**基于主成分分析、支持向量机和神经网络的放贷决策模型**

**摘要**

本文依据以往债务人个人信息和违约情况的历史记录，通过主成分分析、支持向量机等方法，得到决定债务人个人信用的各个主成分，建立可供保理商参考的债务人个人信用和是否放贷的放贷决策模型，给出了由债务人的个人信息和历史信用记录决定是否放贷、放贷额度大小的系统化方法。

针对以往的大量客户的个人信息和历史违约记录，本文首先根据已经标准化的各项个人信息的参数，在对所有条目已知的客户的表格项中进行主成分分析，得到了各项个人信息中对违约记录具有不同影响大小因素的主成分元素集合。基于此，将每个元素对于历史违约记录的影响权重大小从高到低进行排序，取前几个主成分元素作为主成分集合，使得主成分集合中的元素在总体决定权重中的占比超过85%，解决了个人信息中哪些信息主要决定了客户借款后违约的可能性问题。自此，针对预测客户违约的可能性因素，我们已从完整数据集合中找出数据规律，生成了违约预测模型，将得到的主要因素用于接下来的支持向量机中。

针对最后80个class条目未知的客户预测是否违约的问题，本文首先根据前文所提到的违约预测模型建立起的主成分分析元素利用支持向量机建立起了分类模型。在此基础上，根据算法的结果，进行放贷决策的预判。整体只要最终计算出的信用水平不太差则会放贷，具体区别是在基准水平以上则多贷一些，以下则相应减少。否则，若最终的信用水平过低，即低于阈值，则不予贷款。同时，由距离超平面距离的大小决定可能性大小的原理计算导出支持向量机不同输入的概率作为客户违约的可能性。

针对数据有残缺情形下的确定授信额度的分析问题，本小组采用以范数偏差为参数的热卡方法对残缺的参数进行填补，即先将残缺的部分用对应列整体数据的均值作为暂时的带体制，寻找与当前残缺的最接近的一组数据中的对应行中的对应元素作为缺失值的替代值，考虑到若将热卡方法算出来的模拟值直接代入原始数据中经过矩阵变换有可能新矩阵的属性列的分布会改变，即使用原有的数据会有系统性误差，所以用神经网络的办法重新建模，重新得出新的。

针对向公司管理层提供的非技术报告，我们将根据分析出的主成分进行分析，并通过聚类分析寻找其数学上的相关性，并结合列表属性的生活实际进行客观解释，对管理层日后对一组数据决策的迅速决策做出科学指导。

关键词：主成分分析，RBF支持向量机，机器学习，神经网络

1. **问题重述**

随着当今国内外贸易活动的日益活跃，赊销的方式越来越流行，这也对保理业务提

出了新的要求。在大数据和人工智能高速发展的大背景下，确立更加精准可靠的放贷机制成为可能。现实生活中，客户申请放贷的过程中需要提供大量的个人信息，放贷过程中也会产生相关数据。从分析已有的大量案例数据入手，结合相关资料，我们将解决以下几个问题：

1、根据文件数据，寻找数据规律，建立用来预测客户违约可能性的违约预测模型并用实例验证。

2、根据文件数据，分析所有数据之间的联系，建立用于决策授信额度的授信额度估算模型，并用实例验证。

3、结合生活中实际情况，用户提供的信息存在不完整的可能性。结合参数之间的联系，建立有数据缺失情况下的授信额度估算模型。

4、针对上述三个问题的解答，撰写技术报告，展示建模成果。

1. **模型假设**
2. 对于训练条目，数据越完整，分析出来的模型越真实；对于测试条目，数据越完整，分析出来的结果越合理。
3. 根据我们填补缺省值的评价标准，拥有较好填补缺省值性能的填补方法对于数据不完整条目的决策分析所提供的支持越大。
4. 假设残缺数据的客户仍属于完整客户个人信息数据集的完整分布中。
5. **符号说明**

符号 含义 单位

 第i号客户 无

 第i号属性 无

 第i号客户的第j号属性 无

 第i号客户是否违约 无

 总体正确度 无

 第i号客户的授信额度 万元

 第i个主成分 无

1. **问题分析**

**4.1决策任务分解分析**

针对新客户的放贷决策是典型的分类和回归问题。其核心在于分别针对信息完整和存在数据残缺的两种客户的现有个人信息和历史违约记录，并结合已有数据集合进行分析，并给出合理的个人信用评价指标体系，对模型进行检验和评估。

**4.2 数据划分和数据填充分析**

为评价决策模型的优劣，考虑从完整数据情况和缺失数据两种情况来共同确定评价指标体系。对表格中的整体数据条目划分为以下三类：过去信用记录未知（即最后80个class条目未知）的条目；class完整条目中的大部分（用于生成对完整数据情况下的分析的训练集）；其余部分（用于对缺失情况下的目录进行模拟测试，即选择其中某几个进行遮蔽，将这些条目视作存在缺失数据的条目）。从数据完整的情况出发，关心的是对于所有已知数据的合理利用得出符合以往信用记录（即违约情况）的决策。从数据存在缺失的情况出发，关心的是针对利用热卡方法算出来的模拟缺失值进行填补后的条目做出的决策符合以往信用记录的决策。符合以往信用记录的决策意味着，若该条目以往存在着违约记录，则做出的决策应该是在已有信用额度的基础上进行减少或根本不予贷款（若已有信用额度缺失则使用神经网络预测出模拟值再进行计算）；若该条目以往不存在违约记录，则做出的决策应该是在已有信用额度的基础上增加或至少不变才是合理的。综合考虑完整数据情况和缺失数据两种情况给出评价指标体系。

