陕西师范大学校内数学建模竞赛

承　诺　书

我们仔细阅读了陕西师范大学校内数学建模竞赛的竞赛规则。

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与本队以外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的，如果引用别人的成果或其它公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们愿意承担由此引起的一切后果。

我们授权陕西师范大学校内数学建模竞赛组委会，可将我们的论文以任何形式进行公开展示（包括进行网上公示，在书籍、期刊和其他媒体进行正式或非正式发表等）。

参赛题号（从 A/B 中选择一项填写）：B

参赛队号：202327005023

参赛队员：　队员 1 姓名：宋月瑶

队员 2 姓名：杨阳

队员 3 姓名：丁梓睿

（除本页外不允许出现个人信息）

基于餐厅消费数据的隐形资助研究

摘要

隐形资助的精准展开与落实对高校的资助工作有着重要意义，本文利用多种模型分析，达到增强隐形资助准确率的效果。针对问题一，建立K－$Means$聚类模型；针对问题二，建立基于RBF神经网络算法的预测模型；针对问题三，建立了基于K－$Means$和RBF神经网络算法的综合预测模型。针对问题四，建立了规划模型，构建出细粒度资助额度分配算法。

对于问题一：首先，我们进行了数据的预处理：将寒暑假期间的所有数据删除；使用Matlab软件得到离群值，删除后用平均值补全法进行缺失补全；使用最大频率补全法对附件4-7的数据进行缺失补全；并根据性别差异、用餐结构等因素，提取出十个特征。其次，借助MATLAB软件分别计算男女生每个特征变量的特征值并使用SPSS对数据进行聚类分析，将男生与女生各分为三类，通过使用SPSS软件作图并计算平均数、最大值等，得到了每类内部主要消费行为较为稳定，且不同类间差距较大的结论。最后，我们将饮食种类按价格升序排列，使用SPSS软件对食物种类与价格的关系作图，找到两个价位分界点，为饮食种类分类，根据上面六类群体找到每类学生代表，分别计算他们吃高价位、中价位、低价位食物占他们总刷卡次数的比例，得到六类的学生间的饮食种类差距较大的结论。

对于问题二：我们建立了基于RBF神经网络算法的预测模型。首先利用数据8中已知的4000个数据用于训练神经网络，用剩余97个数据进行模型的误差分析检验。并借助MATLAB软件计算附件9中同学的特征变量，代入上述模型，补全了附件9。其次，再次使用Matlab软件，根据该组同学不同年份的特征变量，分别得到第二年、第三年的贫困程度隐形认定等级。最后，分析得到前一年是否享受补助对后一年的贫困程度隐形认定等级有影响的结论。

对于问题三：我们建立了基于K－$Means$和RBF神经网络算法的综合预测模型。首先，在原有的10个特征变量值的基础上增加了代表三个食物种类的特征变量值，得到新的特征变量数据。其次用RBF神经网络算法和K－$Means$聚类分别计算贫困等级，按1:1的比例综合结果，得到相关同学的预测结果。最后，分析预测结果变化，得到（未预测）结论。

对于问题四：我们建立了规划模型，构建出细粒度资助额度分配算法。首先，我们对第三学年附件4-7涉及的学生数据导出，先将贫困等级为0的学生剔除，并且将周消费低于15元的学生默认记为不在学校吃饭，也剔除。其次，为使每位学生受补助后的年消费水平相似，列举出了目标函数与四个约束条件。然后，得到了具体的分配额度，具体数据见支撑材料所示。最后，观察学生补助后的年消费水平，发现补助后学生的年消费水平非常统一，证明了上述算法的公平性与合理性。

基于K－$Means$和RBF神经网络算法的综合预测模型,我们考虑进行适当修改与优化：先使用$Canopy$－K－$Means$算法计算出准确的初始聚类中心，再用K－$Means$对RBF神经网络进行训练。形成$Canopy$－K－$Means$和RBF神经网络算法的综合预测模型。

关键词　平均值补全法；K－$Means$聚类；RBF神经网络算法；线性规划

1 问题重述

如何精准判定高校家庭经济困难学生、提高资助精度，对高校资助工作来说既是重点也是难点。近年来，许多高校选择了隐形资助的手段来进行贫困生认定工作，这种工作方式既最大限度地保证了资助工作的公平性，又保护了学生的自尊心。隐形资助的具体方式即为通过大数据挖掘的形式，以学生在餐厅消费金额、消费品类与消费次数等信息间接反映经济状况，并以此为依据识别出家庭经济困难学生，进行隐形实施的自主。因此，建立合适的贫困生识别模型对高校的资助工作价值很高。

问题一要求我们对附件1-3的该组学生不同学年的日三餐餐厅消费金额数据记录、附件4-7的其中部分同学的饮食种类信息数据进行预处理，并且针对处理后的数据建立模型，挖掘不同代表性群体，并定量分析该群体三学年的主要消费行为特征变化规律、饮食种类变化规律等。

问题二已知附件8给出部分同学第一学年后经其它方式认定的贫困程度等级，其中等级2准确但可能不全、其它等级认定可能有少量偏差，要求我们根据附件1-3建立数学模型依据消费行为，预测贫困程度，补全附件9。并且进一步结合第1问研究结论预测该组同学第二、第三学年的贫困程度隐形认定等级，分析相关变化。

问题三要求我们根据问题二所建立的模型，结合附件4-7的饮食种类数据，改进预测模型，比较分析相关同学的预测结果变化情况。

问题四要求我们通过以上贫困生本质特征挖掘，构建差异化资助额度分配算法。以第三学年为例，对附件4-7涉及的同学，进行资助总金额10万、资助人员80名的资助额度分配，给出具体资助结果，并对资助结果的公平合理性进行评估。

2 问题分析

2.1 问题分析1

针对第一问，由三个小问组成。

第一小问属于数据预处理问题，要求我们完成删除不相关数据、缺失补全、特征提取等操作。根据附件1-3的数据，我们观察发现数据出现明显异常值，并且在一定时间内出现大幅的下降，因此我们需进行删除不相关数据与缺失补全的操作。在删除不相关数据时，我们考虑到寒暑假期间在校人数少，因此，寒暑假期间的数据没有代表性，会造成我们后续建模出现的较大误差，因此我们将寒暑假期间的所有数据全部删除，完成了不相关数据的删除。在处理异常值时，我们首先使用MATLAB软件画出了箱线图，得到了离群值。识别出离群值后，我们将离群值视为缺失值，进行缺失补全操作。根据文献[1]我们得到了处理缺失数据的几种方法，分别为：删除法、常量补全法、平均值补全法、K最近距离邻居补全法等。其中，删除法会导致大量数据的丢失，导致整体数据精度大大降低；常量补全法对补全值进行简单估计时，主观性过大，会导致噪声数据的引入，导致数据补全后效果较差； K最近距离邻居补全法在搜索最近的邻居记录时，每个数据都需要遍历整个数据集，而我们拥有的数据集内数据过多，导致效率低下。在经过综合考量后，我们选择了平均值补全法对缺失数据进行补全。由于附件4-7的数据中存在“待定”,“一层待定”这样的不确定的值，因此我们需要进行缺失补全工作。此外，附件4-7的数据存在部分空白，因此我们依然需要对数据进行缺失补全操作。由于我们需要补全的数据为文本数据，无法确定平均值，所以我们选择用最大频数代替平均值，即使用最大频率补全法对数据进行处理。根据文献[2]可知，由已知数据体现出的消费行为规律是由少部分消费行为属性决定的，结合附件0与实际情况，我们知道消费行为受性别影响。对照附件1-7，我们要将吃饭次数和消费金额作为属性纳入考虑范围，并且吃早餐、吃午餐、吃

晚餐是三次不同的消费行为属性。对于时间的选择，考虑到以日均为选择，数据体量过大，效率降低，以月均为选择，数据过少，研究性低，故选用周为计量单位。基于上述原因，我们选择出若干个特征变量。至此，数据预处理工作告一段落。

第二小问为分类问题，要求我们根据处理后的数据建立模型，挖掘不同代表性群体，分析该群体三年消费行为变化规律。文献[3]中，针对不同学生的学习成绩，用层次聚类法和K − $Means$均值聚类法进行聚类分析，按照分数将学生进行了分类，与本文研究内容方向相似。结合文献[4]，K − $Means$聚类算法优点为收敛快、适应性强、可以灵活选择距离度量方式等，缺点为K值选择难度较大及离群值对聚类结果影响显著。鉴于我们的聚类目的为将学生的消费行为特征分为三类，故取K值为3，且在上述数据预处理时，我们已经对离群值进行处理，极大程度消除了离群值对聚类结果产生的影响，接着考虑到聚类效率问题，我们最终选择K − $Means$均值聚类法对学生的消费行为特征分类，得出聚类结果。由聚类结果数据，以第一类的男生与第二类女生为例，用Excel分别挑选出所有第一类的男生和第二类的女生，找到距离类中心最近的一个记为该类的代表，分析两位代表。首先，我们做出两位代表的周总消费金额趋势图，通过具体分析代表的主要消费行为特征变化规律，可以得到该类整体的主要消费行为特征变化规律。同理，分别得到第二类男生、第三类男生、第一类女生、第三类女生的主要消费行为特征变化规律。

根据附件4-7的数据，我们可以得出总体的饮食种类共有167类。为将饮食种类进行分类，我们将价格作为分类的主要影响因素。首先，我们按照食物的价格对食物种类进行排序，找到出现明显变化的点，记为价位分界点。其次，根据分界点，将167种食物分为不同类别。接着，在第二小问的六个分类的基础上，对每个分类的代表的饮食种类进行分析，得出若干规律。

2.2 问题分析2

针对第二问，由两个小问组成。

第一小问属于预测问题，要求我们根据附件1-3与附件8数据建立模型，预测附件9中学生的贫困程度。文献[5]建立了BP神经网络算法的高校贫困生预测模型，此文献是利用学生每月消费次数，每月消费金额，月早午晚餐次数和金额、食堂消费，超市消费，其他类型消费等数据对高校学生的贫困程度进行了预测，与我们的研究方向相似，且我们拥有的数据体量庞大，能够较好地训练数据。再结合文献[6]，我们了解到RBF神经网络具有在任意精度下逼近任意非线性映射的能力，并且能够达到最佳逼近精度。其缺点是相比bf神经网络，运算速度略有降低，考虑到高校贫困生判定工作的重要程度，我们最终选择RBF神经网络作为我们的预测模型。根据附件8中的已知数据进行训练数据，并且完成误差分析。针对附件9中的1000个待填登记数据，我们首先提取出这一千名学生的序号，并通过前面处理过的数据，得出每名学生各自的十个特征值，使用MATLAB软件计算出他们的贫困程度隐形认定等级，补全了附件9。

第二小问也属于预测问题，要求我们预测该组同学第二、第三年的贫困程度隐形认定等级，分析相关变化。首先，我们根据附件2，即第二年的数据，使用MATLAB软件计算出第二年该组同学各自的十个特征值，计算出他们的贫困程度隐形认定等级。其次，我们根据附件3，即第三年的数据，使用MATLAB软件计算出第三年该组同学各自的十个特征值，计算出他们的贫困程度隐形认定等级。最后，绘制可视化图，计算分析出若干变化。

2.3 问题分析3

第三问属于预测问题，要求我们结合附件4-7的饮食种类数据，改进模型，并分析相关同学的预测结果变化情况。文献[7]指出K − $Means$算法是一种基于误差平方和准则的算法，同时也是一种基于样本间相似性度量的间接聚类方法，属于非监督学习方法。因此我们考虑将K − $Means$聚类

与RBF神经网络算法相结合。由于用K − $Means$算法训练神经网络的方法过于困难，我们将该方法简化为用K−$Means$聚类和RBF神经网络算法分别计算学生的贫困等级，按照1：1的比例综合结果，得到该组同学最终的贫困程度隐形认定等级。首先，考虑到要结合附件4-7的饮食种类数据，而我们通过第一问将饮食种类分为低中高三等级，我们通过对附件4-7中的每位学生的饮食等级做归一化处理，得到新的三个特征值，即用MATLAB软件分别计算每位学生吃高价位、中价位、低价位食物占他们总刷卡次数的比例，作为三个新的特征值，与第一问得到的十个特征值组合，用第一年每名学生的十三个特征向量完成对RBF神经网络的训练与误差分析。接着，使用SPSS软件将第一年每名学生的十三个特征向量数据进行标准化处理，处理后的数据用K − $Means$均值聚类法，得到聚类结果。然后，将K − $Means$聚类结果与RBF神经网络算出的结果按照1:1的比例综合结果，得到第一年最终的贫困程度隐形认定等级。最后，用相同的方法处理第二年与第三年的数据，得到第二年与第三学年该组同学的贫困程度隐形认定等级，根据由K − $Means$算法与RBF神经网络预测模型算出的结果中，同一学生三年的贫困程度隐形认定等级分析，得出若干相关结论。

