175484 175486

178506 186572 180677 204041 176775 197922 191306 200366 175685 175546

179451 187291 180983 204043 177857 197938 191829 200546 176114 175580

179454 187596 181691 204182 178281 198362 191853 200886 176319 175602

179633 187647 181788 204275 178337 198722 193135 201118 176821 175608

180339 187792 182063 204934 178446 199083 193885 201345 177088 176200

181387 187795 182231 205017 179719 199929 194776 201407 177101 176210

181817 188263 182496 120108 179723 201137 196620 201487 177861 176585

184046 188518 182794 161823 181989 201154 197039 201542 178365 176740

184282 188811 182958 164340 182487 202510 197281 201825 178709 177002

185207 188872 183006 166415 182597 203095 197312 201877 179692 177022

185428 188946 183468 170283 183637 203407 197531 202127 179694 178798

185463 189057 183553 172257 185160 203682 198305 202392 179709 179546

185469 189340 183700 172325 185585 203747 198452 204198 180100 179610

185783 189931 184127 175613 185921 204194 199652 204456 180108 179791

186291 190327 184311 175657 186561 204479 199952 204495 180126 180001

187214 190773 184764 175679 188187 204638 200618 205182 180274 180125

188322 190865 184823 175924 188226 205037 200795 8471 180866 181085

188685 191665 185864 176331 188522 159773 200889 56617 180888 181189

188723 191824 186099 176382 189384 160311 201107 175428 181389 181289

189381 191854 186142 176958 189411 167084 201956 175521 181464 181786

189591 191882 186264 176963 190363 173990 204481 175560 181498 182469

190000 192083 186339 178266 192104 175879 205064 175630 182146 182560

190031 192191 187348 179519 192254 176055 162463 175635 182206 182721

190233 192239 187683 179789 192725 176094 175373 175839 183120 182792

190409 192315 187721 180011 192848 176480 175810 175842 183354 182795

191153 193291 188115 180407 193155 176949 176057 176069 183650 182806

191172 194181 188325 180513 194332 177064 176316 176401 183665 182814

191791 194355 188570 181115 194499 177488 176464 176762 183683 183042

191811 194750 188849 181688 194510 179112 176520 176885 184091 183071

191960 195013 189033 182717 194584 179994 176627 176887 184144 183085

192076 195221 189249 182728 194748 181950 176728 176919 184284 183108

192234 195232 189354 183070 196594 181962 177851 176936 184799 183173

192306 195840 189430 183127 196661 183038 179128 176980 185236 183182

194290 196061 189460 183427 197628 184604 180012 176994 185354 183350

195348 196265 189503 184580 198117 185038 180595 178216 185362 184085

195387 196707 189587 185890 198414 185466 180790 178235 185441 184117

195471 196740 189596 186707 198911 186824 180948 178731 185673 184180

195767 198782 189660 187734 199364 187409 181196 180235 185766 184493

196220 198962 189778 187754 199753 187993 181537 180531 185900 184791

196714 199106 189960 188157 199799 188664 182495 181107 185982 185584

197355 199557 190170 188347 199820 188950 182625 181425 186184 185888

198238 199687 190360 188422 200072 188993 183093 181682 186250 186133

198365 200187 190553 188485 200101 189850 183109 181815 186721 186179

198605 200227 190571 188704 200162 191054 183231 181841 187266 186735

198977 200621 190637 188743 202393 191163 184066 181893 187290 186741

198981 200750 190737 188854 202785 191216 184752 182198 187983 187045

199373 200933 191094 188923 202824 191537 185016 182769 188566 187074

199669 201240 191227 189140 203205 192042 185258 183194 188609 187201

199717 201384 191287 189963 204837 192051 185371 183331 188709 187255

199763 201668 191446 190951 205155 192170 185563 183384 188747 187592

200154 201814 191778 191190 208988 192614 185928 183853 188787 187708

202211 202157 191987 191260 138397 193676 186354 184157 189268 188801

202212 202316 191989 191403 149168 194280 186664 185771 189320 189107

202454 203084 192613 191551 175780 194483 186733 185840 189542 189601