**4.3数据归一化处理与分析算法**

客户在决策系统中提供的参数由户籍所在地、婚姻状况、年龄等构成。统计附录中的数据，发现可用于除了部分数值太小之外需要处理，所有的数据条目均可用于分析。先对数据完整的情况进行决策，即对数据进行归一化处理得到归一化之后的矩阵，将此矩阵用于主成分分析，并以此作为支持向量机的分析预测的输入，并得出支持向量机的决策模型，并以此作为计算新的决策的依据。通过比较计算出的结果和阈值进行比较，如果大于阈值则确定给予贷款服务，并根据其距离超平面距离的大小决定可能性大小的原理计算导出支持向量机不同输入的概率作为客户违约的可能性。

**4.4对缺失数据项决策的单独处理**

再对数据存在缺失值的情况进行决策，即先用热卡算法进行缺失值模拟出模拟值进行填补，接着，运用bp神经网络进行处理。将拥有完整数据的条目的V1到V28个参数作为bp神经网络的输入层，将class即违约历史记录作为输出层，通过正向和反向两个子过程，以Widrow-Hoff学习规则为原则，根据梯度下降法，分别调整输入层和隐含层、隐含层和输出层的权值和阈值，不断优化，最终得到整个神经网络的整体结构，以及每个结点的权值和阈值，至此，即已经获得了针对数据存在缺失值的放贷决策模型，其中的计算结果就已经包含了授信额度和客户违约的可能性。

最后，是我们提出的针对缺失值的统一填补算法（如我们采用的热卡算法）在不同缺失情况下的评价体系，并记录在我们的非技术报告中。

1. **模型建立与求解**

**5.1基于主成分分析和支持向量机的数据完整情况下放贷决策模型（授信额度估算和违约可能性预测）**

**5.1.1模型建立**

根据表格将客户的28个个人信息属性分别表示为，例如“户籍所在地”；每个属性决定权大小为，其中n个对客户违约情况具有主要决定权的属性表示为，对应属性决定权大小为，其中这n个属性对放贷决策情况的决定权之和占总体决定权的比重（累计贡献率）大于85%，即



通过分析之前找出来的n个具有主要决定权的属性，试图寻找一种分类方式,其中此时的为之前寻找出的具有主要决定权的属性，通过将是否违约的两类客户正确分开，并使的两类客户的个人属性集合通过这种分类方式所计算出的差异性尽可能大，从而实现合理划分。

这时，再将属性完整的待预测客户的主要属性输入到中，通过比较计算结果和阈值的大小，判断出的值，即做出是否违约的判断。

授信额度的分析单独使用神经网络进行分析。将所有条目中的28个属性作为参数传入输入层，将授信额度作为标签传入输出层，然后根据Widrow-Hoff学习规则从正向和反向两个子过程进行回归学习模拟，最终生成完整的神经网络图，将此网络结构图作为授信额度的分析模型，回归分析出新用户class未知的80个客户的授信额度预测值。