2.4 问题分析4

第四问属于预测问题，要求我们出细粒度资助额度分配算法，并以第三学年为例，对附件4-7涉及的学生进行总额10万元，总资助人员80人的分配并讨论算法的公平合理性。由于以第三学年为例，对附件4-7涉及的学生，因此，我们对第三学年附件4-7涉及的学生数据导出，先将贫困等级为0的学生剔除，并且将周消费低于15元的学生默认记为不在学校吃饭，也剔除。其次，为使每位学生受补助后的年消费水平相似，列举出了目标函数与四个约束条件。然后，得到了具体的分配额度，具体数据见支撑材料所示。最后，观察学生补助后的年消费水平，发现补助后学生的年消费水平非常统一，证明了上述算法的公平性与合理性。

3 模型假设

1、

2、

3、

4、

5、

4 定义与符号说明

符号定义	符号说明
C_i	第i个初始聚类中心
w_{ji}	数据点关于第i个聚类的权重
C_j	更新得到的第j个聚类中心
$G\left(x^i, c^j\right)$	隐层第i个单元的输出
c^j	隐含层第j 个节点的数据中心
σ^j	RBF 函数的宽度
$\|x^i - c^j\|$	x^i 到 c^j 的欧式距离
ω_{jk}	隐含层第j 个节点到输出层第 k个节点之间的权值

5　模型的建立与求解

5.1　数据的预处理

5.1.1　对附件1-3的处理

对附件1-3的数据，我们用Matlab软件画出了学生的三年三餐消费数据图，如下所示：

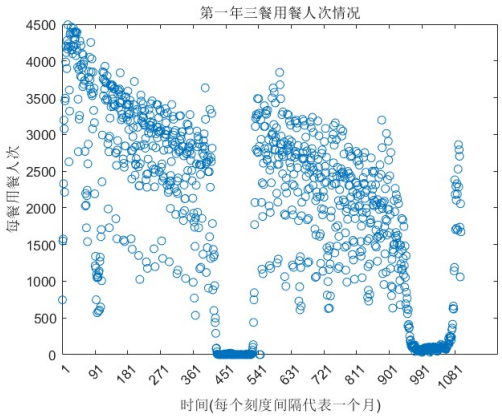

图 1: 第一年三餐用餐人次情况

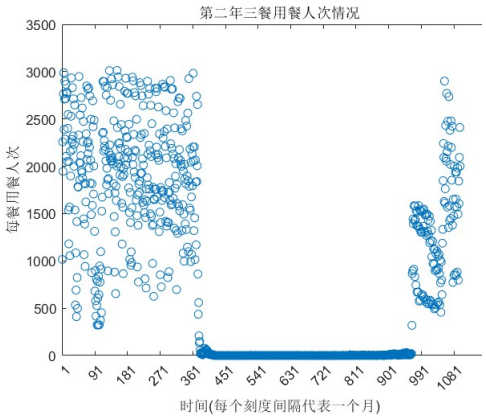

图 2: 第二年三餐用餐人次情况

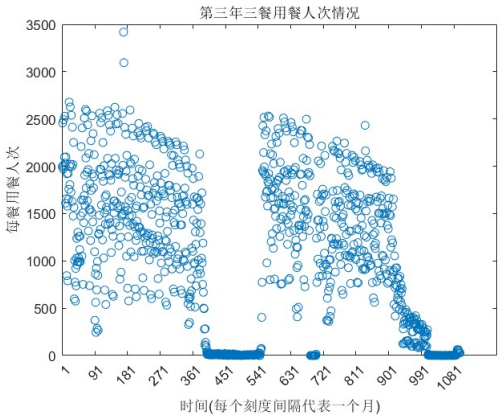

图 3: 第三年三餐用餐人次情况

因此我们使用Matlab软件，计算出本次用餐人数与上一次用餐人数相比的下降值最大，且与之后用餐人数差距小的点作为放假的时间节点。同理计算出本次用餐人数与上一次用餐人数相比的上升值最大，且与之后用餐人数差距小的点作为开学的时间节点。得到假期为：2019年1月19日至2019年2月23日、2019年7月16日至2019年8月7日、2020年1月10日至2020年7月18日、2021年1月11日至2021年3月2日、2021年8月2日至2021年8月31日，将此短时间内数据删除。需要注意的是，2021年4月16日至2021年4月21日在校人数极少，我们猜测学校在这段时间放假，因此也将数据删除。特别地，考虑到2020年疫情爆发，大部分学校学生不满足返校条件，致使学校学生极少，故将2020年1月10日至2020年7月18日数据全部删除。至此，我们删除了所有不相关数据。接着，我们再次使用Matlab软件对数据的异常值进行处理。首先画出第一年序号为1的同学的早餐金额箱线图，结果如下图所示：

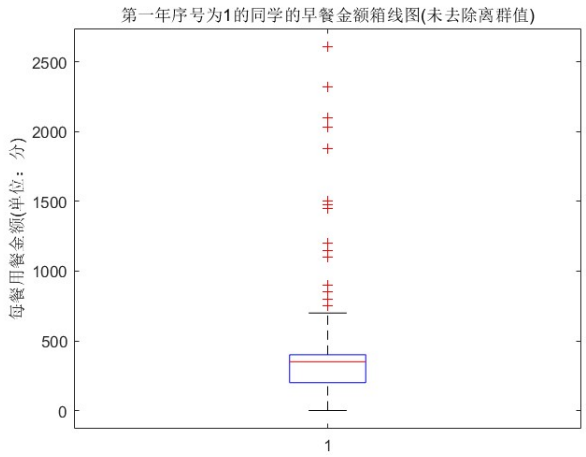

图 4: 未去除离群值箱线图

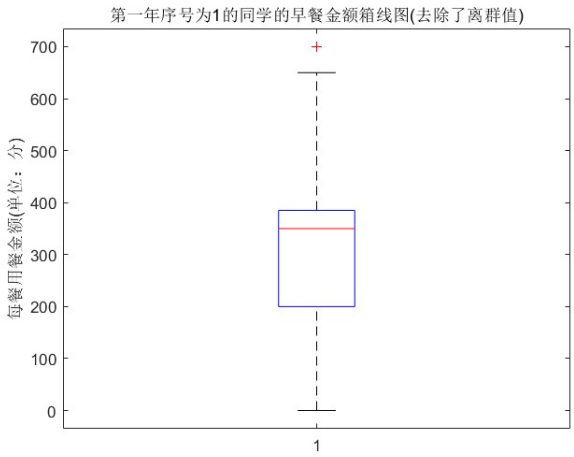

图 5: 已去除离群值箱线图

据上图，我们得到了异常值部分即上边缘上方数据。对于异常值，我们的处理方法为：首先将异常值视为缺失值，用均值补全的方法进行处理。即将已知数据分为若干组，并且分别计算各组的均值，对存在缺失数据的组，用这组的非缺失数据的均值来填充该缺失值。由上述操作完成了对附件1-3数据的处理，得到了第一年三餐消费数据修改、第一年三餐消费数据修改、第一年三餐消费数据修改，具体数据见支撑材料。

5.1.2 对附件4-7的处理

观察附件4-7，我们发现食物种类中存在“待定”、“一层待定”这样的不确定的值，我们将其看作为缺失值。同时数据本身也存在一些缺失值，因此我们对所有缺失值运用最大频率补全法进行缺失补全工作。首先，我们根据缺失值的价格，对整个数据分别进行分类，将同价位的所有食物种类列举在一起。其次，找到同价位的食物种类中出现频数最高的食物，并用其补全缺失值。至此，我们完成了对附件4-附件7的数据的处理，得到了1-100000改、100001-200000改、200001-300000改、300001-331258改，具体数据见支撑材料。

5.1.3 特征提取

根据文献[2]可知，由已知数据体现出的消费行为规律是由少部分消费行为属性决定的，结合附件0与实际情况，我们知道消费行为受性别影响。因此性别为我们选出的第一个特征变量。对照附件1-7，我们要将吃饭次数和消费金额作为属性纳入考虑范围。我们认为学生一天内的吃饭次数是变化值，故吃早餐、吃午餐、吃晚餐是三次不同的消费行为属性。对于时间的选择，考虑到以日均为选择，数据体量过大，效率降低，以月均为选择，数据过少，研究性低，故选用周为计量单位。我们判断每位学生每天的吃饭次数和其每顿饭的消费总金额之间无对应关系，故分开考虑。基于上述原因，我们选择出以下10个特征变量：性别、周早餐消费总金额、周午餐消费总金额、周晚餐消费总金额、周早餐次数、周午餐次数、周晚餐次数、周总消费金额、周总吃饭次数、日均消费金额。至此，我们完成了对所有数据的特征提取。

5.2 问题1第2小问的K − $Means$聚类模型建立与求解

5.2.1 K − $Means$聚类模型的建立

我们需要解决的问题是根据上述处理后的数据，建立模型，挖掘不同代表性群体，并且定量分析该群体三学年的主要消费行为特征变化规律。K − $Means$聚类算法优点为收敛快、适应性强、可以灵活选择距离度量方式等. 缺点为K值选择难度较大及离群值对聚类结果影响显著，但我们可以固定K值，并且通过数据的预处理，我们已经消除了离群值的影响。综合考虑以上所有因素，我们选择建立K − $Means$聚类模型。 K − $Means$聚类算法的工作流程如下图6

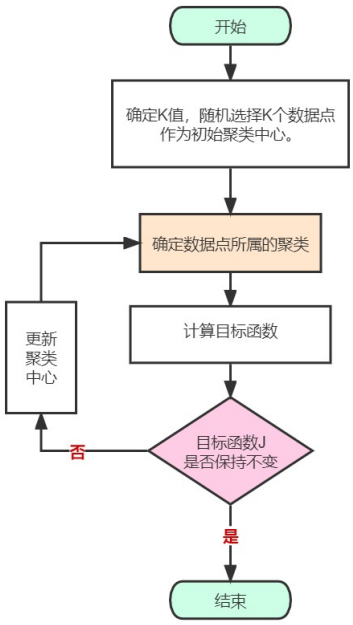

图 6: K − $Means$聚类算法的工作流程

具体步骤如下：

步骤一，在所有数据点中，随机选择K个数据点$C_i, i = 1, 2, 3 \cdots, k$,作为各聚类的初始聚类中心点。

步骤二，确定各数据点所属的聚类，如果数据点X_j被判定属于第i个聚类，则权重值$w_{ji} = 1$;否则为0.其中

$$\sum_{i=1}^{k} w_{ji} = 1, \forall_j = 1, 2, \cdots, n, \sum_{i=1}^{k}\sum_{j=1}^{n} w_{ji} = n, \tag{1}$$

$$w_{ji} = \begin{cases} 1, \|X_j - C_i\| \leq \|X_j - C_m\|, \forall m \neq j \\ 0, \text{ 其他.} \end{cases} \tag{2}$$

步骤三，计算目标函数J, 如果J保持不变，则代表聚类结果已经稳定，迭代结束，否则说明聚类结果非最优，进入下一步。目标函数J计算公式如下：

$$J = \sum_{i=1}^{k} J_i = \sum_{i=1}^{k}\sum_{j=1}^{n} w_{ji} \|X_j - C_i\|^2 \tag{3}$$

步骤四，以$C_j = \frac{\sum_{j=1}^{n} w_{ji} X_j}{\sum_{j=1}^{n} w_{ji}}$为条件更新聚类中心点，回到步骤二。

5.2.2 K − *Means*聚类模型的求解

首先，根据我们所选取的特征使用Matlab软件计算每个特征变量的特征值，特别地，我们将男生与女生的数据分开进行计算。（具体数据见支撑材料），大部分特征变量的特征值可以由附件0-7直接得到，一些特殊特征变量的特征值的计算方式如下所示：

1、 周早餐消费总金额指一周内学生吃早餐的总金额和吃早餐次数的比值。同理可以得到周午餐消费总金额与周晚餐消费总金额的数据。

2、 日均消费金额指某一段时间内学生在校总消费与消费天数的比值。

其次，使用SPSS软件对上述数据进行带入求解，聚类结果。其中，SPSS软件的计算分别得出对男生和女生分别聚类得到的每个聚类中的个案数目，如下所示：

每个聚类中的个案数目

聚类	1	1403.000
	2	954.000
	3	1585.000
有效		3942.000
缺失		.000

图 7: 男生每个聚类的个案数目

每个聚类中的个案数目

聚类	1	466.000
	2	202.000
	3	535.000
有效		1203.000
缺失		.000

图 8: 女生每个聚类的个案数目

最后，为将聚类效果可视化呈现，我们以数据中前一百个女生为例，展示聚类结果如下图所示：

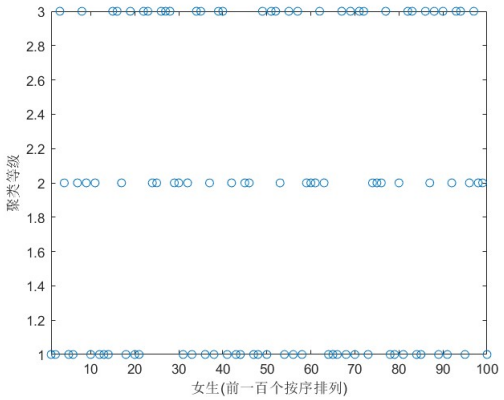

图 9: 前一百个女生的聚类效果

5.2.3 结果

由聚类结果数据首先，用Excel将每一类的成员分开，找到每类中距离类中心最近的一个记为该类的代表。其中男生第一类的代表在附件1-3中序号为2，男生第二类的代表在附件1-3中序号为3，男生第三类的代表在附件1-3中序号为5，女生第一类的代表在附件1-3中序号为1，女生第二类的代表在附件1-3中序号为11，女生第三类的代表在附件1-3中序号为10。接着，我们分别做出这六个人的周总消费金额趋势图，具体周总消费金额趋势图见附录。