203068 203144 192778 192041 175995 195672 186802 185877 189577 189620

203732 203151 192861 192570 176054 195787 187024 187277 189853 190051

204562 203614 193107 192750 176350 196364 187562 187349 189976 190159

204755 203932 193448 193153 176810 196383 187599 187706 190180 190352

205154 204251 193488 194038 177815 196634 187655 188266 190755 190364

208875 204274 194131 194101 179060 198248 187710 188699 190803 190412

41089 204445 194161 194347 180751 199461 188813 188941 191488 190442

59238 204639 194319 194454 181210 199634 189162 189225 192118 190644

160709 204892 194448 194479 181302 200488 189409 190593 192309 190699

167138 204995 194635 195096 181544 200601 190117 190812 192341 190701

169824 205226 194650 195745 183243 202808 190215 192080 192375 191379

171315 28167 194713 195759 184351 203078 190680 192754 192612 191436

175268 76404 194831 195862 184569 203377 190730 192961 192838 191776

175284 107238 194845 197537 185750 204986 190735 193933 193251 191942

175440 112118 194950 197824 186106 205092 190994 194177 193818 191965

175749 160878 194988 198043 186208 131842 191105 194266 194464 192345

175891 171738 195026 198086 186831 161207 191160 194694 194758 192595

175899 175342 195776 198091 186875 165583 191291 194875 195059 192678

175956 175385 196111 198336 186902 173688 191921 194976 195436 192870

176313 175775 196160 198418 187087 175219 192142 195149 195739 192878

176395 175895 196216 198520 187398 176557 192369 195158 195965 193061

176484 175926 196226 198672 187695 176710 192960 195215 196228 193230

176886 175970 196814 198741 188006 177062 193062 195353 196649 193485

177816 176027 196833 198845 188031 177286 193160 195382 196933 193701

178343 176267 197012 199169 188860 178685 193493 195656 197106 193962

179553 176394 197135 199834 189393 179236 194089 195985 197426 193993

179675 176553 197658 200091 189555 179992 194171 196293 197603 194458

194470 191817 188248 187705 189609 195394 201586 198348 187072 194167

194486 192155 188262 187750 189713 196030 159793 198681 187701 197128

194680 192343 188278 188107 189983 196233 177869 198804 189158 198191

195038 192396 188293 188255 190019 197172 179157 199514 189985 199273

195153 192603 188349 188394 190176 197182 179879 199762 190055 199978

195295 192791 188761 188713 190881 197262 180361 201634 193287 200298

195309 193219 188825 188794 191349 197747 180567 202512 201584 201422

195867 193259 189586 188943 191373 198036 181168 202985 202076 201850

196031 194120 189616 189034 191974 198284 181265 202996 203460 203022

196037 194262 189852 189101 192099 198460 181904 203162 176296 128715

196200 194312 190287 189215 192522 198463 182170 203363 179121 176628

196321 194378 190299 190513 192938 198995 184645 203637 179408 178804

197025 194428 190700 190604 193494 199497 188319 205004 179583 179949

197031 194555 190869 191132 193852 199603 188939 205250 180064 181868

197333 195598 190914 191178 194480 199639 190261 60553 180460 183414

197749 195733 191487 191276 195182 199927 194327 176149 181022 185013

197839 196442 191647 191359 195230 201674 194473 176162 185459 186137

197900 196463 191673 191458 195323 201787 194632 176253 186231 188449

198075 196560 191850 193079 195522 201915 195882 177345 187038 188471

198334 197060 192014 193286 196408 202898 195909 179002 188766 189676

198546 197412 192364 193345 196729 203037 196452 179873 189545 190324

198571 197599 192390 193442 197528 203692 199531 181630 190289 191727

198686 197978 192988 193692 197931 204555 199691 181806 190733 192658

198821 198063 193404 194023 198369 204827 202746 181856 192087 198417

198946 198230 193414 194173 198702 204980 203105 182066 194179 199727

199007 198260 193532 194735 199111 175796 204302 183408 194638 204985

199286 198435 193589 194839 200140 176540 204353 186138 195525 180092

199789 199858 193627 194991 200276 177811 204544 187224 196143 180104