**5.1.2模型求解算法**

**5.1.2.1主成分分析算法**

对原始客户数据矩阵进行标准化处理,将变换为：

，

表示第j列属性的均值，表示第j列属性的标准差。通过标准化变化，得到标准化矩阵。求解相关系数矩阵：



计算矩阵的特征值（即为各个属性决定权大小）和对应的特征向量。将从大到小排序，从第一项开始寻找出n个，使得



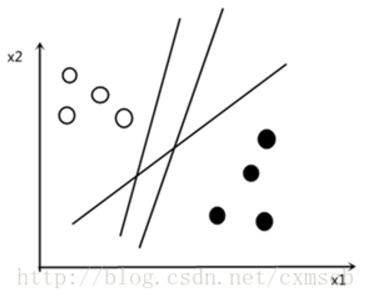
则为分析得出的主成分权重，与对应的为分析出的主成分属性。

**5.1.2.2 RBF支持向量机算法**

根据给定的训练集



式中，l为客户信息条目数,；称为输入空间，输入空间中的每一个点由n个主成分属性组成，下面寻找的一个实值函数,以便用分类函数



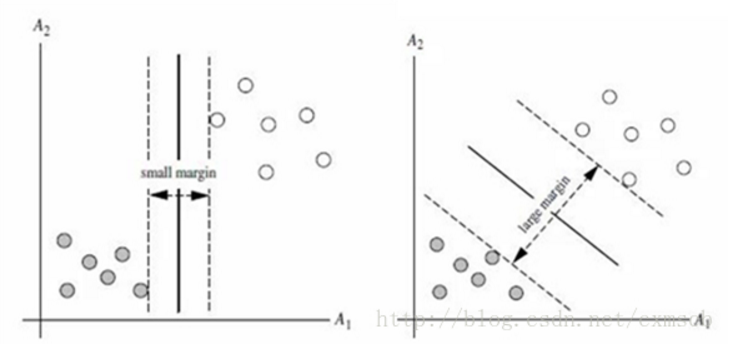


推断任意一个客户的个人信息集合相对应的值的问题为分类问题。此分类方式即为支持向量机分类方式，此分类函数即为超平面。下面给出结合本文定义的规范超平面的定义：

空间中超平面都可以写为的形式，参数乘以任意一个非零常数后得到的是同一个超平面，定义满足条件

的超平面为训练集T的规范超平面。



定义满足成立的称为普通支持向量，普通支持向量间的间隔为。最有超平面即意味着最大化，称为分类边界，于是寻找最优超平面的问题可以转化为如下的二次规划问题：

 ,

该问题的特点是目标函数是的凸函数，并且约束条件都是线性的。

引入函数



式中：为乘子。

根据对偶的定义，通过对原问题中各变量的偏导置零，得

，

，

代入函数化为原问题的对偶问题：

 ，



求解上述最优化问题，得到最优解，计算

，

由互补条件知



可得只有当为支持向量的时候，对应的才为正，否则皆为0。选择的一个正分量，并以此计算



于是构造分类超平面,并由此求得决策函数



得到分类函数

，

从而对未知样本进行分类。

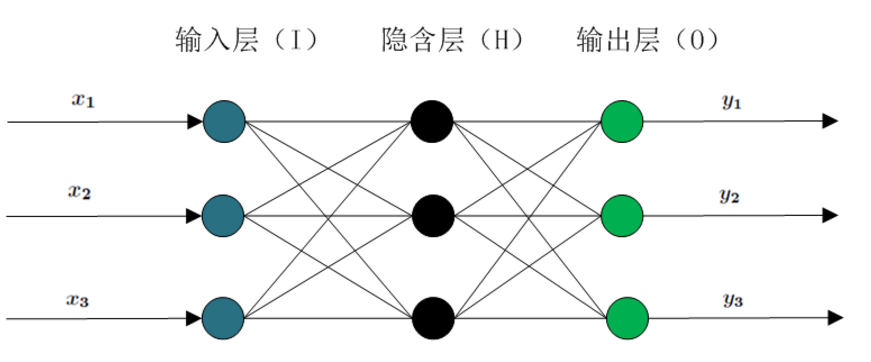
**5.1.2.3神经网络分析算法**

本论文主要应用的神经网络分支是BP(Back Propagation)神经网络，下面对BP神经网络算法进行阐述。

BP神经网络分为两个过程：

1. 工作信号正向传递子过程
2. 误差信号反向传递子过程

在BP神经网络中，单个样本有m个输入，有n个输出，在输入层和输出层之间通常还有若干个隐含层。实际上，1989年证明了对于任何闭区间的一个连续函数都可以用一个隐含层的BP网络来逼近，即万能逼近定理。所以一个三层的BP网络就可以完成任意的m维到n维的映射。即这三层分别是输入层（I），隐含层（H），输出层（O）。如图所示：



1. 工作信号正向传递子过程

现在设节点和节点之间的权值为，节点的阀值为，每个节点的输出值为，而每个节点的输出值是根据上层所有节点的输出值、当前节点与上一层所有节点的权值和当前节点的阀值还有激活函数来实现的。具体计算方法如下





其中为激活函数，一般选取型函数或者线性函数。

正向传递的过程比较简单，按照上述公式计算即可。在BP神经网络中，输入层节点没有阈值。

1. 反向传递子过程

在BP神经网络中，误差信号反向传递子过程比较复杂，它是基于Widrow-Hoff

学习规则的。假设输出层的所有结果为，误差函数如下



而BP神经网络的主要目的是反复修正权值和阀值，使得误差函数值达到最小。Widrow-Hoff学习规则是通过沿着相对误差平方和的最速下降方向，连续调整网络的权值和阀值，根据梯度下降法，权值矢量的修正正比于当前位置上E(w,b)的梯度，对于第个输出节点有



假设选择激活函数为



对激活函数求导，得到



那么接下来针对有



同样，对于，有



这就是著名的IMG_256学习规则，通过改变神经元之间的连接权值来减少系统实际输出和期望输出的误差，这个规则又叫做Widrow-Hoff学习规则或者纠错学习规则。

上面是对隐含层和输出层之间的权值和输出层的阀值计算调整量，而针对输入层和隐含层和隐含层的阀值调整量的计算更为复杂。假设是输入层第k个节点和隐含层第i个节点之间的权值，那么有