这里我们选择第一类男生周消费金额变化趋势、第二类女生周消费金额变化趋势，如下：

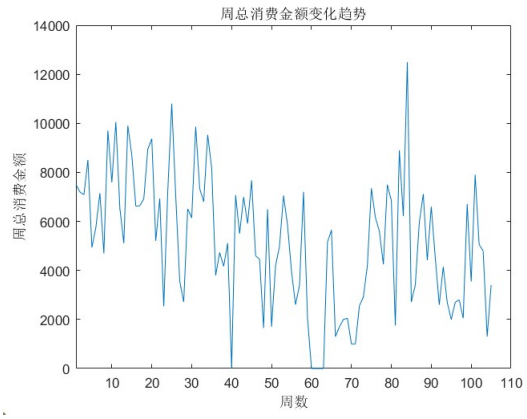

图 10: 第一类男生周消费金额变化趋势

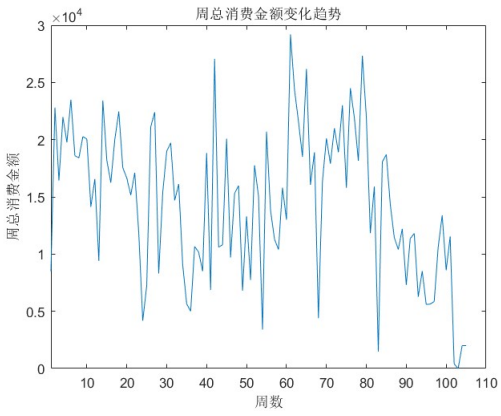

图 11: 第二类女生周消费金额变化趋势

根据上面的周消费金额变化趋势图，可以得到以下结论：第一类男生代表的周消费金额主要在20到100元之间波动，第二类女生代表的周消费金额在100到250元之间波动，可见第一类男生消费能力较弱，第二类女生消费能力较强。且第一类男生消费变化较小，第二类女生消费变化略大。

针对六类学生分析得：

我们分别计算了六类代表的平均数为5176、8315、2416、2478、14495、6925；最大值为12497、23140、13602、12500、29200、17400；总和为543481、873091、253639、260228、1521984、727166。男生第一类和男生第二类消费基本稳定，男生第二类消费总金额更高，而男生第三类周总消费金额前期波动后期逐渐下降，可能的原因是该类男生经常点外卖或者在校外吃饭。女生第二类和女生第三类消费水平基本稳定，而女生第二类消费总金额更高。可以得出结论：第三类男生为男生整体中最需要资助的，第二类男生为最不需要资助的；第一类女生为最需要资助的，第二类女生为最不需要资助的。

5.3 问题1第3小问的求解

第三小问为分类问题，要求我们根据处理后的数据挖掘出不同代表性群体，并且定量分析该群体三年的饮食种类变化规律。考虑到对食物种类进行分类时，价格因素是最大的影响因素，并且在一定程度上可以反映出学生的贫困程度，因此我们选择由价格作为分类导向。

步骤一，我们根据附件4-7的数据，用Excel导出了总体的饮食种类，共为167类。

步骤二，我们将饮食种类按照价格升序进行排列，使用SPSS软件，对食物种类与价格的关系画出图像，如下所示：

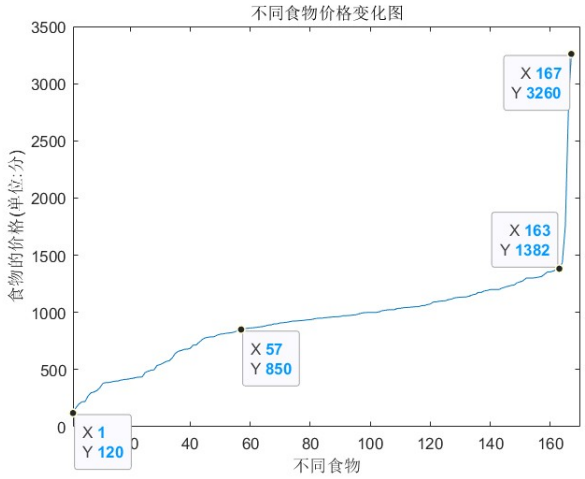

图 12: 不同食物价格变化图

由上图可以看出，饮食种类按照价格的分布有两个明显的变化点，即为图中标注的第57类与第163类，将这两个变化点记为价位分界点。由此得到分类：第1-56种食品记为第一类，第57-163种食物记为第二类，第164-167类食物记为第三类。第一类食物的价格范围为1.2元至8.5元，属于低价位；第二类食物的价格范围为8.5元至13.82元，属于中价位；第三类食物的价格范围为13.82元至32.6元，属于高价位。

步骤三，由第二小问知，所有学生被分为了六类，我们在这六类学生中分别找到了一名代表。需要注意的是，由于附件4-7提供的数据不是完整的、包含每个学生刷卡次数的数据，因此我们对代表每一类的学生进行了重新选择，并非第二问选择的。为了达到找到代表的目的，我们将附件4-7中的学生序号导出，按照每类中距离类中心最近的一个记为该类的新代表。操作得到，男生第一类的代表在附件1-3中序号为3062，男生第二类的代表在附件1-3中序号为76，男生第三类的代表在附件1-3中序号为3207，女生第一类的代表在附件1-3中序号为1275，女生第二类的代表在附件1-3中序号为5001，女生第三类的代表在附件1-3中序号为3272。

步骤四，根据附件4-7的数据，我们得到每类代表的刷卡记录属于饮食种类分类中的类数，使用SPSS软件分别做出六个图，将结果可视化，具体图见附录，在此只列举第一类男生与第三类女生的饮食种类变化趋势，如下：

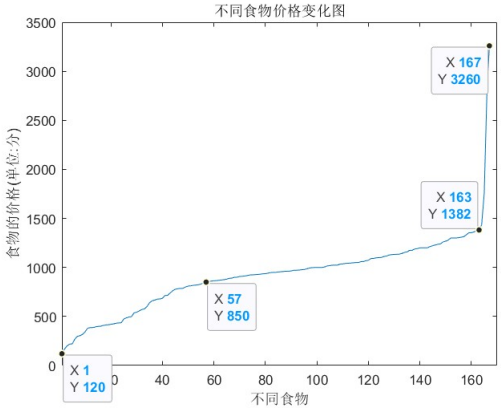

图 13: 第一类男生饮食种类图

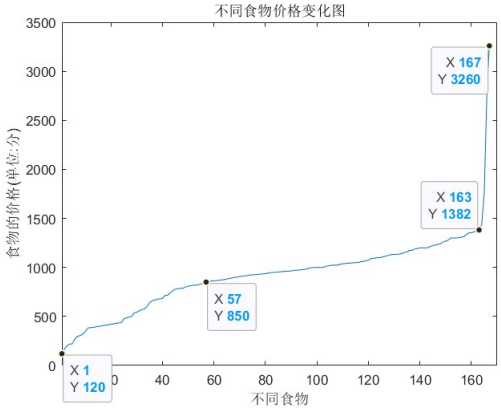

图 14: 第三类女生饮食种类图

由上图可以看出，第一类男生极少吃高价位食物，多数时间吃中价位与低价位食物。第三类女生吃高价位食物的次数明显多于第一类男生，可见第三类女生在食堂的消费能力比第一类男生较

高。且第三类女生吃中价位食物的次数也多于第一类男生。

步骤五，分别计算六位学生代表吃高价位、中价位、低价位食物占他们总刷卡次数的比例，结果见下表：

表 1: 消费占比

代表	低价位消费占比	中价位消费占比	高价位消费占比
男生第一类	66.88103%	32.47588%	0.64309%
男生第二类	82.43752%	16.29579%	1.26669%
男生第三类	60.57392%	39.23893%	0.187149%
女生第一类	87.82743%	11.86441%	0.30817%
女生第二类	82.18479%	17.71312%	0.10209%
女生第三类	74.96839%	23.13527%	1.89633%

基于上表可以清楚看出我们所分的六类的学生间的饮食种类差距较大。男生第一类、男生第二类与男生第三类饮食种类都以低价位和中价位食物为主，相比之下，男生第三类高价位食物比男生第二类吃的更多，男生第一类的刷卡次数远低于男生第二类和男生第二类，可能经常不在学校吃饭；女生第一类、女生第二类和女生第三类饮食种类都以低价位和中价位食物为主，且刷卡次数相近，相比之下，女生第二类中价食物吃的更少，更有可能是贫困生。

5.4 问题2的基于RBF神经网络算法的预测模型建立与求解

5.4.1 基于RBF神经网络算法的预测模型的建立

我们需要解决的问题是根据附件1-3与附件8数据建立预测模型，对附件9中学生的贫困程度进行预测。

首先，我们对上一问得到的特征值进行观察，发现有394人的特征值为0，即判断出这些人不在学校吃饭，因而无法进行贫困程度的判定，故记为贫困程度为0。

其次，根据附件9中数据，发现需要预测1000位同学的贫困程度，再看附件8，发现我们已知4415位同学的贫困程度，剔除394位已知的贫困程度为0的同学后，我们已知判定等级的学生有4097人。考虑到基于RBF神经网络算法的预测模型需要大量数据对神经网络进行训练，因此我们先用4000个数据作为训练数据集，剩余97个数据用于模型的误差分析检验。

需要特殊解释的是：

RBF网络中, 隐层最常用的基函数是高斯函数

$$G\left(X,T_i\right)=G\left(\left\|X-T_i\right\|\right)=\exp\left(-\frac{\left\|X-T_i\right\|^2}{2\sigma_i^2)}\right)\quad i=1,2,\cdots,M$$

其中$G\left(X,T_i\right)$为隐层第i个单元的输出,X为p维输入矢量,T_i为隐层第i个单元高斯函数的中心,σ_i为第i个隐节点的归一化参数,即宽度,M为隐层节点数.该网络中需要学习的参数有三类,即RBF的中心T_i宽度σ_i和连接权重Wij.

为达到尽量精确的数据，设置误差允许值为0。

具体步骤如下所示：

第一步，假设训练数据集$X=\{x^1,x^2,x^3,\cdots,x^m\}$, 其中第$i$个训练样本为$x^i=\{x_1^i,x_2^i,x_3^i,\cdots,x_n^i\}$, 其中$i=(1,2,3,\cdots,m)$, 即样本数量为$m$, 特征数为$n$。如果隐含层节点个数为$s$, 则第$i$个训练样

本x^i所对应的隐含层的第j个节点 $j=(1,2,3\cdots,s)$ 的 RBF 函数, 如式（4）所示:

$$G\left(x^i,c^j\right)=\exp\left(-\frac{\left\|x^i-c^j\right\|^2}{2\left(\sigma^j\right)^2}\right) \tag{4}$$

其中 c^j 为隐含层第j 个节点的数据中心, σ^j 为该 RBF 函数的宽度, $\|x^i-c^j\|$为 x^i 到 c^j 的欧式距离。

第二步, 如果输出层节点的个数为 t, 输出层对隐含层输出的节点应用线性函数, 如式(5) 所示:

$$y_k=\sum_{i=1}^{s}\omega_{jk}G\left(x^i,c^j\right) \tag{5}$$

其中 $k=(1,2,3,\cdots,t),s$ 为隐含层节点个数, ω_{jk}为隐含层第j 个节点到输出层第 k个节点之间的权值。

当迭代次数达到3950次时，MSE=0.0697154，而当迭代次数达到4000次时，MSE=0.0697217.可以看出，当迭代次数超过3950次时，MSE值反而增大，即误差变大，故得到当迭代次数为3950次时，效果最优。第三步，导入97个数据用于模型的误差分析检验，选出部分贫困标签预测值和相对误差作为展示（完整数据见支撑材料），如下：

表 2: 部分贫困标签预测值和相对误差

贫困标签预测值	相对误差	贫困标签预测值	相对误差
0	0.142956691	0	0.295073108
0	0.125753133	0	0.001609297
1	0.759155242	0	0.282101481
2	0.042975555	0	0.02708873
0	0.095840661	6	2.525666832