200158 199941 194069 196092 200689 179964 208857 187573 196176 182055

200352 201620 194188 196195 201235 180743 175785 190530 197043 185216

200609 201925 194641 196238 201259 181438 176322 191179 197332 187117

201057 201968 194745 196562 201732 182210 176323 191376 199714 187805

201185 202623 195131 196610 201942 183437 178743 192010 199782 190912

201353 202893 195223 196625 202318 184657 181030 192646 202323 192446

201390 203142 195284 196700 202354 186919 182003 193749 202602 193145

201418 203255 195389 196722 202483 187222 182474 194032 203228 193375

201510 203381 195518 196738 202495 188017 182882 194366 203668 193680

201651 204423 195553 196749 202693 188311 183343 196022 204109 196163

202378 204492 196178 197453 202888 189073 183415 197702 204484 197594

202886 172930 196304 198431 203024 189694 184725 198708 204970 198084

202903 175401 196352 198526 203116 190355 184926 199365 28033 199201

202939 175708 196367 198892 203695 190356 184971 199960 175882 200161

202951 175765 197212 199104 204496 191244 185385 201826 176820 201053

203590 176083 197213 199542 204530 193497 185609 202629 178251 201598

204079 176329 197247 199716 207573 193685 186140 207576 179263 203547

204863 176950 197276 200698 208425 195712 187048 179149 179884 177855

208882 176975 197877 201417 175516 195727 188718 179968 181335 181631

175264 177285 197966 201691 175597 195808 191980 181042 181442 184148

175665 177783 198525 202042 175820 196311 192158 181354 181732 186946

175883 179033 198822 202129 176999 196956 193078 182276 182132 193778

176092 179204 199017 202989 177091 197331 193829 182301 182736 194235

176160 179249 199313 203165 177883 198061 194466 182399 182823 195852

176379 179303 199396 203705 179126 198611 195112 183137 183232 197918

176406 179485 199999 204756 179209 198770 196148 184423 184124 199832

176567 179634 200483 205006 179285 199076 196495 186532 184256 201557

176597 179676 200971 205115 179524 199558 196644 187852 185098 175292

177121 179812 201432 205313 179835 200261 196653 188241 186955 175622

177374 179981 201859 158458 179959 200422 198996 189423 189440 175864

178357 180010 202295 175682 180702 200905 199488 191938 190341 179550

178773 180301 202319 175938 180957 202482 199562 192050 191262 179662

179246 180318 202542 176373 181080 203711 200696 194108 191694 181645

179843 180730 203679 176591 181248 204266 201481 196223 191934 185453

179854 180859 203786 176834 181891 205172 202095 199471 193186 186322

179902 180891 204156 177115 182153 175526 203101 204344 193400 190472

180000 180946 204252 177887 182172 176097 203127 204436 193883 191421

180997 181007 204466 179174 182551 176310 203787 175655 195027 191800

181416 181426 204968 179269 183157 176433 205059 177889 195668 192093

181557 181624 205010 179420 183281 176648 164465 178672 197389 192174

181747 181671 175482 179575 183526 176876 176516 180077 197574 192360

181967 181708 175674 179781 183530 180546 176638 181099 197646 192655

182274 181980 176077 179797 184108 181504 179039 182900 197773 192660

182410 182335 176498 179956 184336 181580 179182 184308 197882 193547

182828 182483 176622 179958 184488 181860 179836 184715 198353 194098

183150 182486 176899 179993 185161 183249 180890 184817 198374 194471

183272 182513 177099 179997 187200 183451 182950 185596 198717 194569

183428 182963 177301 180249 187373 183513 183113 185637 199731 194777

183492 183141 177366 180674 187662 183595 184058 187469 200861 194983

183773 183244 179413 180766 188071 183605 184315 189080 201433 195189

184294 183348 179701 180878 188237 184189 184338 189923 202022 197019

184498 183543 179735 182001 188300 184647 185519 190195 202894 197275

185007 183550 180049 182207 188597 185029 186704 190896 203136 197496

185518 183998 180784 182746 188605 185576 187364 191946 203307 198730

186051 184006 180841 183095 188800 186389 187597 192643 203676 199816