有了上述公式，根据梯度下降法，那么对于隐含层和输出层之间的权值和阀值调整如下





而对于输入层和隐含层之间的权值和阈值调整同样有





通过将29个属性作为参数输入输入层，真实信用额度作为结果输入输出层，通过神经网络网络多次学习调节不同节点的权重和阈值，形成完整的神经网络图，最终逼近真实的结果。

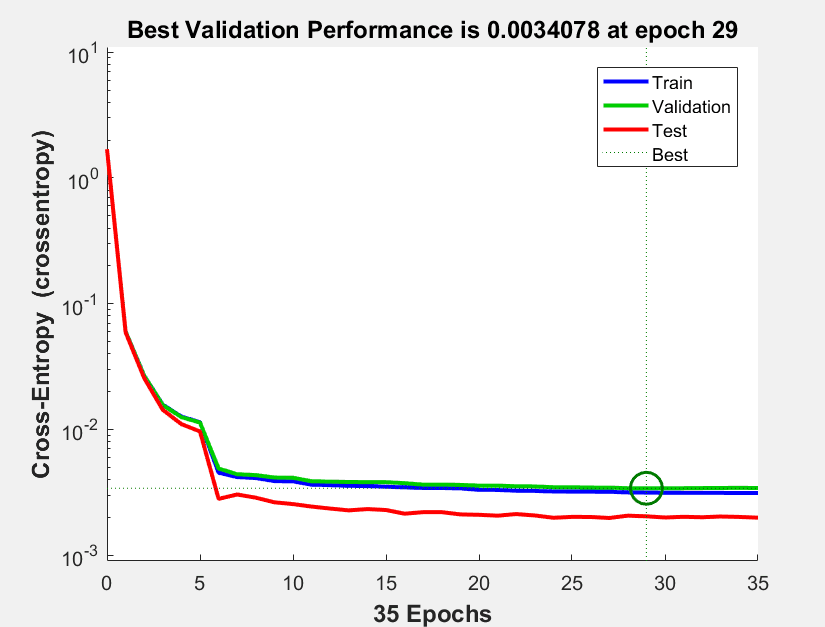
**5.1.3模型求解**

**5.1.3.1神经网络分析违约可能性结果**

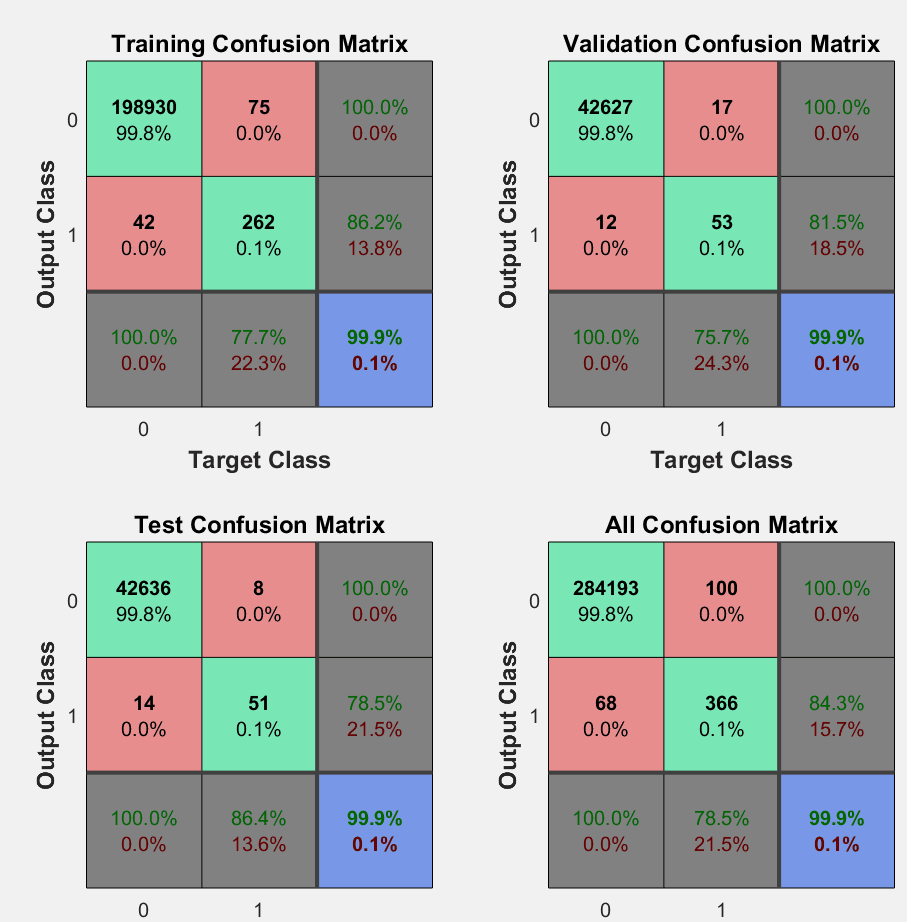
利用神经网络，将29个客户属性作为参数输入集，将实际违约结果作为参数作为输出集，根据纠错学习规则，不断调整神经网络节点中的权重值和阈值，使得输入集最终拟合成输出集。即通过比较输出结果与阈值的大小对客户违约的可能性进行分类判断直接做出违约的决策结果。

用所有数据训练的神经网络，该网络在测试集上的准确率99.9%以上。

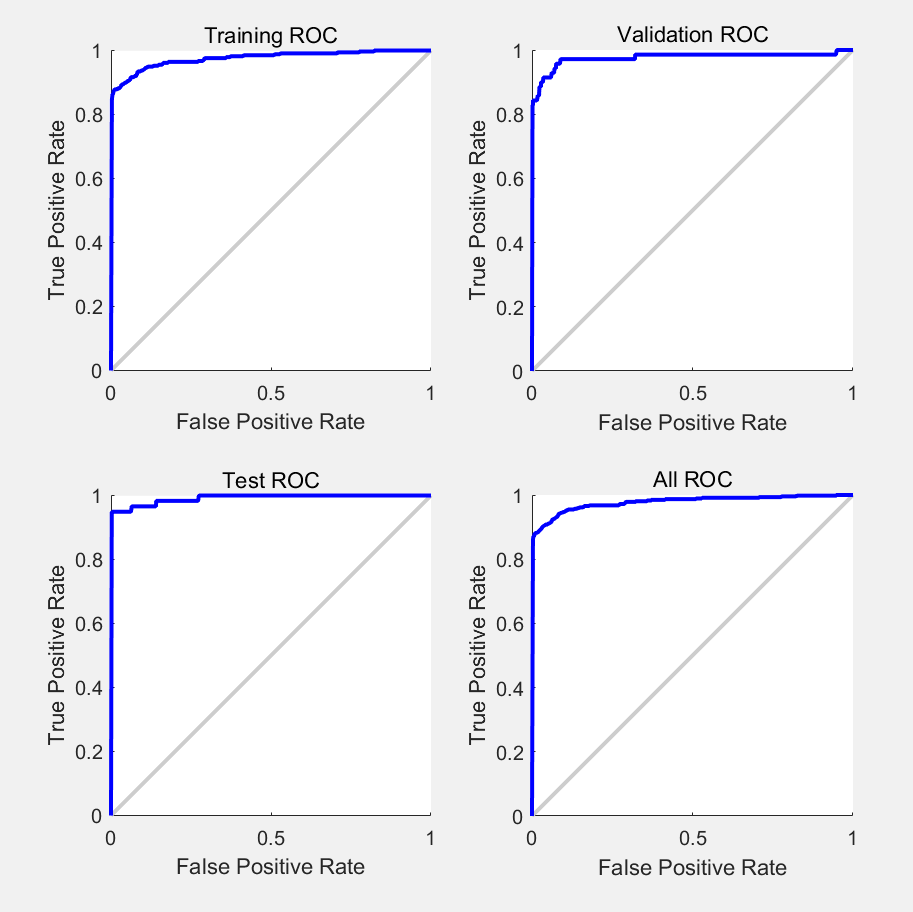
训练这个神经网络一共使用了284727个数据，其中70%作为训练集，15%作为验证集，15%作为测试集，