由此表可以看出，RBF神经网络固然存在不可避免的误差，但对于这些误差量较小，也就意味着基于RBF神经网络算法的预测模型预测精度较高。

5.4.2 基于RBF神经网络算法的预测模型的求解

首先，将附件9的1000名学生的特征值进行提取，并将得到的数据使用Matlab软件导入进行计算，得出结果即为附件9中学生的贫困等级，。详细数据见附件9。

其次，对该组同学第二年的特征值数据导入Matlab，得到该组同学第二年的贫困程度隐形认定等级，这里选出部分第二年预测结果作为展示（完整数据见支撑材料），如下表：

表 3: 第二年预测结果（部分）

序号	预测结果	序号	预测结果
1	1	4	1
10	0	11	1
13	3	19	1
21	1	23	0
28	1	29	1
30	2	33	2
37	2	39	2
42	0	46	2

最后，对该组同学第二年的特征值数据导入Matlab，得到该组同学第二年的贫困程度隐形认定等级，这里选出部分第二年预测结果作为展示（完整数据见支撑材料），如下表：

表 4: 第三年预测结果（部分）

序号	预测结果	序号	预测结果
1	0	4	0
10	1	11	1
13	1	19	0
21	0	23	1
28	2	29	1
30	1	33	0
37	1	39	2
42	2	46	0

5.4.3 相关变化分析

根据上面列出的表格中数据，列举前八个学生的预测结果三年的变化，得到下图：

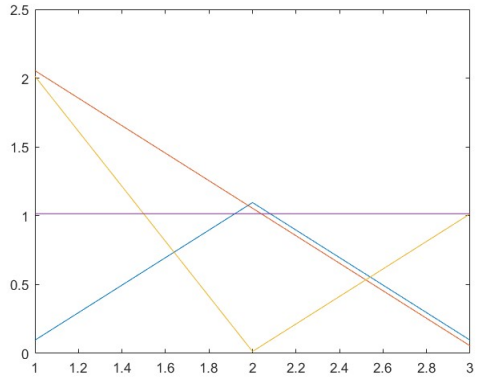

图 15: 1-4同学的贫困程度变化图

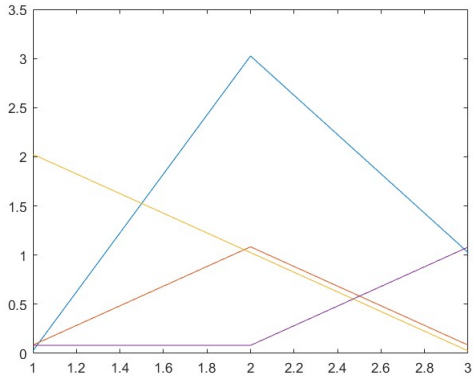

图 16: 5-8同学的贫困程度变化图

在本图中，由于部分图像出现重叠情况，导致可视化程度降低，因此我们对部分折线进行了轻微的移动。

对数据和可视化图进行综合分析，我们得到以下结论：

在第一学年享受了补助的学生，有63%的学生在第二学年的贫困等级降级了，有18%的学生在第二学年的贫困等级不变，有18%的学生在第二学年的贫困等级升级了，在第一学年未享受补助的学生，有83%的学生在第二学年的贫困等级升级了，仅有16%的学生在第二学年的贫困等级未发生改变。在第二年享受了补助的学生，有69%的学生在第三学年的贫困等级降级了，有23%的学生在第三学年的贫困等级未发生变化，在第二学年未享受补助的学生，有75%的学生在第三学年的贫困等级升级了。由此可见，对大学生进行贫困资助是非常有必要并且有用的，由少数学生反贫可知，对于贫困学生的界定还需要更加准确，才不至于漏选贫困学生而使他们更加贫困。

5.5 问题3的基于K − $Means$和RBF神经网络算法的综合预测模型的建立与求解

5.5.1 基于K − $Means$和RBF神经网络算法的综合预测模型的建立

K − $Means$聚类存在一定误差，且RBF神经网络算法也脱离不了误差的产生。综合考虑，我们建立了基于K − $Means$和RBF神经网络算法的综合预测模型。

具体操作步骤如下：

第一步，重新整理构建数据。首先，用MATLAB软件分别计算每位学生吃高价位、中价位、低价位食物占他们总刷卡次数的比例，作为三个新的特征值，与第一问得到的十个特征值组合，得到共13个特征值，并使用SPSS软件，对所有学生的这13个特征值进行归一化处理。

第二步，用前文使用过的K − $Means$算法对第一步处理好的数据进行聚类，得到所有学生第一年的贫困程度隐形认定等级聚类结果。

第三步，在附件4-7中共涉及301人，其中在附件8中标注出贫困标签的人是253个，48个人没有贫困标签。因此，我们在第一步处理好的数据中，选择出了200个人的13个特征值作为训练数据，剩下53个人的数据作为检验数据。

第四步，我们设置误差允许值为0，将数据带入RBF神经网络算法的预测模型中进行运算。

第五步，观察发现，当迭代次数达到200次时，基本满足误差允许值，迭代停止。进入误差分析检验，选出部分贫困标签预测值和相对误差作为展示（完整数据见支撑材料），如下：

表 5: 部分贫困标签预测值和相对误差

序号	贫困标签预测值	相对误差	序号	贫困标签预测值	相对误差
5116	0	0.002759228	5119	0	0.003671699
5120	0	0.012180366	5131	0	0.026858641
5135	1	0.096381754	5143	0	0.254612964
5138	2	0.033952538	5136	0	0.02708873
5151	1	0.095229563	5155	1	0.117416376

由上表格可以看出，训练后的RBF神经网络存在误差，但依然能在一定程度上保证预测精度。

第六步，按照1：1的比例对K − $Means$聚类得到的贫困等级与RBF神经网络算法预测出的贫困等级进行综合，得到最后的贫困程度隐形认定等级。

5.5.2 基于K − $Means$和RBF神经网络算法的综合预测模型的求解

由上述步骤进行预测后，得到如下结果：

表 6: 第一年预测结果（前50）

序号	预测结果	序号	预测结果	序号	预测结果	序号	预测结果
5111	1	5112	1	5113	2	5114	1
5115	1	5116	1	5117	1	5118	1
5119	1	5120	1	5121	1	5122	0
5123	1	5124	2	5125	1	5126	1
5128	2	5129	1	5130	0	5131	1
5132	1	5133	1	5134	2	5135	1
5136	1	5137	1	5138	2	5140	1
5141	3	5145	2	5146	2	5147	1
5148	1	5142	2	5143	1	5144	2
5149	0	5150	1	5151	1	5152	1

用相同的办法对第二年学生的贫困程度隐形认定等级进行计算，得到下列结果：

表 7: 第二年预测结果（前50）

序号	预测结果	序号	预测结果	序号	预测结果	序号	预测结果
76	1	1275	1	2434	1	2932	2
3012	1	3062	1	3186	1	3207	2
3239	2	3272	1	3287	1	3321	1
3340	1	3347	2	3366	2	3432	2
3509	3	3814	1	3985	1	4332	1
4332	1	4439	0	4609	2	4649	2
4890	1	4891	1	4892	1	4893	2
4894	2	4895	2	4896	2	4897	1
4898	2	4899	0	4900	1	4901	3
4902	1	4903	2	4904	1	4905	1

再用相同的办法对第二年学生的贫困程度隐形认定等级进行计算，得到下列结果：

表 8: 第三年预测结果（前50）

序号	预测结果	序号	预测结果	序号	预测结果	序号	预测结果
1275	1	2434	1	2738	1	2932	1
3012	2	3062	1	3186	1	3207	0
3239	1	3272	1	3287	0	3321	0
3340	1	3347	1	3366	2	3432	2
3509	0	3814	1	3985	1	4322	1
4439	2	4609	2	4809	2	4629	2
4649	2	4690	1	4890	1	4891	1
4892	1	4893	1	4894	2	4895	2
4896	1	4897	1	4898	2	4899	2
4900	1	4901	1	4902	1	4903	1

5.5.3 相关变化分析

5.6 问题4的规划模型的建立与求解

5.6.1 规划模型的建立

这是一个补助金分配规划问题，我们构建细粒度资助额度分配算法，意在为附件4-7中涉及的学生进行资助总额10万、资助人员80名的额度分配。设x_i为分配给每个学生的资助金额，c_i为每个学生的年消费水平总额。 S为所有学生一年的可消费总额，包括总资助金额和学生的年消费水平总额。目标函数是使每个学生的年消费水平总额与被资助金额之和的方差最小。，即

$$\min \quad \sum_{i=1}^{80}((x_i+c_i)-\frac{S}{80})^2\frac{1}{80}$$

约束条件分为四类。 1.总资助金额的约束，即

$$\sum_{i=1}^{80}x_i=10^5$$

2.平均水平的约束，即

$$x_i+c_i-\frac{1}{80}S\geqslant 0$$

3.资助金额的约束，即

$$0 < x_i < 10^4$$

4.所有学生一年的可消费总额的约束，即

$$S = \sum_{i=1}^{80} c_i + 10^5$$

综合上述内容，得到以下规划模型。

$$\min \quad (\overline{x_i + c_i}) \tag{6}$$

$$\begin{aligned} s.t. \quad & \sum_{i=1}^{80} x_i = 10^5 \\ & x_i + c_i - \tfrac{1}{80} S \geqslant 0 \\ & 0 < x_i < 10^4 \\ & S = \sum_{i=1}^{80} c_i + 10^5 \end{aligned} \tag{7}$$

其中，$i = (1, 2, 3, \cdots, n)$。

5.6.2 规划模型的求解与公平合理性评估

使用MATLAB软件计算得到如下结果：

表 9: 补助后年消费水平结果

序号	年消费水平（补助前）	补助金	年消费水平（补助后）
3012	10302.81	19272	29574.81
3366	19756.52	9819	29575.52
3432	7478.59	22097	29575.59
4439	13101.55	16474	29575.55
4649	19489.61	10086	29575.61
4894	5633.18	23942	29575.18
4895	7150.5	22425	29575.5
4899	9516.92	20058	29574.92
4918	6528.06	23047	29575.06
4921	8290.17	21285	29575.17
4923	24721.06	4854	29575.06
4927	15023.4	14552	29575.40
4929	24970.54	4605	29575.54
4944	6416.83	23158	29574.83

根据本算法，我们的目的是让不同贫困程度隐形认定等级的学生在被补助后，得到较为近似的年消费水平。未达到这一目的，我们给不同的同学进行不同额度的资助，满足了细粒度的要求。观察上表第四列，即学生补助后的年消费水平，发现补助后学生的年消费水平非常统一，证明得出了上述算法的公平性与合理性。

6 模型的评价及推广

6.1 误差分析

6.1.1 针对于K − $Means$聚类模型的误差分析

6.1.2 针对于基于RBF神经网络算法的预测模型的误差分析

6.1.3 针对基于K − $Means$和RBF神经网络算法的综合预测模型的误差分析

6.1.4 针对于规划模型的误差分析

6.2 模型优点

1、

2、 基于RBF神经网络算法的预测模型在学习了大量数据后，模型的计算结果准确，精度较高。

3、 基于K − $Means$和RBF神经网络算法的综合预测模型具有较强的创新型。

4、

6.3 模型缺点

1、 基于K − $Means$和RBF神经网络算法的综合预测模型在进行最终的结果预测时，选用1:1的比例，体现了决策者的主观性，过于主观化。

2、

6.4 模型的推广

1、

2、

3、

参考文献

[1] 沈雪.基于贝叶斯方法的缺失数据补全研究【D】.重庆大学,2011(01).

[2] 任志愿.基于校园一卡通消费数据的学生行为分析研究与应用【D】.电子科技大学,2020(07).

[3] 肖铮,聚类分析在学生群体分类中的应用研究,,山东工业技术,2020,42-46,42-46

[4] 刘亮;许灵;刘斯文,基于K-Means聚类的高校困难学生贫困等级划分研究—— 以蚌埠学院为例,,白城师范学院学报,2017,38-41+64,38-41+64

[5] 温上海.基于神经网络算法的高校贫困生预测模型研究【J】.网络安全技术与应用,2021,No.244(04):36-37.

[6] 党开放,杨利彪,林廷圻.一种新型的广义RBF神经网络及其训练方法【J】.计算技术与自动化,2007,No.101(01):9-13.

[7] 庞振,徐蔚鸿.一种基于改进k-means的RBF神经网络学习方法【J】.计算机工程与应用,2012,48(11):161-163+184.