186217 184035 180929 183330 188882 186457 188110 192865 204778 200174

186558 184749 181588 184392 189172 186593 188737 193674 205003 201838

186623 184806 181968 184418 189658 186881 189060 195238 205029 204582

186786 184979 182407 185477 190149 186948 189703 197441 176441 205125

187112 185748 182416 185488 190336 188126 191526 199203 178990 205304

188457 185769 182463 185671 190681 188621 192154 199351 181046 179155

189360 185990 183174 185855 190816 189376 193487 199515 181252 180413

189468 186273 183597 185918 191542 190196 193791 200715 183268 180668

190040 186350 183614 185927 191762 192173 194232 201263 183579 181684

190042 186415 184133 186428 192753 193197 194457 201628 184413 182978

190086 186495 184258 187144 193322 195521 194482 202905 185080 187128

190148 186535 185442 187543 193534 196423 195058 203836 185881 188040

190523 186821 186085 187642 194284 198288 195245 204692 187477 189180

191047 186871 186603 188147 194307 198591 195913 204701 188430 189482

191229 187105 187367 188622 194423 199386 197723 176058 188823 190437

191338 187212 187378 189339 194770 201274 197795 176061 190477 191663

191365 187285 187648 189404 195034 201470 197913 179764 192084 195012

195822 177787 192449 202156 192308 204584 201221 199413 194453 190331

196134 180213 194561 204691 196121 110045 180176 202775 195999 192931

196378 181311 195695 169626 198115 181648 181485 204351 203129 193138

196943 182713 195964 181106 199776 187235 183473 175797 203244 195556

197174 183917 197367 185489 200166 187607 186537 176050 204581 199056

197374 184961 198391 187649 202168 192564 186801 183259 204883 200211

197411 188631 199065 188623 202641 196692 192571 183483 8744 202518

201593 188786 199499 191096 203263 198111 193018 185642 185282 169431

202190 188834 200678 191457 204078 199842 197529 186637 185794 176765

170147 190350 201844 192240 204512 200978 198151 187663 186373 179057

182664 184369 190198 192064 192998 197972 201420 204202 204669

**次要嫌疑ID号**

179573 183021 185779 189955 193086 205217 175250 176370 179991

181530 184505 186866 191501 193321 163536 175370 176487 180694

184524 185528 187236 191896 201532 175732 175946 177859 180990

185011 186408 188389 192061 201897 179489 176954 180579 183144

185541 186701 190245 192074 204021 179827 179048 181162 183164

188769 188014 190386 192282 205137 180911 179910 182156 187410

188874 188401 191002 192630 158904 181208 181653 183226 188250

189594 188862 192048 192698 176609 182326 182903 183442 188637

194133 196076 194558 192857 179090 183782 185558 184191 188981

194606 196970 195236 192921 180033 184243 187049 184675 190113

199442 199567 199097 192997 184995 184993 188314 185703 192224

202031 199976 203612 193208 187999 185938 188473 187532 193327

202442 200021 175223 193302 190528 186873 188779 187550 194317

176165 200709 175683 193672 191723 187084 190163 187604 196087

176390 161050 179260 194211 192872 187362 190269 189971 197804

178679 175633 180134 194827 194854 188497 190362 190426 198128

181144 176136 180871 196435 195419 188591 190871 191408 202740

187186 176481 181467 198028 199311 188697 191851 191503 203208

193720 177082 182102 198324 201860 188821 192264 193886 203331

198705 178332 182404 199372 201865 191138 192283 194590 203754

203620 179019 185394 199374 202352 191255 192927 198121 204110

174420 182321 186769 200191 202558 193616 194065 198570 175617

179945 182831 189370 200611 204714 193702 195033 199618 182211

188213 183112 190147 201550 205264 193895 196459 200213 184084

189593 183349 190939 202365 175237 195241 197120 201147 187150

193462 183652 191920 202462 175248 195357 199591 203641 189371

194579 185627 194441 202981 175310 195858 199956 204684 191878

197210 188033 194572 203663 175349 197346 201784 176391 192426

197466 190371 194876 204163 175416 198361 204656 177314 196177