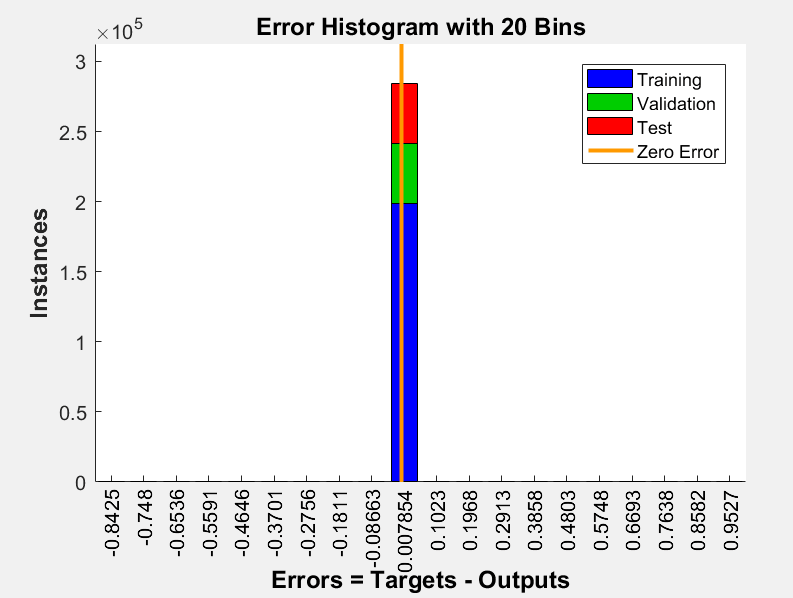
这张图片是训练过程中评价函数的值的变化，值越小，表示误差越小，可以看到我们的神经网络一共训练了35epoch，在第29epoch的时候到达最好的效果，此时的交叉熵为0.0034078。



这个是混淆矩阵，（看左下角）我们在测试集上的准确率为99.9%以上，同时可以看到，我们的神经网络具有非常小的概率（小于0.01%）把正常客户预测为违约客户，我们有13.6%的概率把违约客户预测为正常客户。

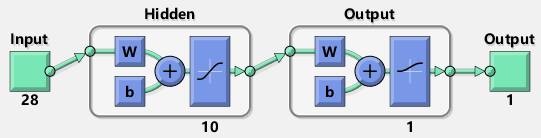


这是模型的ROC，这个和上一个的含义类似。



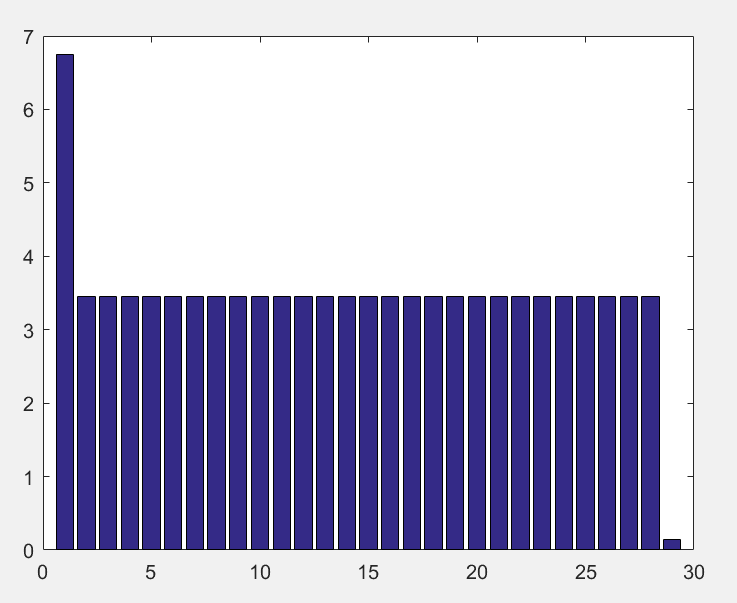
此为error的分布，含训练，验证，测试集的个数。

神经网络结构如下：



**5.1.3.2主成分分析与支持向量机预测class未知客户违约结果**

根据输入的用户信息属性数据，通过主成分分析，得出如下图结果，得出24个主成分实现了降维效果（原先属性有29项）



**主成分表**

**权重值排序 主成分**

1. 授信额度
2. 婚姻状况

3 卡均使用率

4 经营合同风险性质

5 同意交易对手频次

6 交易对手数量

7 利息保障倍数

8 毛利率

1. 退租率
2. 出租率
3. 月均还款占比
4. 月均银行账户资金留存
5. 年龄
6. 固定资产
7. 大额进出额交易频率
8. 贷款类查询记录（含本人）
9. 合伙人数量
10. 第三方征信风险得分
11. 申请人公检法记录
12. 户籍所在地
13. 申请人占股
14. 近5年内贷款逾期次数
15. 有效信用卡数
16. 新开金融类账户数量