附录 A		
支撑文件清单		
文 件 夹名	文件名	含义
数据	1-100000 改.xlsx	问题一对附件 4 的预处理结果
	100001-200000 改.xlsx	问题一对附件 5 的预处理结果
	200001-300000 改.xlsx	问题一对附件 6 的预处理结果
	300001-331258 改.xlsx	问题一对附件 7 的预处理结果
	kmeanRbf 方法确定的贫困标签 1.xlsx	问题三预测的学生第一年贫困程度
	kmeanRbf 方法确定的贫困标签 2.xlsx	问题三预测的学生第二年贫困程度
	kmeanRbf 方法确定的贫困标签 3.xlsx	问题三预测的学生第三年贫困程度
	kmeans 聚类结果男总.htm	问题一基于”特征提取男总.xlsx”的进行 kmean 聚类的报告
	kmeans 聚类结果男总.xlsx	问题一基于”特征提取男总.xlsx”的进行 kmean 聚类的结果
	kmeans 聚类结果女总.htm	问题一基于”特征提取女总.xlsx”的进行 kmean 聚类的报告
	kmeans 聚类结果女总.xlsx	问题一基于”特征提取女总.xlsx”的进行 kmean 聚类的结果
	不同食物及其价格.xlsx	问题一对处理后的附件 4-7 的文件进行食物价格统计
	不在食堂吃饭的人 1.xlsx	问题二对第一年的数据进行特征提取时筛除的人
	不在食堂吃饭的人 2.xlsx	问题二对第二年的数据进行特征提取时筛除的人
	不在食堂吃饭的人 3.xlsx	问题二对第三年的数据进行特征提取时筛除的人
	不在食堂吃饭的人总.xlsx	问题二对汇总的三年的数据进行特征提取时筛除的人
	差异化资助情况.xlsx	问题四 80 个人资助金额具体分配情况
	第二年个人贫困程度.xlsx	问题二 rbf 神经网络预测学生第二年贫困程度结果
	第二年三餐消费数据改.xlsx	问题一对附件 2 预处理的结果
	第三年个人贫困程度.xlsx	问题二 rbf 神经网络预测学生第三年贫困程度结果
	第三年三餐消费数据改.xlsx	问题一对附件 3 预处理的结果
	第一年三餐消费数据改.xlsx	问题一对附件 1 预处理的结果
	附件 4-7 出现序号.xlsx	统计的附件 4-7 出现的学生序号
	每日用餐人数统计 1.xlsx	问题一对第一年三餐用餐人次的统计
	每日用餐人数统计 2.xlsx	问题一对第二年三餐用餐人次的统计
	每日用餐人数统计 3.xlsx	问题一对第三年三餐用餐人次的统计
	贫困标签预测值 rbf3_2.xlsx	问题三 rbf 神经网络预测学生第二年贫困程度结果

附录B　提取六个代表性学生的三年的周总消费金额-Consumptionchange1.m

```
clc, clear;
%周总消费金额变化趋势
filename = 数据第一年三餐消费数据改["..\\.xlsx";数据第二年三餐消费数据改
    "..\\.xlsx"; 数据第三年三餐消费数据改"..\\.xlsx"];
range = ["B2:AGG5416"; "B2:ST5416"; "B2:AEW5416"];
%将三年的数据汇总
T = readmatrix(filename(1), 'Range', range(1));
[m, n] = size(T);
for i = 2:3
    t = readmatrix(filename(i), 'Range', range(i));
    [~, nt] = size(t);
    T(:, n + 1:n + nt) = t;
    n = n + nt;
end

xm = readmatrix数据特征提取男总("..\\.xlsx", 'Range', 'A2:J3943');
xw = readmatrix数据特征提取女总("..\\.xlsx", 'Range', 'A2:J1204');

%第一列为序号, 第二列为等级
dm = [1, 1; 11, 2; 10, 3];
dw = [2, 1; 3, 2; 5, 3];
%提取周总消费金额变化趋势
s = zeros(6, n / 21, 1);
for i = 1:6
    if i >= 4
        t = dw(i - 3, 1);
    else
        t = dm(i, 1);
    end

    for j = 1:21:n
        for k = j:j+20
            s(i, ceil(j / 21)) = s(i, ceil(j / 21)) + T(t, k);
        end
        s(i, ceil(j / 21)) = s(i, ceil(j / 21));
    end
end

%绘图
[ms, ns] = size(s);
for i = 1:ms
    figure;
    plot(1:ns, s(i, :));
    xlim([1 110]);
    xlabel周数("");
```

```
    ylabel周总消费金额("");
    title周总消费金额变化趋势("");
end

writematrix(s, 数据周总消费金额变化趋势"..\\.xlsx");
```

附录C　问题一对附件1、2、3进行预处理-dataPer1_123.m

```
clc, clear;
filename = 题目附件第一年三餐消费数据["..\\1.xlsx";题目附件
    "..\\2 第二年三餐消费数据.xlsx";题目附件
    "..\\3 第三年三餐消费数据.xlsx"];

tname = 数据第一年三餐用餐人次情况["..\\";数据第二年三餐用餐人次情况
    "..\"; 数据第三年三餐用餐人次情况"..\"];
%统计每餐用餐人次
for k = 1:3
    T = readcell(filename(k), 'Range', 'A1:APG5416');
    x = cell2mat(T(2:end, 2:end));
    switch k
        case 1
            T1 = T;
        case 2
            T2 = T;
        case 3
            T3 = T;
    end

    [m, n] = size(x);
    c = sum(x > 0);

    a = T(1, 2:end);
    ac = string(zeros(1, numel(a)));
    for i = 1:numel(a)
        ac(i) = cell2mat(a(i));
    end
    ta = table(ac', c');
%     writetable(ta, 数据每日用餐人数统计['..\\', num2str(k), '.xlsx']);

    %     figure;
    %     plot(1:n, c, 'o')
    %     xticks(1:90:n);
    %     xlabel时间每个刻度间隔代表一个月("()");
    %     ylabel每餐用餐人次("");
    %     title(tname(k))
end

s = ["2019.01.19", "2019.02.22", "2019.02.29", "2019.07.15", "2019.08.25", "-";
```

```
    "2020.01.06", "2020.07.18", "-",       "-",         "-",         "-";
    "2021.01.10", "2021.03.02", "2021.04.16", "2021.04.21", "2021.08.01", "2021.08.31"];
lo = zeros(3, 6, 1);
%去除寒暑假的影响
for i = 1:3
    switch i
        case 1
            cp = T1;
        case 2
            cp = T2;
        case 3
            cp = T3;
    end

    for j = 1:5
        for k = 1:1099
            if contains(cp{1, k}, s(i, j))
                lo(i, j) = k;
            end
        end
    end
end
lo = lo - 3;

% , 2018.09.01~2019.1.18  2019.2.23~2019.02.28, ,
            2019.03.01~2019.7.14       2019.08.25~2019.08.31
ct1(2:5416, :) = [T1(2:end, 1:lo(1, 1)), T1(2:end, lo(1, 2) + 4:lo(1, 3)), T1(2:end,
    lo(1, 3) + 4:lo(1, 4)), T1(2:end, lo(1, 5) + 4:end)];
ct1(1, :) =    [T1(1, 1:lo(1, 1)),  T1(1, lo(1, 2) + 4:lo(1, 3)),  T1(1, lo(1, 3) +
    4:lo(1, 4)),  T1(1, lo(1, 5) + 4:end)];

%, 2019.09.01~2020.01.05 2020.07.19~2020.08.31
ct2(2:5416, :) = [T2(2:end, 1:lo(2, 1)), T2(2:end, lo(2, 2) + 4:end)];
ct2(1, :) = [T2(1, 1:lo(2, 1)), T2(1, lo(2, 2) + 4:end)];

%2020.09.01~2021.01.09,  2021.03.03~2021.04.15, 2021.04.22~2021.07.31
ct3(2:5416, :) = [T3(2:end, 1:lo(3, 1)), T3(2:end, lo(3, 2) + 4:lo(3, 3)), T3(2:end,
    lo(3, 4) + 4:lo(3, 5))];
ct3(1, :) = [T3(1, 1:lo(3, 1)), T3(1, lo(3, 2) + 4:lo(3, 3)), T3(1, lo(3, 4) + 4:lo(3,
    5))];


%处理离群值
for p = 1:3
    if p == 1
        tx = cell2mat(ct1(2:end, 2:end));
    elseif p == 2
        tx = cell2mat(ct2(2:end, 2:end));
    else
        tx = cell2mat(ct3(2:end, 2:end));
```

```
end

[mt, nt] = size(tx);
for i = 1:mt
    %统计早餐，午餐，晚餐
    x1 = zeros(nt / 3, 1, 1); x2 = zeros(nt / 3, 1, 1); x3 = zeros(nt / 3, 1, 1);
    k1 = 1;                 k2 = 1;                 k3 = 1;
    for j = 1:nt
        flag = mod(j, 3);
        if flag == 1
            %早餐
            x1(k1, 1) = tx(i, j); k1 = k1 + 1;
        elseif flag == 2
            %午餐
            x2(k2, 1) = tx(i, j); k2 = k2 + 1;
        else
            %晚餐
            x3(k3, 1) = tx(i, j); k3 = k3 + 1;
        end
    end

    %分别对早餐，午餐，晚餐进行离群值处理
    for k = 1:3
        if k == 1
            x = x1;
        elseif k == 2
            x = x2;
        else
            x = x3;
        end
        % 计算分位数的函数需要安装了统计机器学习工具箱MATLAB
        Q1 = prctile(x, 25); % 下四分位数
        Q3 = prctile(x, 75); % 上四分位数
        IQR = Q3 - Q1; % 四分位距
        lb = Q1 - 1.5 * IQR; % 下界
        ub = Q3 + 1.5 * IQR;% 上界
        tmp = (x < lb) | (x > ub);
        ind = find(tmp);
        %将异常值视作缺失值
        x(ind) = nan;

        %处理缺失值用平均值填充,
        avg = 0;   cnt = 0;
        for j = 1:length(x)
            if ~isnan(x(j))
                avg = avg + x(j);
                cnt = cnt + 1;
            end
        end
        avg = round(avg / cnt);
```

```
        for j = 1:length(x)
            if isnan(x(j))
                x(j) = avg;
            end
        end
        % 作图对比

        if k == 1
            fg = x1;
        elseif k == 2
            fg = x2;
        else
            fg = x3;
        end

        %     yyaxis left
        %     plot(1:nt - 1, x);
        %     ylim([0 ub + 1000])
        %     yyaxis right
        %     plot(1:nt - 1, fg);
        %     ylim([0 ub + 1000])

        %         figure;   boxplot(fg);    ylabel每餐用餐金额单位：分("()");
        %         title第一年序号为的同学的早餐金额箱线图未去除离群值("1()");
        %         figure;   boxplot(x);     ylabel每餐用餐金额单位：分("()");
        %         title第一年序号为的同学的早餐金额箱线图去除了离群值("1()");

        if k == 1
            x1 = x;
        elseif k == 2
            x2 = x;
        else
            x3 = x;
        end
    end

    k1 = 1;                 k2 = 1;                 k3 = 1;
    for j = 1:nt
        flag = mod(j, 3);
        if flag == 1
            %早餐
            tx(i, j) = x1(k1, 1); k1 = k1 + 1;
        elseif flag == 2
            %午餐
            tx(i, j) = x2(k2, 1); k2 = k2 + 1;
        else
            %晚餐
            tx(i, j) = x3(k3, 1); k3 = k3 + 1;
        end
    end
```

```
    end

    if p == 1
        ct1(2:end, 2:end) = num2cell(tx);
    elseif p == 2
        ct2(2:end, 2:end) = num2cell(tx);
    else
        ct3(2:end, 2:end) = num2cell(tx);
    end
end


%记录数据
writecell(ct1, 数据第一年三餐消费数据改"..\\.xlsx");
writecell(ct2, 数据第二年三餐消费数据改"..\\.xlsx");
writecell(ct3, 数据第三年三餐消费数据改"..\\.xlsx");
```

