大数据背景下基于机器学习和深度学习的家庭困难学生认定方法

张家绪　东北林业大学理学院 信息与计算科学 2019 级
马佳欣　东北林业大学理学院 信息与计算科学 2020 级

2021 年 3 月

目录

1 背景介绍

近年来，随着我国高校招生规模的不断扩大，高校贫困生也伴随着一定程度的增长，而贫困生认定工作作为国家帮扶贫困家庭政策的一部分，对于保障贫困家庭享受高等教育福利，公平享受社会主义发展给全国人民带来的福祉，具有重要作用。然而，我国的贫困生认定工作仍然存在一定程度的不足，主要表现为：

1. 生源不同，学生之间的贫困情况不同，难以制作统一的判别标准；
2. 学生的实际生活水平往往与学生提供的生活水平相差较大，判别精度不足；
3. 目前仍缺乏完善的、可以量化的判别指标，传统的判定模型往往以牺牲精度为代价，以换取更高的可解释性。

本文提出的模型将针对以上问题进行建模与优化。

进入大数据时代，面对海量的数据，如何做好数据挖掘、信息处理工作，进而使大数据为人所用，正在逐渐变成一个重要的研究课题。人工智能为信息挖掘带来了巨大的便利，机器学习、深度学习技术克服了传统方法难以针对大数据进行建模的弊端，使得从海量数据中识别家庭贫困学生成为可能 [1]。近年来，不断有高校采用基于校园卡消费、移动支付 APP 月账单、手机话费消费的方法进行数据挖掘，以期识别家庭贫困学生与消费之间的关联模式，获得的极好的效果。本文主要基于机器学习、深度学习方法，采用多个弱学习模型相融合的思路，建立了一个有监督学习模型，获得了大数据背景下基于机器学习和深度学习的家庭困难学生识别方法。

2 模型简介

在完成信息采集、数据清洗之后，作者主要基于模型融合理论，结合多元统计分析方法进行特征工程和数据挖掘，并训练了支持向量机、随机森林、XGBoost、神经网络四个弱分类求解器进行有监督学习，最后合成一个融合判别器进行优化建模。本文采用的融合判别器在测试集上获得了 88.71% 的准确率。

3 基于机器学习和深度学习的家庭困难学生判定模型

3.1 获取数据、数据清洗

原始数据基于问卷星平台由问卷调查形式获得，问卷的问题参考了文献 [3]。不妨记样本空间为 Ω，有效样本空间为 $\hat{\Omega} \subseteq \Omega$，截至建模时共收到样本 $|\Omega| = 472$ 件。该问卷

以 xlsx 格式的电子表格保存，占用空间 67.9KB，共包含 21 个字段。考虑原始字段集 $\mathcal{A}$，字段个数 $|\mathcal{A}|$，我们有：

$$\mathcal{A} := \{a_i | a_i \text{为原始字段}, 1 \leqslant i \leqslant |\mathcal{A}|, i \in \mathbb{N}\},$$

此处字段个数 $|\mathcal{A}| = 21$，字段集 $\mathcal{A}$ =｛“序号”,“提交答卷时间”,“所用时间”,“来源”,“来源详情”,“来自 IP”,“1、您的性别”,“2、年级”,“3、每月生活费开销大部分消费在”,“4、家庭所在地”,“5、家庭受非义务教育情况”,“6、父母工作类型”, “7、家中是否有重大疾病或者需常年服药维持”,“8、家庭是否处于贫困地区”,“9、家庭劳动力人口”,“10、突发事件”,“11、家庭负债情况 (以万为单位)”,“12、家庭人均月收入”,“13、是否在校申请贫困”,“14、每个月生活费 (不包括自己兼职得到)”,“15、每个月基础消费 (吃饭)”｝。