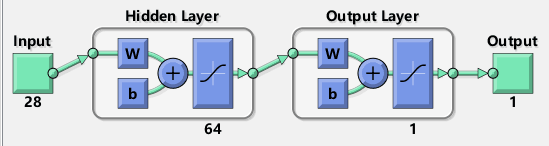
根据分析出的24个主成分，立用支持向量机计算出的80个class未知客户的违约结果如下。（注：0表示不违约，1表示违约）

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Class | ID | Class |
| U01 | 1 | U41 | 1 |
| U02 | 0 | U42 | 0 |
| U03 | 0 | U43 | 0 |
| U04 | 0 | U44 | 0 |
| U05 | 0 | U45 | 0 |
| U06 | 1 | U46 | 0 |
| U07 | 0 | U47 | 0 |
| U08 | 0 | U48 | 0 |
| U09 | 1 | U49 | 0 |
| U10 | 1 | U50 | 0 |
| U11 | 0 | U51 | 0 |
| U12 | 1 | U52 | 0 |
| U13 | 0 | U53 | 0 |
| U14 | 0 | U54 | 1 |
| U15 | 1 | U55 | 0 |
| U16 | 1 | U56 | 0 |
| U17 | 1 | U57 | 0 |
| U18 | 0 | U58 | 0 |
| U19 | 0 | U59 | 0 |
| U20 | 1 | U60 | 0 |
| U21 | 1 | U61 | 0 |
| U22 | 0 | U62 | 1 |
| U23 | 1 | U63 | 1 |
| U24 | 1 | U64 | 0 |
| U25 | 0 | U65 | 0 |
| U26 | 0 | U66 | 0 |
| U27 | 0 | U67 | 0 |
| U28 | 0 | U68 | 1 |
| U29 | 0 | U69 | 0 |
| U30 | 0 | U70 | 1 |
| U31 | 0 | U71 | 0 |
| U32 | 0 | U72 | 1 |
| U33 | 1 | U73 | 0 |
| U34 | 0 | U74 | 0 |
| U35 | 0 | U75 | 0 |
| U36 | 1 | U76 | 0 |
| U37 | 0 | U77 | 0 |
| U38 | 0 | U78 | 0 |
| U39 | 0 | U79 | 1 |
| U40 | 0 | U80 | 0 |

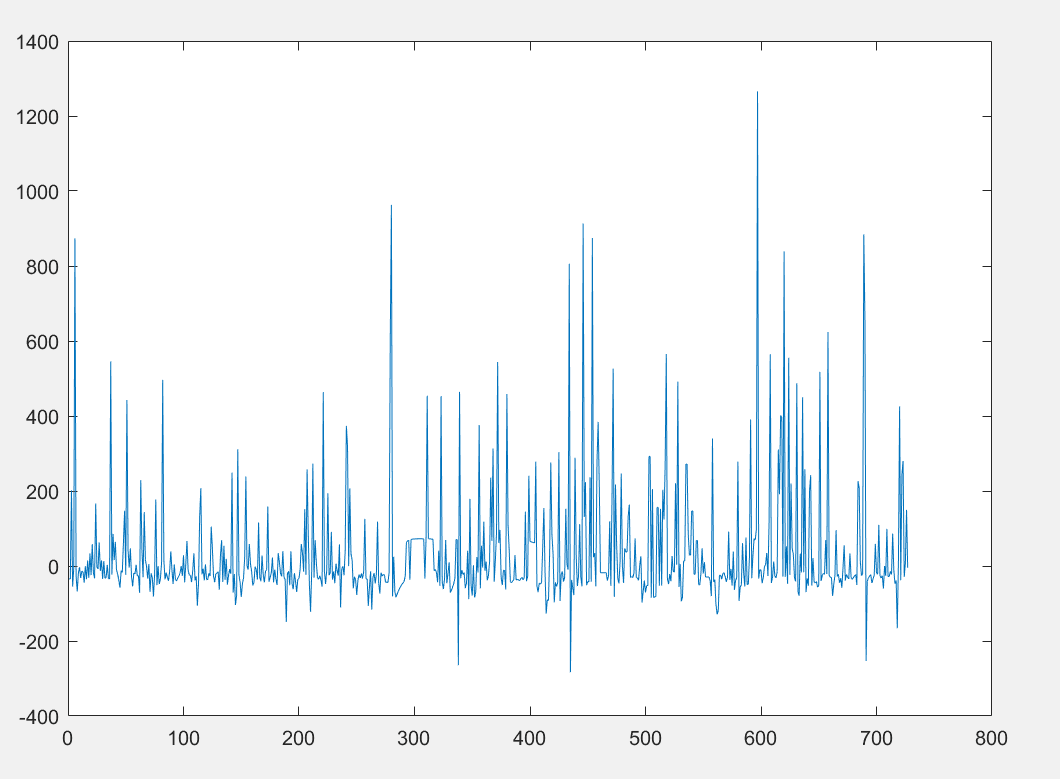
**5.1.3.3 BP神经网络进行授信额度估算**

为了实现对授信额度的估算，我们训练了回归的神经网络，根据已有数据估算授信额度。

BP（Back Propagation）神经网络传输函数为函数；训练函数为函数；学习规则为；所使用的误差评判标准为均方误差。神经网络的模型结构和参数如下图所示：



为了对所建立的神经网络模型准确度进行验证，我们划分出所提供的数据集中的最后727条数据进行预测。对于这些数据的误差可视化结构如下图所示，经计算真实值和预测值之间的误差（均方误差）为33.659760694303344。



**5.2基于神经网络的数据残缺情况下客户放贷决策模型（授信额度估算和违约可能性预测）**

**5.2.1模型建立**

针对数据残缺的情况，根据“将待解决的问题转换为已解决问题”的数学思想，即考虑使用数据完整的情况中神经网络作为模板解决数据残缺情况下客户放贷决策问题。考虑残缺数据项的客户应属于完整数据集合的整体分布，故考虑采用某种方式（即后面提到的然卡方式）填补缺省数据项，使得填补后的数据客户条目项仍属于完整分布。