附录D　问题一对附件4 5 6 7进行预处理-dataPer1_4567.m

```
clc, clear;
%附件缺失值补全4-7
filename1 = 题目附件["..\\4 1-100000.xlsx"; 题目附件"..\\5 100001-200000.xlsx";题目附件
    "..\\6 200001-300000.xlsx"; 题目附件"..\\7 300001-331258.xlsx"];

filename2 = 数据改["..\\1-100000.xlsx"; 数据改"..\\100001-200000.xlsx";数据改
    "..\\200001-300000.xlsx"; 数据改"..\\300001-331258.xlsx"];

range = ["A1:D100001"; "A1:D100001"; "A1:D100001"; "A1:D31259"];

for p = 1:4
    if p == 4
        clear
        p = 4;
        filename1 = 题目附件["..\\4 1-100000.xlsx"; 题目附件"..\\5 100001-200000.xlsx";题
            目附件
           "..\\6 200001-300000.xlsx"; 题目附件"..\\7 300001-331258.xlsx"];

        filename2 = 改["1-100000.xlsx"; 改"100001-200000.xlsx"; 改"200001-300000.xlsx";
            改"300001-331258.xlsx"];

        range = ["A1:D100001"; "A1:D100001"; "A1:D100001"; "A1:D31259"];
    end
    T1 = readcell(filename1(p), 'Range', range(p));
    [m, ~] = size(T1);
    mx = 150;  k = 1;
    x = zeros(mx, 2); x(1, 1) = cell2mat(T1(2, 1));

    %记录每个人的记录范围
```

```
for i = 3:m
    t = cell2mat(T1(i, 1));
    if x(k, 1) ~= t
        x(k, 2) = i - 1;
        k = k + 1;
        x(k, 1) = t;
    end
end
x(k, 2) = m;

xf = 食物{"", 0}; xf(2, :) = {string(T1{2, 4}), 1};
xfk = 3;     sf(1, :) = 食物""; sf(2, :) = string(T1{2, 4});

%填补缺失值
for i = 2:mx
    if x(i, 1) == 0
        break;
    end
    clearvars xf xfk sf
    xf = 食物{"", 0}; xf(2, :) = {string(T1{2, 4}), 1};
    xfk = 3;     sf(1, :) = 食物""; sf(2, :) = string(T1{2, 4});
    %记录每种食物的频数
    for j = 3:x(i, 2)
        fn = string(T1{j, 4});
        %判断是不是新的食物
        if sum(contains(sf, fn)) > 0
            %不是新的食物
            lo = find(contains(sf, fn) > 0);
            xf(lo, 2) = num2cell(cell2mat(xf(lo, 2)) + 1); %频数+1
        else
            %是新的食物
            xf(xfk, 1) = cellstr(fn); xf(xfk, 2) = num2cell(1);
            sf(xfk, 1) = fn;          xfk = xfk + 1;
        end
    end

    %用频数最高的食物填充缺失值
    a = cell2mat(xf(:, 2));
    lo = find(a == max(a));
    maxName = sf(lo);
    for j = 2:x(i, 2)
        if ismissing(string(T1{j, 4}))
            T1(j, 4) = cellstr(maxName);
        end
    end
end

%用字符串记录食物在文件中的索引
Ts(1, 1) = 食物"";
for i2 = 2:m
```

```
    Ts(i2, 1) = cell2mat(T1(i2, 4));
end

%用数字记录食物在文件中的价格
Tp(1, 1) = 0;
for i2 = 2:m
    Tp(i2, 1) = str2double(T1{i2, 3});
end

%处理待定为对应价格出现次数最多的食物, 记录每个待定的价格dn
dn = zeros(1000, 1, 1); dk = 2;
%记录所有的待定的索引lod
lod = find(Ts == 待定""); lodfl = zeros(length(lod), 1, 1);
%每一行记录同一价格待定的索引dx
dx = zeros(1000, 1000, 1);
%替换不同价位的待定
for id = 1:length(lod)
    %lodfl说明该待定已经被处理=1
    if lodfl(id, 1, 1) == 1
        continue;
    end
    %记录待定的价格
    td = str2double(T1{lod(id, 1), 3});

    if sum(sum(dn == td)) == 0 %新的待定
        dn(dk) = td;  dk = dk + 1;
        %存储价位的待定的索引ldtd
        ld = find(Ts == 待定"" & Tp == td);
        %存储价位的其他食物的索引lfdtd
        lfd = find(Ts ~= 待定"" & Tp == td);

        %循环找出价位的频数最高的食物td
        pona(1, 1) = 食物""; ponu(1, 1) = 0; pok = 2;
        for po = 1:length(lfd)
            ttp = T1{lfd(po, 1), 4};

            if sum(sum(pona == ttp)) == 0 %新的食物
                pona(pok, 1) = ttp; ponu(pok, 1) = 1; pok = pok + 1;
            else %旧的食物
                lp = find(pona == ttp);
                ponu(lp, 1) = ponu(lp, 1) + 1;
            end
        end
        %价位的频数最高的食物的索引td
        lopo = find(ponu == max(ponu));
        lona = pona(lopo);
        %将待定的值赋为频数最高的食物
        for lpp = 1:length(ld)
            T1(ld(lpp), 4) = cellstr(lona);
            lodfl(ld(lpp), 1) = 1;
```

```
        end
    end
end

clearvars dn dk lod lodfl dx id td ld lfd pona ponu pok ttp po lp lopo lona


%处理一层待定
%处理一层待定为对应价格出现次数最多的食物, 记录每个待定的价格dn
dn = zeros(1000, 1, 1); dk = 2;
%记录所有的一层待定的索引lod
lod = find(Ts == 一层待定""); lodfl = zeros(length(lod), 1, 1);
%每一行记录同一价格一层待定的索引dx
dx = zeros(1000, 1000, 1);
%替换不同价位的一层待定
for id = 1:length(lod)
    %lodfl说明该一层待定已经被处理=1
    if lodfl(id, 1) == 1
        continue;
    end
    %记录一层待定的价格
    td = str2double(T1{lod(id, 1), 3});

    if sum(sum(dn == td)) == 0 %新的一层待定
        dn(dk) = td;  dk = dk + 1;
        %存储价位的一层待定的索引ldtd
        ld = find(Ts == 一层待定"" & Tp == td);
        %存储价位的其他食物的索引lfdtd
        lfd = find(Ts ~= 一层待定"" & Tp == td);

        %循环找出价位的频数最高的食物td
        pona(1, 1) = 食物""; ponu(1, 1) = 0; pok = 2;
        for po = 1:length(lfd)
            ttp = T1{lfd(po, 1), 4};

            if sum(sum(pona == ttp)) == 0 %新的食物
                pona(pok, 1) = ttp; ponu(pok, 1) = 1; pok = pok + 1;
            else %旧的食物
                lp = find(pona == ttp);
                ponu(lp, 1) = ponu(lp, 1) + 1;
            end
        end
        %价位的频数最高的食物的索引td
        lopo = find(ponu == max(ponu));
        lona = pona(lopo);
        %将待定的值赋为频数最高的食物
        for lpp = 1:length(ld)
            T1(ld(lpp), 4) = cellstr(lona);
            lodfl(ld(lpp), 1) = 1;
        end
```

```
        end
    end


    %更新Ts
    %用字符串记录食物在文件中的索引
    Ts(1, 1) = 食物"";
    for i2 = 2:m
        Ts(i2, 1) = cell2mat(T1(i2, 4));
    end
    %统一同一个食物的价格
    px = zeros(m, 1, 1);
    for i3 = 2:m - 1
        if px(i3) == 1
            continue;
        end

        sp = Ts(i3, 1);
        px(i3) = 1; %标记食物已经被处理
        ls = find((Ts == sp) == 1);

        %计算同一食物的价格平均值
        avg = 0;
        for j = 1:length(ls)
            px(ls(j, 1)) = 1;
            avg = avg + str2double(T1{ls(j, 1), 3});
        end
        avg = round(avg / length(ls));
        %将价格统一替换为平均值
        for j = 1:length(ls)
            T1(ls(j, 1), 3) = cellstr(num2str(avg));
        end
    end


    writecell(T1, filename2(p));
end
```

附录E　问题三对第一，二，三年数据的特征的分别提取-featuerExtractionSingle3_123.m

```
%第三问提取特征
clc, clear;

filename = 数据特征提取["..\\1.xlsx"; 数据特征提取"..\\2.xlsx"; 数据特征提取"..\\3.xlsx"];
range = ["A1:K5022"; "A2:K4301"; "A2:K3686"];
```

```
filenf = 数据改["..\\1-100000.xlsx"; 数据改"..\\100001-200000.xlsx";数据改
    "..\\200001-300000.xlsx"; 数据改"..\\300001-331258.xlsx"];
rangf = ["A1:D100001"; "A2:D100001"; "A2:D100001"; "A2:D31259"];

p = readmatrix数据附件出现序号("..\\4-7.xlsx", 'Range', 'F2:F302');
fp = readcell数据不同食物及其价格("..\\.xlsx", 'Range', 'A1:B167');

%是低价食品，为平价食品，为高价食品1~5253~163164~167
%记录食物名字
for i = 1:167
    fps(i, 1) = string(fp{i, 1});
end

xf = readcell(filenf(1), 'Range', rangf(1));
%将所有食物记录汇总
[mf, nf] = size(xf);
for i = 2:4
    xfp = readcell(filenf(i), 'Range', rangf(i));
    [mfp, nfp] = size(xfp);
    xf(mf + 1:mf + mfp, :) = xfp;
    mf = mf + mfp;
end

%表明食物的种类
for i = 1:mf
    lo = find(xf{i, 4} == fps);
    if lo <= 52
        xf(i, 5) = num2cell(0);
    elseif lo <= 163
        xf(i, 5) = num2cell(1);
    else
        xf(i, 5) = num2cell(2);
    end
end

%按时间排序
tm = xf(2:end, :);
a2 = string(cell2mat(tm(:, 2)));
[~, ind] = sort(a2);
tm = tm(ind, :);
xf(2:end, :) = tm;

%第二年数据起始点
k1 = 173973;
%第三年数据起始点
k2 = 227883;
%保存序号
xfpi = zeros(mf, 1, 1); xfpi(1, 1) = -1;
xfpn(1, 1) = 时间"";
for i = 2:mf
```

```
    xfpi(i) = xf{i, 1};
    xfpn(i, 1) = xf{i, 2};
end
clearvars xfp mfp nfp ind a2;


for year = year:3
    xt = readmatrix(filename(year), 'Range', range(year));

    %存储最终的特征值
    tt(:, 1) = p;
    [mt, nt] = size(tt);
    %循环每一个人
    for i = 1:mt
        %序号
        temp = tt(i, 1);
        % 在已经提取的人里找
        lo = find(temp == xt(:, 1));
        if isempty(lo)
            tt(i, 2:11) = zeros(10, 1, 1);
        else
            tt(i, 2:11) = xt(lo, 2:11);
        end

        %查找序号对应的食物记录
        lop = find(temp == cell2mat(xf(2:end, 1)));
        lop = lop + 1;

        %记录每种食物出现的次数
        times = zeros(3, 1, 1);
        for j = 1:length(lop)
            t = lop(j);

            if year == 1
                if t >= k1
                    break;
                end
                times(xf{t, 5} + 1) = times(xf{t, 5} + 1) + 1;
            elseif year == 2
                if t < k1
                    continue;
                end
                if t >= k2
                    break;
                end
                times(xf{t, 5} + 1) = times(xf{t, 5} + 1) + 1;
            else
                if t < k2
                    continue;
                end
```

```
                times(xf{t, 5} + 1) = times(xf{t, 5} + 1) + 1;
            end
        end

        times = times / sum(times);
        tt(i, 12:14) = times;
    end

    writematrix(tt, ['数据特征及食物提取..\\', num2str(year), '.xlsx']);
end
```