观察可得字段 $\{a_{19}\}$（“13、是否在校申请贫困”）即为有监督学习的标签字段。

3.1.1 数据清洗

首先针对原始数据进行数据清洗。作者基于 3-σ 原则删去了各个字段中包含离群值的样本，例如在原字段子集 $\{a_{20}, a_{21}\}$ =｛“14、每个月生活费 (不包括自己兼职得到)”,“15、每个月基础消费 (吃饭)”｝ 两个字段下填写了 {‘0’,‘1000 多’,‘父母给’,‘不知’ ,‘500 万’,‘没算过’,‘不造’,‘1’,‘20000000’,‘100000’, ‘20000000000000000’,‘看心情’} 的样本，显然这样的样本是数据集本身包含的随机噪声，但此类样本作为建模数据并不会提高模型的泛化能力，因此作者视这样的离群值为异常值并对其进行了删除，并对不易直接处理的数据格式（如字符型）进行了格式转换（如数值型），最终获得有效样本 $|\hat{\Omega}| = 452$ 个，有效样本比 $\eta = 95.76\%$。原数据集中未发现缺失值，故未进行缺失值处理。

3.1.2 匿名化处理

为了确保敏感信息被删除或受到保护，防止性别、生源地等不必要或不适用于建模的隐私信息进入建模过程，造成标签泄露，作者删除了原字段子集 $\{a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_{10}\}$ = {“提交答卷时间”,“所用时间”,“来源”,“来源详情”,“来自 IP”,“1、您的性别”,“2、年级”,“4、家庭所在地”}8 个字段，以进行匿名化处理。

3.2 特征工程

作者针对剩余的 13 个字段进行特征工程，以获得能指导建模的特征。特征工程的结果是产生了新的用于建模的特征空间，特征空间与 $\mathcal{A}$ 一样，同为概率空间，记为 $\mathcal{U}$，记特征个数为 $|\mathcal{U}|$。我们有

$$\mathcal{U} := \{u_j | u_j \text{为特征}, 1 \leqslant j \leqslant |\mathcal{U}|, j \in \mathbb{N}\},$$

本文中特征数 $|\mathcal{U}| = 21$，特征 $\mathcal{U}$ =“家庭受非义务教育人数”, “是否重大疾病”,“贫困地区”,“劳动力人口”,“突发事件”,“家庭负债”,“人均月收入”,“是否在校申请贫困”,“生活费”,“基础消费”,“网购”,“温饱问题”,“娱乐”,“谈恋爱”,“护肤或者化妆品”,“务农”, “失地农民”,“失业”,“个体工商户”,“城镇农民工”,“其它”}。

特征工程主要采用了建立稀疏特征子空间、离散化连续变量 2 种数据处理方法。

3.2.1 基于 one-hot 编码建立稀疏特征子空间

原始字段子集 $\{a_9, a_{12}\}$ = {“3、每月生活费开销大部分消费在”,“6、父母工作类型”}属于多选字段，受访对象可根据自身实际情况自由选择 1 个及以上的选项。考虑 a_9 的选项空间

$$\mathcal{C}_9 = \{c_9^{(1)}, c_9^{(2)}, c_9^{(3)}, c_9^{(4)}, c_9^{(5)}\},$$

此处 $\mathcal{C}_9$ = {“网购”,“温饱问题”,“娱乐”,“谈恋爱”,“护肤或者化妆品”}，作者将选项转换为 one-hot 编码，将样本子特征嵌入到一个新的向量空间，建立了基于月生活费开销选项的稀疏特征子空间

$$\mathcal{U}_k := \{u_{j_k} | u_{j_k} \text{为特征}, 11 \leqslant j_k \leqslant 15, j_k \in \mathbb{N}\},$$

$\mathcal{U}_k$ =“网购”,“温饱问题”,“娱乐”, “谈恋爱”,“护肤或者化妆品”}。这是一个 452×5 维的 0-1 二值化稀疏特征子空间，样本选择了某个选项记为 1，否则记为 0，如表1所示。

表 1: 基于月生活费开销的 0-1 二值化稀疏特征子空间

样本序号	网购	温饱问题	娱乐	谈恋爱	护肤或者化妆品
11	1	1	0	0	1
21	0	1	0	0	1
31	1	1	1	1	0

使用同样的方法处理原字段 a_{12}，即父母职业类型对应的选项空间

$$\mathcal{C}_{12} = \{c_{12}^{(1)}, c_{12}^{(2)}, c_{12}^{(3)}, c_{12}^{(4)}, c_{12}^{(5)}, c_{12}^{(6)}\},$$