201161 192585 194913 204212 175807 199920 204661 178722 196965

202416 196073 197285 204430 175911 202030 204783 179171 197838

204593 198291 199253 204866 176062 202549 204832 179328 203675

205100 199856 201233 176135 176293 203892 160905 179505 204147

161724 202191 202244 179341 176384 163918 175863 180019 204800

177841 204118 203030 181605 177049 175755 178707 180343 175646

179346 176522 203550 182289 178221 176521 179457 182285 178363

180887 182984 205318 182768 179023 177396 181132 183458 178698

180996 184234 176045 185390 179099 177458 182336 184910 178832

183644 184983 179078 187941 179541 180292 183303 185184 179572

183978 185023 179488 192029 179851 180697 184770 185237 182141

186581 185785 179563 193785 180298 186218 185436 187184 182921

186635 185825 179800 194539 180454 188710 185732 187736 186726

187081 187779 180185 194727 180953 193686 186610 187966 190007

188544 188393 180592 195199 181103 194925 187009 188335 192302

191881 190872 181285 195376 182786 195746 187345 188581 195075

198806 191176 181861 195711 182840 196441 188388 190041 196004

201847 191323 182809 196136 182883 196812 189176 190334 198027

204487 193331 182895 196911 183011 197906 189352 190338 203021

205086 193884 187025 197206 183015 199744 189691 190629 203760

177792 195206 189377 197390 183480 200122 190078 190698 204336

180145 199317 189602 197437 184007 200529 194237 192851 164271

183571 201702 190346 198116 184179 200533 197417 194207 175299

183670 201926 191508 199016 184197 201948 197652 195014 175735

187484 175872 191848 199524 184398 202774 197974 195646 180717

188036 176074 191874 199702 184417 203303 199653 195857 181211

193394 176467 192077 199918 185420 205087 200559 196006 181529

195718 181699 192508 200755 187431 177837 202464 196307 184192

197967 182390 194424 200765 188261 178047 202473 196476 185342

198412 184592 194537 201275 188441 179470 202590 197272 190832

201939 186251 195661 201493 188959 180708 202735 198001 190837

204405 186686 196894 202470 189228 181488 203073 199752 190985

205007 187825 197761 202565 189349 181735 203688 199955 192568

207572 196251 199152 170191 189494 182383 203845 200023 192689

101135 198574 201034 175828 189944 182512 160623 200523 193649

182534 200587 203221 176524 190557 187715 175702 200940 195447

185689 202600 173435 177421 190780 187938 176219 201133 197156

185996 203099 175819 177779 191633 193730 176469 175880 198379

186345 203973 175913 177914 191787 194118 176946 176264 200590

186351 173114 176034 179391 193157 195650 179043 178231 201978

192863 175287 176129 179911 193268 196091 179216 179574 204365

195061 176304 176202 181244 193681 196093 179665 180073 205145

196784 179077 176435 187296 194006 196686 179710 180190 178775

198011 179887 176514 188020 194041 199289 179840 180608 183537

204891 180072 176961 189347 194337 200931 179855 181350 184059

176639 184700 177124 190139 194359 202119 180808 182406 186120

177086 185611 178803 190353 194615 204085 180864 182499 186344

181288 185901 179706 190443 194686 204475 182880 183134 187102

183258 186442 180835 191491 195699 204493 184464 183271 191861

187950 186832 181491 194329 196016 176223 184703 186060 192413

191086 187632 181533 195673 196109 178258 185603 186599 194792

192491 188015 182028 196005 196171 178721 187079 186660 194889

192985 189947 182220 198783 196389 179200 187352 187699 195958

196545 190469 182881 198819 196656 182423 188069 189626 196410

196947 190484 183087 199982 197672 184699 189302 190944 200312

197217 191402 183136 200069 198866 185040 190190 192438 201610

197679 192106 183682 200903 199359 188424 190262 192493 202102

198338 193766 184406 176962 200633 188846 190387 192758 204814

198884 197058 184767 180161 200635 189159 190438 194141 167143

203213 199818 184825 180691 200742 194705 190464 194923 172493

203816 202675 185509 180727 200828 195703 190956 195806 172656