附录F　问题二对第一，二，三年数据的特征的分别提取-featureExtractionSingle2_123.m

```
%特征提取
%性别
%周早餐消费总金额 周午餐消费总金额周晚餐消费总金额
%周早餐次数 周午餐次数周晚餐次数
%周总消费金额 周总吃饭次数日均消费金额

clc, clear;
Tx = readmatrix题目附件("..\\0 性别标签.xlsx", 'Range', "A2:B5416");
filename = 数据第一年三餐消费数据改["..\\.xlsx"; 数据第二年三餐消费数据改"..\\.xlsx";数据第三年三餐消费数据改
    "..\\.xlsx"];
range = ["B2:AGG5416"; "B2:ST5416"; "B2:AEW5416"];
%将三年的数据汇总
% for i = 2:3
%     t = readmatrix(filename(i), 'Range', range(i));
%     [~, nt] = size(t);
%     T(:, n + 1:n + nt) = t;
%     n = n + nt;
% end

for k = 1:3
    T = readmatrix(filename(k), 'Range', range(k));
    [m, n] = size(T);


    %第一行为 特征值的序号
    x = zeros(m + 1, 11, 1); x(2:end, 1) = 1:m; x(2:end, 2) = Tx(:, 2); x(1, 2:11) = 1:10;
    for i = 1:m
        s1 = 0;    s2 = 0;    s3 = 0;
        s4 = 0;    s5 = 0;    s6 = 0;
        s7 = 0;    s8 = 0;    s9 = 0;

        for j = 1:n
```

```
        %周早餐消费总金额 周午餐消费总金额周晚餐消费总金额
        %周早餐次数 周午餐次数周晚餐次数
        %周总消费金额 周总吃饭次数日均消费金额

        f = mod(j, 3);
        %i == 早餐1:
        if f == 1
            s1 = s1 + T(i, j);
            if T(i, j) > 0
                s4 = s4 + 1;
            end
        elseif f == 2
            %i == 午餐2:
            s2 = s2 + T(i, j);
            if T(i, j) > 0
                s5 = s5 + 1;
            end
        else
            %i == 晚餐0:
            s3 = s3 + T(i, j);
            if T(i, j) > 0
                s6 = s6 + 1;
            end
        end
    end
    s1 = s1 / (n / 21);   s2 = s2 / (n / 21);   s3 = s3 / (n / 21);
    s4 = s4 / (n / 21);   s5 = s5 / (n / 21);   s6 = s6 / (n / 21);
    s7 = s1 + s2 + s3;    s8 = s4 + s5 + s6;    s9 = s7 / 7;

    i = i + 1;
    x(i, 3) = s1;          x(i, 4) = s2;        x(i, 5) = s3;
    x(i, 6) = s4;          x(i, 7) = s5;        x(i, 8) = s6;
    x(i, 9) = s7;          x(i, 10) = s8;        x(i, 11) = s9;
    i = i - 1;
end


c = x(2:end, 2:end);
%部分人不在食堂吃饭
p1 = find(sum(c, 2) == 0); lenp1 = length(p1);      %女生0
p2 = find(sum(c, 2) == 1); lenp2 = length(p2);      %男生1
p = zeros(lenp1 + lenp2, 2, 1);
p(1:lenp1, 1:2) = [p1, zeros(lenp1, 1, 1)];
p(lenp1 + 1: lenp1 + lenp2, 1:2) = [p2, ones(lenp2, 1, 1)];

%去除不在食堂吃饭的人
xp(1, :) = x(1, :); ip = 2;
for i = 2:m + 1
    if sum(p(:, 1) == x(i, 1)) == 0
        xp(ip, :) = x(i, :); ip = ip + 1;
```

```
        end
    end
    %更新x
    clearvars x;
    x = xp;
    [m, n] = size(x);

    im = 2; iw = 2;
    xm(1, 1:10) = [0, 2:10]; xw(1, 1:10) = [0, 2:10];
    for i = 2:m
        if x(i, 2) == 1 % 女生0 男生1
            xm(im, 1:10) = [x(i, 1), x(i, 3:end)]; im = im + 1;
        else
            xw(iw, 1:10) = [x(i, 1), x(i, 3:end)]; iw = iw + 1;
        end
    end

    writematrix(x, string(['数据特征提取..\\', num2str(k), '.xlsx']));
%     writematrix(xm, string数据特征提取男(['..\\', num2str(k), '.xlsx']));
%     writematrix(xw, string数据特征提取女(['..\\', num2str(k), '.xlsx']));
    writematrix(p, string(['数据不在食堂吃饭的人..\\', num2str(k), '.xlsx']));
end
```

附录G　问题三对汇总的三年数的特征的提取-featureExtractionTotal1.m

```
 %特征提取
%性别
%周早餐消费总金额 周午餐消费总金额周晚餐消费总金额
%周早餐次数 周午餐次数周晚餐次数
%周总消费金额 周总吃饭次数日均消费金额

clc, clear;
filename = 数据第一年三餐消费数据改["..\\.xlsx";数据第二年三餐消费数据改
    "..\\.xlsx"; 数据第三年三餐消费数据改"..\\.xlsx"];
range = ["B2:AGG5416"; "B2:ST5416"; "B2:AEW5416"];
% sname = 性别["", 周早餐消费总金额"", 周早餐消费总金额"", 周早餐消费总金额""]
%将三年的数据汇总
T = readmatrix(filename(1), 'Range', range(1));
[m, n] = size(T);
for i = 2:3
    t = readmatrix(filename(i), 'Range', range(i));
    [~, nt] = size(t);
    T(:, n + 1:n + nt) = t;
    n = n + nt;
end
```

```
Tx = readmatrix题目附件("..\\0 性别标签.xlsx", 'Range', "A2:B5416");

% n = n - 1;

%第一行为 特征值的序号
x = zeros(m + 1, 11, 1); x(2:end, 1) = 1:m; x(2:end, 2) = Tx(:, 2); x(1, 2:11) = 1:10;
for i = 1:m
    s1 = 0;    s2 = 0;    s3 = 0;
    s4 = 0;    s5 = 0;    s6 = 0;
    s7 = 0;    s8 = 0;    s9 = 0;

    for j = 1:n
        %周早餐消费总金额 周午餐消费总金额周晚餐消费总金额
        %周早餐次数 周午餐次数周晚餐次数
        %周总消费金额 周总吃饭次数日均消费金额

        f = mod(j, 3);
        %i == 早餐1:
        if f == 1
            s1 = s1 + T(i, j);
            if T(i, j) > 0
                s4 = s4 + 1;
            end
        elseif f == 2
            %i == 午餐2:
            s2 = s2 + T(i, j);
            if T(i, j) > 0
                s5 = s5 + 1;
            end
        else
            %i == 晚餐0:
            s3 = s3 + T(i, j);
            if T(i, j) > 0
                s6 = s6 + 1;
            end
        end
    end
    s1 = s1 / (n / 21);   s2 = s2 / (n / 21);   s3 = s3 / (n / 21);
    s4 = s4 / (n / 21);   s5 = s5 / (n / 21);   s6 = s6 / (n / 21);
    s7 = s1 + s2 + s3;    s8 = s4 + s5 + s6;    s9 = s7 / 7;

    i = i + 1;
    x(i, 3) = s1;          x(i, 4) = s2;        x(i, 5) = s3;
    x(i, 6) = s4;          x(i, 7) = s5;        x(i, 8) = s6;
    x(i, 9) = s7;          x(i, 10) = s8;        x(i, 11) = s9;
    i = i - 1;
end
```

```
c = x(2:end, 2:end);
%部分人不在食堂吃饭
p1 = find(sum(c, 2) == 0); lenp1 = length(p1);      %女生0
p2 = find(sum(c, 2) == 1); lenp2 = length(p2);      %男生1
p = zeros(lenp1 + lenp2, 2, 1);
p(1:lenp1, 1:2) = [p1, zeros(lenp1, 1, 1)];
p(lenp1 + 1: lenp1 + lenp2, 1:2) = [p2, ones(lenp2, 1, 1)];

%作图
% figure;
% y = sum(c, 2);
% % col = rand(m, 3);
% col(:, 1) = linspace(1, 100, m)';
% col(:, 2) = linspace(1, 100, m)';
% col(:, 3) = linspace(1, 100, m)';
% % scatter(1:m, y, 25, col);
% scatter(1:m, y, []);
% ylabel人按序号排("()");
% xlabel特征值总和("");
% title每个人的特征值总和("");
% xlim([0 5415]);
% xticks(1:100:m);

%去除不在食堂吃饭的人
xp(1, :) = x(1, :); ip = 2;
for i = 2:m + 1
    if sum(p(:, 1) == x(i, 1)) == 0
        xp(ip, :) = x(i, :); ip = ip + 1;
    end
end
%更新x
clearvars x;
x = xp;
[m, n] = size(x);

im = 2; iw = 2;
xm(1, 1:10) = [0, 2:10]; xw(1, 1:10) = [0, 2:10];
for i = 2:m
    if x(i, 2) == 1 % 女生0 男生1
        xm(im, 1:10) = [x(i, 1), x(i, 3:end)]; im = im + 1;
    else
        xw(iw, 1:10) = [x(i, 1), x(i, 3:end)]; iw = iw + 1;
    end
end

writematrix(x, 数据特征提取总"..\\.xlsx");
writematrix(xm, 数据特征提取男总"..\\.xlsx");
writematrix(xw, 数据特征提取女总"..\\.xlsx");
writematrix(p, 数据不在食堂吃饭的人总"..\\.xlsx");
```

附录H　问题一统计食物的价格并对食物进行分类-foodPriceClass1.m

```
%统计所有的食物，并且划分为三个种类
clc, clear
filename = 数据改["..\\1-100000.xlsx"; 数据改"..\\100001-200000.xlsx";数据改
    "..\\200001-300000.xlsx"; 数据改"..\\300001-331258.xlsx"];

range = ["A2:D100001"; "A2:D100001"; "A2:D100001"; "A2:D31259"];

it = 1; Ts(1, 1) = 食物""; Tp = [];
for k = 1:4
    T = readcell(filename(k), 'Range', range(k));
    [m, n] = size(T);
    for i = 1:m
        if sum(Ts == T{i, 4}) == 0
            %记录新的食物和价格
            Ts(it, 1) = T{i, 4}; Tp(it, 1) = str2double(T{i, 3}); it = it + 1;
        end
    end
end
s = 致远一层超市[""; 深圳校区图书馆""; 游泳馆""];
Ts1(1, 1) = 食物""; Tp1 = []; ik = 1;
for i = 1:length(Ts)
    if sum(Ts(i, 1) == s) == 0
        Ts1(ik, 1) = Ts(i, 1); Tp1(ik, 1) = Tp(i, 1); ik = ik + 1;
    end

end
TT = 食物{"", 0};
TT(1:length(Ts1), 1) = cellstr(Ts1(:, 1));
TT(1:length(Tp1), 2) = num2cell(Tp1(:, 1));
[~,ind] = sort(Tp1);
TT = TT(ind,:);

figure;
plot(1:length(Tp1), cell2mat(TT(:, 2)));
xlim([1 170]);
xlabel不同食物("");
ylabel食物的价格单位分("(:)");
title不同食物价格变化图("");

writecell(TT, 数据不同食物及其价格"..\\.xlsx");
```

附录I　问题一绘制对同一食物价格处理前后对比图-foodPriceCompare1.m

```
%绘制同一食物处理前后的价格对比图
clc, clear;
%未处理的
filename = 题目附件"..\\4 1-100000.xlsx";
T = readcell(filename, 'Range', 'A1:D100001');
[m, ~] = size(T);
sn = 大灶稀饭"";

Ts(1, 1) = 食物"";
for i = 2:m
    Ts(i, 1) = T{i, 4};
end

lo = find(Ts == sn);
pr = zeros(length(lo), 1, 1);
for i = 1:length(lo)
    pr(i, 1) = str2double(T{lo(i, 1), 3});
end

%处理的
filename = 改"1-100000.xlsx";
T = readcell(filename, 'Range', 'A1:D100001');
[m, ~] = size(T);

Ts(1, 1) = 食物"";
for i = 2:m
    Ts(i, 1) = T{i, 4};
end

lo = find(Ts == sn);
px = str2double(T{lo(1, 1), 3});
pt = zeros(length(pr), 1, 1);
for i = 1:length(pr)
    pt(i, 1) = px;
end

figure
x = 1:length(pr);
plot(x, pr, x, pt, 'r');
ylabel食物价格单位：分("()");
xlabel食物出现次数("");
title大灶稀饭的价格处理前后对比前蓝后红("()");
xlim([0 750])
```

附录J　问题一绘制六个代表学生的三年食物种类变化图-foodPricePeoplePicture1.m

```
%六个代表，食物种类到，画个图，1~36
clc, clear
filename = 数据改["..\\1-100000.xlsx"; 数据改"..\\100001-200000.xlsx";数据改
    "..\\200001-300000.xlsx"; 数据改"..\\300001-331258.xlsx"];

range = ["A1:D100001"; "A1:D100001"; "A1:D100001"; "A1:D31259"];
T = readcell(filename(1), 'Range', range(1));
Ts = readcell数据不同食物及其价格("..\\.xlsx", "Range", 'A1:B167');
[m, n] = size(Ts);
for i = 1:m
    sn(i, 1) = string(Ts{i, 1});
    sp(i, 1) = Ts(i, 2);
end

%第一列为序号，第二列为等级
dm = [3062, 1; 76, 2; 3207, 3];
dw = [1275, 1; 5001, 2; 3272, 3];

%是低价食品，为平价食品，为高价食品1~5253~163164~167
for i = 1:6
    if i == 5
        continue;
    end
    if i >= 4
        t = dw(i - 3, 1);
    else
        t = dm(i, 1);
    end
    lo = find(t == cell2mat(T(2:end, 1)));
    fn(1, 1) = 食物""; fk = 1;
    for j = 1:length(lo)
        fn(fk, 1) = string(T{lo(j), 4}); fk = fk + 1;
    end
    %记录对应食品的等级
    for j = 1:fk - 1
        lf = find(sn == fn(j));
        if lf < 53
            fp(j, 1) = 1;
        elseif lf < 164
            fp(j, 1) = 2;
        else
            fp(j, 1) = 3;
        end
    end
```

```
    %绘图
    figure;
    plot(1:length(fp), fp, 'o');
    xlabel食物等级("");
    ylabel刷卡记录单位：次("()")
    title饮食等级图("");
end

%单独处理女生5001
t = dw(2, 1);
T = readcell(filename(2), 'Range', range(2));
lo = find(t == cell2mat(T(2:end, 1)));
fn(1, 1) = 食物""; fk = 1;
for j = 1:length(lo)
    fn(fk, 1) = string(T{lo(j), 4}); fk = fk + 1;
end
%记录对应食品的等级
for j = 1:fk - 1
    lf = find(sn == fn(j));
    if lf < 53
        fp(j, 1) = 1;
    elseif lf < 164
        fp(j, 1) = 2;
    else
        fp(j, 1) = 3;
    end
end

%绘图
figure;
plot(1:length(fp), fp, 'o');
xlabel食物等级("");
ylabel刷卡记录单位：次("()")
title饮食等级图("");
```