接着，可考虑使用神经网络模型同时解决放贷决策和违约可能性大小。即将28项个人数据集合作为输入集，将class和个人信用额度作为输出集，通过神经网络中纠错学习规则不断修改网络中结点的权重值和阈值，使得输入集通过神经网络的输出吻合输出集结果。至此，数据残缺情况下的客户放贷决策模型已经建立，即上文提到的这个神经网络图。输入含残缺项的客户条目，即可同时获得放贷决策和违约可能性大小。

**5.2.2模型求解算法**

**5.2.2.1 热卡填补残缺值算法**

热卡填充原理：对于一个包含空值的对象，热卡填充法在完整数据中找到一个与它最相似的对象，然后用这个相似对象的值来进行填充。

在缺少数据项的客户条目中，假设缺失属性为，先用缺失属性值所在属性列的其余所有完整条目的对应属性值的平均值暂时填充得到条目，根据热卡填充原理，需要寻找一个总体水平和残缺条目最相近的条目，并用中对应的缺失属性填补，从而得到了用热卡方法填补的数据条目，以用于神经网络分析计算。

总体水平最相近条目的算法定义如下：

对其余的所有完整条目，计算与暂时用均值填补的条目的均方误差和达到最小的数据条目即可，即

对完整条目，，寻找到对应达到最小的条目，即为与残缺条目总体水平最相近的条目。然后填充中的对应缺失数据项至中即可。

**5.2.2.2 残缺值填补性能评估算法**

为了进一步验证上述算法在填补残缺值时的性能，我们提出了一种评估算法。

该评估算法建立在以下三条假设的基础上：

假设1：所有属性值所占权重相等，即：

C:\Users\liuch\AppData\Local\Temp\ksohtml\wps8427.tmp.png

假设2：缺省一组属性值所占权重，等于组内各个属性权重值之和，即：

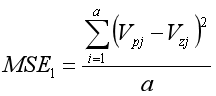
设属性集C:\Users\liuch\AppData\Local\Temp\ksohtml\wps8428.tmp.png

则有C:\Users\liuch\AppData\Local\Temp\ksohtml\wps8429.tmp.png

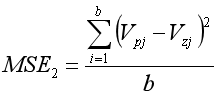
假设3：缺省属性的个数按照占总属性个数的百分比分为5%、10%、20%三档

假设有缺省5%属性的数据C:\Users\liuch\AppData\Local\Temp\ksohtml\wps842A.tmp.png组，缺省10%属性的数据b组，缺省30%属性的数据c组

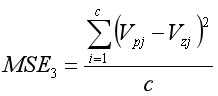
缺省5%属性的填充均方差



缺省10%属性的填充均方差



缺省20%属性的填充均方差



总体数据填充均方差

C:\Users\liuch\AppData\Local\Temp\ksohtml\wps843D.tmp.png

C:\Users\liuch\AppData\Local\Temp\ksohtml\wps843E.tmp.png

注：C:\Users\liuch\AppData\Local\Temp\ksohtml\wps843F.tmp.png指测试数据中某一行有属性残缺的客户数据中残缺的第C:\Users\liuch\AppData\Local\Temp\ksohtml\wps8440.tmp.png个属性，C:\Users\liuch\AppData\Local\Temp\ksohtml\wps8441.tmp.png指与填充值对应的真实值。

**5.2.2.3 结合热卡算法的神经网络分析算法**

由于本文5.1.2.3和5.2.2.1两小节分别对bp神经网络算法和热卡算法的介绍，下面结合本章节需解决的处理存在数据缺失值的数据项的分析方法。

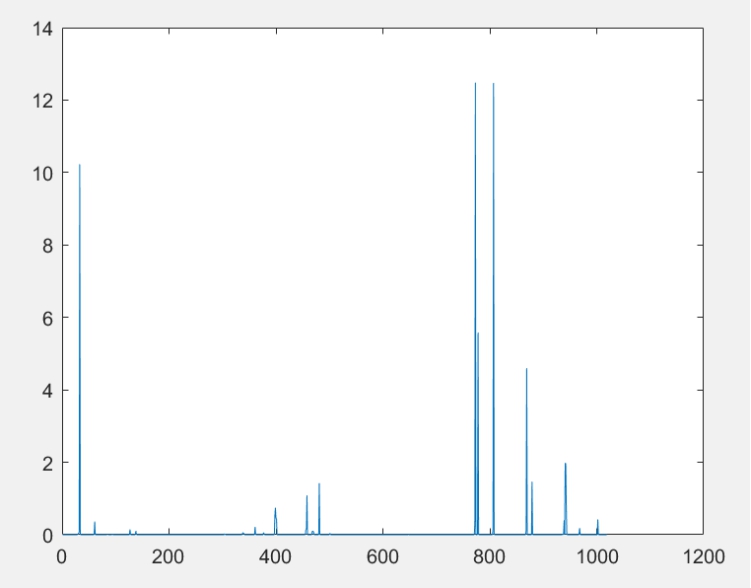
即先利用热卡填补算法填补出与原始数据基本吻合的数据项，然后利用将其视作数据完整的客户条目数据项，并将其中的对应元素与原始完整条目数据集合一起对应放入bp神经网络的输入层和输出层，通过正向传递和反向传递两个子过程，结合Widrow-Hoff学习规则得出完整的神经网络结构，并对其进行信用额度和违约可能的决策预判。