208856 208417 185787 181116 201009 197294 191248 196157 175263

176575 174262 185800 181993 201454 199322 191329 196510 176465

176724 176725 186151 182875 201461 199384 191509 198903 181199

179137 177117 186896 183620 201734 200812 194227 199161 181763

179166 177130 187316 184871 201823 201486 196641 199387 182554

179409 181664 187572 185383 202026 202762 199178 200890 185875

180014 183965 187821 187589 202788 203248 201523 203065 186729

180720 184379 187851 190291 203710 203418 202023 203474 194408

180742 184435 188590 190867 204510 161261 175446 204099 194872

182243 185551 189751 191832 205089 168989 176157 205050 195707

198073 198153 198973 199211 199391 199767 199795 201388 201611

201792 201989 203856 175365 181526 182083 186003 188068 188174

188853 189208 189576 190091 190177 190383 190645 191485 203376

203380 204966

**嫌疑较小的ID号：**

161379 181456 195872 173987 162994 104606 30508 156311 167582

177906 181611 196615 174608 173939 137661 34377 158393 167793

186011 182988 196892 64729 174302 142395 47729 160127 168014

187781 184694 199095 161812 28939 162835 56452 161125 169027

188170 185801 205013 104457 160458 164018 64148 161422 169169

191582 200772 173619 106557 163997 165951 70408 162022 169825

192699 202018 179872 159933 164173 168047 72054 162066 170778

193937 162861 197014 167675 165563 169020 84779 162494 171307

195845 177491 160566 163049 169190 169096 90964 162650 171376

200441 189076 160426 78772 169988 169316 96084 162714 171627

202650 189087 167518 110779 174790 169450 101212 163074 171777

202965 190087 168815 162142 11672 171369 101297 163231 171826

16206 191193 169945 165762 29122 174128 112372 163270 172350

164980 193074 171204 168434 86833 174310 116627 165161 172421

179531 200504 201812 169757 161088 174464 121681 165793 172875

180063 202303 46911 170251 169164 178412 128880 165853 173192

185290 176605 102938 172334 4214 2162 144520 166112 173403

201156 181520 164367 7768 11644 2438 144933 167014 177214

161926 189041 165922 21261 92630 16818 149660 167082 178423

180842 191311 168930 120930 96689 18152 155540 167563 178470

#### 附录10 三类嫌疑ID号的信息挖掘（盒图绘制）

%----------------------------------------------------------------------------------------------------------------------------%

% 重大嫌疑，次要嫌疑，嫌疑较小的信息挖掘

%----------------------------------------------------------------------------------------------------------------------------%

%% 清屏

clc

clear

%% 相关ID号的数据导入

% 重大嫌疑ID号

data1=xlsread('E:\2015深圳杯A题原数据\100组问题数据\统计结果.xls','B1:B2109');

% 次要嫌疑ID号

data2=xlsread('E:\2015深圳杯A题原数据\100组问题数据\统计结果.xls','E1:E912');

% 嫌疑较小ID号

data3=xlsread('E:\2015深圳杯A题原数据\100组问题数据\统计结果.xls','H1:H180');

disp('ID号全部读取成功');

%% 匹配信息的读取

% 9个弱因子的数据读取

disp('9个弱因子的数据读取');

Weakfactor=xlsread('E:\2015深圳杯A题原数据\2015深圳杯A题\神经网络全部样本数据.xlsx','A2:J58013');

disp('Done！');

% 部分欺诈识别因子的数据读取

% 买药频率数据的读取

pinlv=xlsread('E:\2015深圳杯A题原数据\所有ID号频率数据.xlsx','A1:B58017');

disp('买药频率.....done！');

%% 不同类别的数据匹配

% 重大嫌疑类

[x,~]=size(data1);

for i=1:x

curID=data1(i);

place=find(Weakfactor(:,1)==curID);

Weakfactorinf(1,:)=Weakfactor(place,2:10);

place2=find(pinlv(:,1)==curID);

Recfainf=pinlv(place2,2);

data1(i,1)=curID;

data1(i,2:10)=Weakfactorinf(1,:);

data1(i,11)=Recfainf;

end

disp('重大嫌疑.....done！.');

% 次要嫌疑类

[x,~]=size(data2);

for i=1:x

curID=data2(i);

place=find(Weakfactor(:,1)==curID);

Weakfactorinf=Weakfactor(place,2:10);

place2=find(pinlv(:,1)==curID);

Recfainf=pinlv(place2,2);

data2(i,1)=curID;

data2(i,2:10)=Weakfactorinf(1,:);

data2(i,11)=Recfainf;

end

disp('次要嫌疑.....done！.');

% 嫌疑较小类

[x,~]=size(data3);

for i=1:x

curID=data3(i);

place=find(Weakfactor(:,1)==curID);

Weakfactorinf=Weakfactor(place,2:10);

place2=find(pinlv(:,1)==curID);

Recfainf=pinlv(place2,2);

data3(i,1)=curID;

data3(i,2:10)=Weakfactorinf(1,:);

data3(i,11)=Recfainf;

end

disp('嫌疑较小.....done！.');