附录K　问题三汇总kmeans聚类和rbf神经网络的结果-kmeanRbfResult3.m

```
 %将和rbfk-的结果综合mean
clc, clear;
%k-结果变换means
km = [1, 0, 2; 0, 1, 2;  0, 1, 2];

filekm = 数据贫困标签预测值和相对误差["..\\kmeans3_1.xlsx";数据贫困标签预测值和相对误差
    "..\\kmeans3_2.xlsx";数据贫困标签预测值和相对误差
    "..\\kmeans3_3.xlsx"];
rangekm = 'A2:B302';
```

```
filerbf = 数据贫困标签预测值和相对误差["..\\rbf3_1.xlsx";数据贫困标签预测值和相对误差
    "..\\rbf3_2.xlsx";数据贫困标签预测值和相对误差
    "..\\rbf3_3.xlsx"];
rangerbf = ["A1:B53"; "A1:B301"; "A2:B301"];
fT = readmatrix(filerbf(1), 'Range', rangerbf(1));
kT = readmatrix(filekm(1), 'Range', rangekm);
%循环处理结果kmean
for i = 1:length(kT)
    kT(i, 2) = km(1, kT(i, 2));
end
fT(:, 3) = zeros(length(fT), 1, 1) - 1;
for i = 1:length(fT)
    lo = find(fT(i, 1) == kT(:, 1));
    fT(i, 3) = round((fT(i, 2) + kT(lo, 2)) / 2);
end
writematrix([fT(:, 1), fT(:, 3)], 数据"..\\方法确定的贫困标签kmeanRbf1.xlsx");

%第二年
for year = 2:3
    fT = readmatrix(filerbf(year), 'Range', rangerbf(year));
    kT = readmatrix(filekm(year), 'Range', rangekm);
    %循环处理结果kmean
    for i = 1:length(kT)
        if isnan(kT(i, 2))
            %不在食堂里吃饭的人
            kT(i, 2) = 0;
        else
            kT(i, 2) = km(year, kT(i, 2));
        end
    end

    %汇总的结果rbf
    fT(:, 3) = zeros(length(fT), 1, 1) - 1;
    for i = 1:length(fT)
        if isnan(fT(i, 2))
            %不在食堂里吃饭的人
            kT(i, 2) = 0;
        end
        lo = find(fT(i, 1) == kT(:, 1));
        fT(i, 3) = round((fT(i, 2) + kT(lo, 2)) / 2);
    end
    writematrix([fT(:, 1), fT(:, 3)], string(['数据..\\方法确定的贫困标
        签kmeanRbf', num2str(year), '.xlsx']));
end
```

附录L　问题二rbf神经网络预测学生贫困标签-RBF2.m

```
clc, clear
```

```
xt = readmatrix数据特征提取("..\\1.xlsx", 'Range', 'A2:K5022');
t = readmatrix题目附件("..\\8 已知贫困标签.xlsx", 'Range', 'A2:B4416');
% pt = zeros(1000, 2, 1);
pt(:, 1) = readmatrix题目附件("..\\9 问题待补全标签数据2.xlsx", 'Range', 'A2:A1001');
px = readmatrix数据不在食堂吃饭的人("..\\1.xlsx", 'Range', 'A1:A394');

[m, n] = size(xt);
xt(:, n + 1) = -1; n = n + 1;

cnt = 0;%空数据
for i = 1:m
    lo = find(t(:, 1) == xt(i, 1));
    if isempty(lo)
        cnt = cnt + 1;
        continue;
    end
    xt(i, n) = t(lo, 2);
end

%空数据
xk = 1;
for i = 1:m
    if xt(i, n) == -1
        continue;
    end
    tt(xk, :) = xt(i, :); xk = xk + 1;
end
xkk = xk;
for i = 1:m
    if xt(i, n) == -1
        tt(xkk, :) = xt(i, :); xkk = xkk + 1;
    end
end
xkk = xkk - 1; xk = xk - 1;

a = tt(:, 2:end)';
%处理
a(11, :) = a(11, :) + 1;
%规格化处理
pk = 4000;
P = a(1:10, 1:pk); [PN, PS1] = mapminmax(P); %自变量数据规格化到[-1,1]
T = a(11, 1:pk); [TN, PS2] = mapminmax(T); %因变量数据规格化到[-1,1]

% load("net.mat");
net = newrb(PN, TN); %训练网络RBF
% save("net.mat", "net");

%预测样本点自变量规格化
x = a(1:10, pk + 1:xk); xn = mapminmax('apply', x, PS1);
%求预测值，并把数据还原
```

```
yn1 = sim(net, xn);  y1 = mapminmax('reverse', yn1, PS2);
%计算网络预测的相对误差RBF
dt = abs(a(11, pk + 1:xk) - y1) ./ a(11, pk + 1:xk);
delta = mean(abs(a(11, pk + 1:xk) - y1) ./ a(11, pk + 1:xk));

y1 = round(abs(y1 - 1)); a(11, :) = a(11, :) - 1;
writematrix([y1', dt'], 数据贫困标签预测值和相对误差"..\\rbf2_1.xlsx");

%补全附件9
%预测样本点自变量规格化
pp = tt(xk + 1:xkk, 2:11)'; ppn = mapminmax('apply', pp, PS1);
%求预测值，并把数据还原
pn = sim(net, ppn);  yt = mapminmax('reverse', pn, PS2);
%yt = round(yt);
yt = abs(round(yt));
tt(xk + 1:xkk, 12) = yt';

for i = 1:length(pt)
    lop1 = find(pt(i) == tt(:, 1));
    lop2 = find(pt(i) == px(:, 1));
    if isempty(lop1)
        if ~isempty(lop2)
            pt(i, 2) = 0;
            %           disp(i);
        end
    else
        pt(i, 2) = tt(lop1, 12);
    end
end
writematrix(pt, 数据问题标签数据补全"..\\2.xlsx");

%预测第年的贫困程度2,3
filename = 数据特征提取["..\\2.xlsx"; 数据特征提
    取"..\\3.xlsx"]; range = ["A2:K4301"; "A2:K3686"];
filenameP = 数据不在食堂吃饭的人["..\\2.xlsx"; 数据不在食堂吃饭的人"..\\3.xlsx"];
fp = 数据第二年个人贫困程度["..\\.xlsx"; 数据第三年个人贫困程度"..\\.xlsx"];
rangeP = ["A1:A1115"; "A1:A1730"];

for k = 1:2
    xtt = readmatrix(filename(k), 'Range', range(k));
    xp = readmatrix(filenameP(k), 'Range', rangeP(k));
    a = xtt(:, 2:end)';

    P = a(1:10, :); [PN, PS1] = mapminmax(P); %自变量数据规格化到[-1,1]
    %预测样本点自变量规格化
    pp = a; ppn = mapminmax('apply', pp, PS1);
    %求预测值，并把数据还原
    pn = sim(net, ppn); yt = mapminmax('reverse', pn, PS2);
    yt = abs(round(yt));
```

```
    xtt(:, 12) = yt';
    mtt = length(xtt); xpk = 1;
    for i = mtt + 1:5415
        xtt(i, :) = [xp(xpk, 1), zeros(1, 10, 1) - 1, 0]; xpk = xpk + 1;
    end
    writematrix([xtt(:, 1), xtt(:, end)], fp(k));
 end
```

附录M　问题三rbf神经网络预测学生贫困标签-RBF3.m

```
%第三问提取特征, 神经网络RBF
clc, clear;

filename = 数据特征及食物提取["..\\1.xlsx"; 数据特征及食物提取"..\\2.xlsx"; 数据特征及食物提
    取"..\\3.xlsx"];
range = "A1:N301";

year = 1;
tt = readmatrix(filename(year), 'Range', range);
Ts = readmatrix题目附件("..\\8 已知贫困标签.xlsx", 'Range', 'A2:B4416');
for i = 1:length(tt)
    lo = find(tt(i, 1) == Ts(:, 1));
    if isempty(lo)
        tt(i, 15) = -1;
    else
        tt(i, 15) = Ts(lo, 2);
    end
    %不在食堂吃的人认为不贫困
    if sum(tt(i, 3:11)) == 0
        tt(i, 15) = 0;
    end
end
txk = 1;
for i = 1:length(tt)
    if tt(i, 15) ~= -1
        tx(txk, :) = tt(i, :); txk = txk + 1;
    end
end
for i = 1:length(tt)
    if tt(i, 15) == -1
        tx(txk, :) = tt(i, :); txk = txk + 1;
    end
end

%个人有贫困标签253
k1 = 253;
a = tx(:, 2:end)';
%处理
```

```
a(14, :) = a(14, :) + 1;
%训练集个，个作为检验20053
pk = 200;
%规格化处理
P = a(1:13, 1:pk); [PN, PS1] = mapminmax(P); %自变量数据规格化到[-1,1]
T = a(14, 1:pk); [TN, PS2] = mapminmax(T); %因变量数据规格化到[-1,1]

% load("net.mat");  load("PS1.mat");  load("PS2.mat");

netT = newrb(PN, TN); %训练网络RBF
% save("netT.mat", "netT");

%预测样本点自变量规格化
pp = a(1:13, pk + 1:k1); ppn = mapminmax('apply', pp, PS1);
%求预测值，并把数据还原
pn = sim(netT, ppn); y1 = mapminmax('reverse', pn, PS2);

%计算相对误差
dt = abs(a(14, pk + 1:k1) - y1) ./ a(14, pk + 1:k1);
delta = mean(abs(a(14, pk + 1:k1) - y1) ./ a(14, pk + 1:k1));

y1 = abs(round(y1) - 1);
a(14, :) = a(14, :) - 1;
tx(pk + 1:k1, 15) = y1;
writematrix([tx(pk + 1:k1, 1), y1', dt'], 数据贫困标签预测值和相对误差"..\\rbf3_1.xlsx");



%预测其他人的贫困标签
for year = 2:3
    tt = readmatrix(filename(year), 'Range', range);
    a = tt(:, 2:end)';

    %处理
    [~, n] = size(a);
    %规格化处理
    P = a(1:13, 1:n); [PN, PS1] = mapminmax(P); %自变量数据规格化到[-1,1]
    %     T = a(14, 1:pk); [TN, PS2] = mapminmax(T); 因变量数据规格化到%[-1,1]
    %预测样本点自变量规格化
    pp = a(1:13, :); ppn = mapminmax('apply', pp, PS1);
    %求预测值，并把数据还原
    pn = sim(netT, ppn); y1 = mapminmax('reverse', pn, PS2);

    y1 = abs(round(y1) - 1);
    tt(:, 15) = y1;
    writematrix([tt(:, 1), y1'], ['数据贫困标签预测值和相对误
        差..\\rbf3_',num2str(year), '.xlsx']);
end
```

附录N　问题四线性规划方法求解资助公平分配-subsidizeAllocate4.m

```
%基于线性规划的资助分配
clc, clear
Tc = readmatrix数据("..\\方法确定的贫困标签kmeanRbf3.xlsx", 'Range', 'A1:B300');
Tx = readmatrix数据特征及食物提取("..\\3.xlsx", 'Range', 'A1:K3686'); % 第九列是周消费总额
w = readmatrix数据第三年三餐消费数据改("..\\.xlsx", 'Range', 'A2:AEW2');
week = (length(w) - 1) / 21;

%选择个人80
clearvars x;
xnk = 1;
for i = 1:300
    if Tc(i, 2) == 2
        lo = find(Tc(i, 1) == Tx(:, 1));
        if Tx(lo, 9) >= 1500 %剔除数据
            x(xnk, :) = [Tc(i, :), Tx(lo, 9)];
            xnk = xnk + 1;
        end
    end
end
cnt = xnk - 1;
for i = 1:300
    if Tc(i, 2) == 1
        lo = find(Tc(i, 1) == Tx(:, 1));
        if Tx(lo, 9) >= 1500 %剔除数据
            x(xnk, :) = [Tc(i, :), Tx(lo, 9)];
            xnk = xnk + 1;
        end
    end
end
xnk = xnk - 1;

%线性规划
c = x(1:80, 3) * week / 100; s = sum(c);
M = 1e+6; tcAvg = (s + M) / 80;
%创建最小化优化问题
prob = optimproblem;
%创建优化变量，目标函数
t = optimvar('t', 80, 'LowerBound', 0, 'UpperBound', M / 10);
goal = ((t + c) - tcAvg) .* ((t + c) - tcAvg);
prob.Objective = mean(goal);
%创建线性约束
prob.Constraints.con1 = sum(t) == M;
%创建非线性约束
prob.Constraints.con2 = t + c - s / 80 >= 0;
%求解优化问题
```

```
[sol, fval, flag, out] = solve(prob);
res = round(sol.t);

ma = [x(1:80, 1), c, res];
writematrix(ma, 数据差异化资助情况"..\\.xlsx");
```