**5.2.3模型求解**

根据数据完整的客户情况，训练神经网络。具体的神经网络训练情况参见上文5.1.3.1模型求解部分。

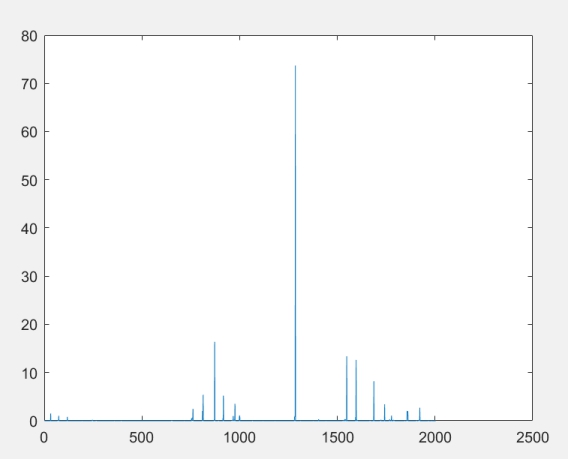
在本题中我们将解决用户提供数据不完整的情况下对用户授信额度的估算。我们采用热卡填充的算法，寻找“最接近”的数据填入残缺数据项。接下来是我们对热卡填充的准确程度的验证。验证过程分为三个部分：缺失5%数据、缺失10%数据和缺失20%数据。

1）共C:\Users\liuch\AppData\Local\Temp\ksohtml\wps175A.tmp.png个数据，缺失5%数据时，随机缺失值1019个，经过填补再计算和原来值之间的差值的平方得到下图：



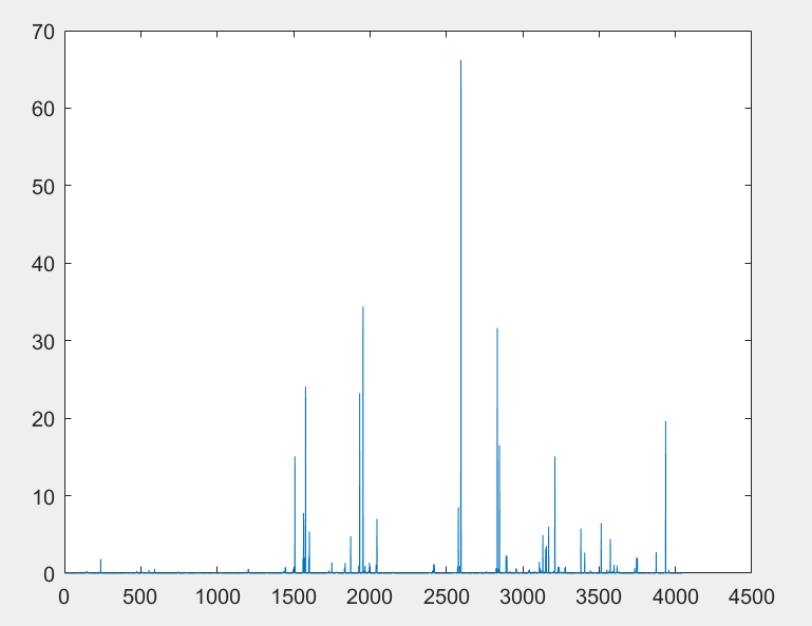
经过计算，计算出其均方误差,可见此时热卡填充效果较好。

2）缺失10%数据时。共C:\Users\liuch\AppData\Local\Temp\ksohtml\wps175D.tmp.png个数据，随机缺失值1019个，经过填补再计算和原来值之间的差值的平方得到下图：



经过计算，计算出其均方误差,可见此时热卡填充效果较好。

3）缺失20%数据时。共C:\Users\liuch\AppData\Local\Temp\ksohtml\wps1760.tmp.png个数据，随机缺失值1019个，经过填补再计算和原来值之间的差值的平方得到下图：



经过计算，计算出其均方误差,此时热卡填充随着缺失值的增大而增加。

C:\Users\liuch\AppData\Local\Temp\ksohtml\wps1773.tmp.png经过如上分析，我们可以得出热卡填充方式的均方误差相对较小，且随着缺失量的增多均方差有变大的趋势的结论，这与我们日常生活中的认知较为符合。

1. **模型评价与改进**

**6.1模型的优点**

1. 模型5.1中使用了主成分分析的方法，实现了数据降维，提高运算的速率。经验证，对客户是否违约的划分准确度达到99%以上，具有较高的准确程度。
2. 经验证，模型5.1中对数据完整情况下对授信额度估算的准确率较高。
3. 模型5.2中使用热卡方法填充缺失值，该算法运算效率较高、复杂度不高且鲁棒性强。经验证，热卡方式填充准确度很高，具体数据可见5.2.3。

**6.2模型的缺点**

1. 模型5.1中，对授信额度进行估算时，经验证在特殊情况下估算偏差较大。
2. 用于预测在数据不完整情况下违约情况的数据样本正负样本比例差距大。这导致在模型5.2中存在一定可能性将违约用户判定为不违约用户（约有13%的概率）
3. 模型5.2中采用热卡填充的方式，难以确定相似标准，受主观因素影响较大。同时依赖大量数据，否则影响填充的准确度。

**6.3模型的改进方向**

1. 对于模型5.1中存在的特殊值影响，可以用适当方法对个别数据进行处理
2. 对于数据中存在的违约正负样本比例差距大的问题，可以进一步进行数据增强进行矫正。
3. 对于模型5.2中采用的热卡填充的缺陷，可以使用基于朴素贝叶斯的缺失填补方法。
4. **参考文献**
5. **附件清单**