%% 信息输出

xlswrite('C:\Users\Public\Desktop\重大嫌疑类.xls',data1,1);

xlswrite('C:\Users\Public\Desktop\次要嫌疑类.xls',data2,1);

xlswrite('C:\Users\Public\Desktop\嫌疑较小类.xls',data3,1);

%-------------------------------------------------------------------------------------------------------------------%

% 信息盒图的绘制

%--------------------------------------------------------------------------------------------------------------------%

%%

clc

clear

%% 数据的导入

% 重大嫌疑

data1=xlsread('C:\Users\Public\Desktop\重大嫌疑类.xls','B1:K2109');

data2=xlsread('C:\Users\Public\Desktop\次要嫌疑类.xls','B1:K912');

data3=xlsread('C:\Users\Public\Desktop\次要嫌疑类.xls','B1:K180');

%% 信息随机抽取

a=randperm(2109);

a=a(1:180);

data1=data1(a,:);

b=randperm(912);

b=b(1:180);

data2=data2(b,:);

%% 绘制盒图

figure

subplot(131);

% f1=boxplot([data1(:,1),data1(:,2),data1(:,3),data1(:,4),data1(:,5),data1(:,6),data1(:,7),data1(:,8),data1(:,9),data1(:,10)]);

f1=boxplot([data1(:,1),data1(:,2),data1(:,3),data1(:,6),data1(:,7),data1(:,8),data1(:,9)]);

% set(gca,'XTickLabel',{'A','B','E','F','G','H','I','J'});

set(gca,'xtick',[0:100:500]);

set(gca,'ylim',[0 500]);

title('重大嫌疑');

subplot(132);

% f2=boxplot([data2(:,1),data2(:,2),data2(:,3),data2(:,4),data2(:,5),data2(:,6),data2(:,7),data2(:,8),data2(:,9),data2(:,10)]);

f2=boxplot([data2(:,1),data2(:,2),data2(:,3),data2(:,6),data2(:,7),data2(:,8),data2(:,9)]);

% set(gca,'XTickLabel',{'A','B','E','F','G','H','I','J'});

set(gca,'xtick',[0:100:500]);

set(gca,'ylim',[0 500]);

title('次要嫌疑');

subplot(133);

% f3=boxplot([data3(:,1),data3(:,2),data3(:,3),data3(:,4),data3(:,5),data3(:,6),data3(:,7),data3(:,8),data3(:,9),data3(:,10)]);

f3=boxplot([data3(:,1),data3(:,2),data3(:,3),data3(:,6),data3(:,7),data3(:,8),data3(:,9)]);

% set(gca,'XTickLabel',{'A','B','E','F','G','H','I','J'});

set(gca,'xtick',[0:100:500]);

set(gca,'ylim',[0 500]);

title('嫌疑较小');

figure(2)

subplot(131)

f4=boxplot([data1(:,4),data1(:,5)]);

set(gca,'ylim',[0 1000]);

title('重大嫌疑');

subplot(132)

f5=boxplot([data2(:,4),data2(:,5)]);

set(gca,'ylim',[0 1000]);

title('次要嫌疑');

subplot(133)

f6=boxplot([data3(:,4),data3(:,5)]);

set(gca,'ylim',[0 1000]);

title('嫌疑较小');

#### 附录11 三类ID号的人员统计程序

主函数main

%-------------------------------------------------------------------------%

% 数据挖掘部分程序

% 对于已知的划分区间，统计重大嫌疑/次要嫌疑/嫌疑较小三类的人数

%-------------------------------------------------------------------------%

%% 清屏

clc

clear all

close all

%% 数据导入

zhongyao=xlsread('C:\Users\Lenovo\Desktop\2015深圳杯数据挖掘部分\重大嫌疑类.xlsx','A1:J3201');

ciyao=xlsread('C:\Users\Lenovo\Desktop\2015深圳杯数据挖掘部分\次要嫌疑类.xlsx','A1:J912');

jiaoxiao=xlsread('C:\Users\Lenovo\Desktop\2015深圳杯数据挖掘部分\嫌疑较小类.xlsx','A1:J180');

disp('数据导入完毕！')

clc

%% 数据的随机抽取（150个）

[x1,~]=size(zhongyao);

[x2,~]=size(ciyao);

[x3,~]=size(jiaoxiao);

suiji1=randperm(x1);

suiji2=randperm(x2);

suiji3=randperm(x3);

chouqu1=suiji1(1:150);

chouqu2=suiji2(1:150);

chouqu3=suiji3(1:150);

zhongyaodata=zhongyao(chouqu1,:);

ciyaodata=ciyao(chouqu2,:);

jiaoxiaodata=jiaoxiao(chouqu3,:);

disp('数据随机抽取完毕!');

%% 年龄的统计(合格)

nianl=zeros(6,3);

for i=1:150

a1=tongji(zhongyaodata(i,3),1);

a2=tongji(ciyaodata(i,3),1);

a3=tongji(jiaoxiaodata(i,3),1);

nianl(a1,1)=nianl(a1,1)+1;

nianl(a2,2)=nianl(a2,2)+1;

nianl(a3,3)=nianl(a3,3)+1;

end

%% 买药频率的统计

pinlv=zeros(6,3);

for i=1:150

a1=tongji(zhongyaodata(i,4),2);

a2=tongji(ciyaodata(i,4),2);

a3=tongji(jiaoxiaodata(i,4),2);

pinlv(a1,1)=pinlv(a1,1)+1;

pinlv(a2,2)=pinlv(a2,2)+1;

pinlv(a3,3)=pinlv(a3,3)+1;

end

%% 买药总价的统计

zongjia=zeros(6,3);

for i=1:150

a1=tongji(zhongyaodata(i,5),3);

a2=tongji(ciyaodata(i,5),3);

a3=tongji(jiaoxiaodata(i,5),3);

zongjia(a1,1)=zongjia(a1,1)+1;

zongjia(a2,2)=zongjia(a2,2)+1;

zongjia(a3,3)=zongjia(a3,3)+1;

end

%% 买药总数的统计

zongshu=zeros(6,3);

for i=1:150

a1=tongji(zhongyaodata(i,6),4);

a2=tongji(ciyaodata(i,6),4);

a3=tongji(jiaoxiaodata(i,6),4);

zongshu(a1,1)=zongshu(a1,1)+1;

zongshu(a2,2)=zongshu(a2,2)+1;

zongshu(a3,3)=zongshu(a3,3)+1;

end

%% 医嘱的统计

yizhu=zeros(5,3);

for i=1:150

a1=tongji(zhongyaodata(i,10),5);

a2=tongji(ciyaodata(i,10),5);

a3=tongji(jiaoxiaodata(i,10),5);

yizhu(a1,1)=yizhu(a1,1)+1;

yizhu(a2,2)=yizhu(a2,2)+1;

yizhu(a3,3)=yizhu(a3,3)+1;

end

子函数 tongji.m

% 子函数 年龄段划分

function a=tongji(x,L)

%%

if L==1 % 第一类 年龄

if (x<=20&&x>=0)

a=1;

elseif (x>=21&&x<=40)

a=2;

elseif(x>=41&&x<=60)

a=3;

elseif (x>=61&&x<=80)

a=4;

elseif(x>=81&&x<=100)

a=5;

elseif(x>100)

a=6;

end

end

%%

if L==2 % 第二类 买药频率

if (x<=4&&x>=0)

a=1;

elseif (x>=5&&x<=8)

a=2;

elseif(x>=9&&x<=12)

a=3;

elseif (x>=13&&x<=16)

a=4;

elseif(x>=17&&x<=20)

a=5;

else

a=6;

end

end

%%

if L==3

if (x<=80&&x>=0)

a=1;

elseif (x>=81&&x<=160)

a=2;

elseif(x>=161&&x<=240)

a=3;

elseif (x>=241&&x<=320)

a=4;

elseif(x>=321&&x<=400)

a=5;

else

a=6;

end

end

%%

if L==4

if (x<=80&&x>=0)

a=1;

elseif (x>=81&&x<=160)

a=2;

elseif(x>=161&&x<=240)

a=3;

elseif (x>=241&&x<=320)

a=4;

elseif(x>=321&&x<=400)

a=5;

else

a=6;

end

end

%%

if L==5

if (x<=0.2&&x>=0)

a=1;

elseif (x>0.2&&x<=0.4)

a=2;

elseif(x>0.4&&x<=0.6)

a=3;

elseif (x>0.6&&x<=0.8)

a=4;

elseif(x>0.8&&x<=1)

a=5;

end

end

end