$\mathcal{C}_{12}$ = {“务农”,“失地农民”,“失业”,“个体工商户”,“城镇农民工”,“其它”}，作者仍然采用 one-hot 编码将字段转换为特征，建立基于 one-hot 编码的稀疏特征子空间 $\mathcal{U}_k$,

$$\mathcal{U}_l := \{u_{j_l} | u_{j_l} \text{为特征}, 16 \leqslant j_l \leqslant 21, j_l \in \mathbb{N}\},$$

此处 $\mathcal{U}_k$ =“务农”,“失地农民”,“失业”,“个体工商户”,“城镇农民工”,“其它”}，如表2所示。

对于其他以二选一形式出现的字段，如原始字段子集 $\{a_{14}, a_{15}, a_{17}\}$={“7、家中是否有重大疾病或者需常年服药维持”,“8、家庭是否处于贫困地区”,“10、突发事件”}，它们

表 2: 基于父母职业类型的 one-hot 稀疏特征子空间

样本序号	务农	失地农民	失业	个体工商户	城镇农民工	其它
11	1	0	0	0	1	0
21	0	0	0	0	1	0
31	0	0	1	0	0	1

多是以 “是” 或 “否” 的选项出现的。针对这样的二选一字段，同样建立转换映射 T_{binary}，如公式(3.1)所示：

$$T_{binary}(c_n^{(j)}) = \begin{cases} 1, \text{if } c_n^{(j)} = \text{“是”} \\ 0, \text{if } c_n^{(j)} = \text{“否”} \end{cases}, \quad \forall n \in \{14, 15, 17\}, j \in \{1, 2\}; \tag{3.1}$$

采用同样的方法建立基于 one-hot 编码的稀疏特征子空间

$$\mathcal{U}_m := \{u_2, u_3, u_5\},$$

其中 $\mathcal{U}_m$ =“是否重大疾病”,“贫困地区”,“劳动力人口”,“突发事件”}。

3.2.2 连续特征离散化

考虑原字段集中的连续型字段子集 $\{a_{11}, a_{17}, a_{18}, a_{20}, a_{21}\}$ = {“5、家庭受非义务教育情况”,“11、家庭负债情况 (以万为单位)”,“12、家庭人均月收入”,“14、每个月生活费 (不包括自己兼职得到)”,“15、每个月基础消费 (吃饭)” }，观察可知其中的 $\{a_{20}, a_{21}\}$ 为可直接用于构建模型的连续型特征，故直接将其转换为特征 $\{u_9, u_{10}\}$，其中 u_9 为 “生活费”，u_{10} 为 “基础消费”。对于剩下的三个字段 $\{a_{11}, a_{17}, a_{18}\}$，观察可得三类字段均对应分箱型选项（即选项为数值区间），故需进一步进行连续特征离散化处理，以便于模型处理连续型变量字段。

针对字段 a_{11}（家庭受非义务阶段教育人事字段），选项空间 $\mathcal{C}_{11} = \{c_{11}^{(1)}, c_{11}^{(2)}, c_{11}^{(3)}\}$ 分别对应单选项 {‘1 人’,‘2 人’,‘3+’}。该字段的选项属于上界截断型（即在最大值 3 处截断），因此不妨构造离散化映射 $T_{person} : a_{11} \to u_1$ 如公式(3.2)所示：

$$T_{person}(c_{11}^{(i)}) = \begin{cases} 1, \text{if } c_{11}^{(i)} = \text{‘1 人’} \\ 2, \text{if } c_{11}^{(i)} = \text{‘2 人’} \\ 3, \text{if } c_{11}^{(i)} = \text{‘3+’} \end{cases}, \quad \forall i \in \{1, 2, 3\}; \tag{3.2}$$

根据映射后的像集构建随机变量 u_1 =“家庭受非义务教育人数” 作为一个离散特征。

针对字段 a_{17}（“家庭负债数字段”），选项空间 $\mathcal{C}_{17} = \{c_{17}^{(1)}, c_{17}^{(2)}, c_{17}^{(3)}, c_{17}^{(4)}\}$ = { ‘0’, ‘0-5’, ‘5-10’,‘10+’}。针对此分箱型字段（即区间型字段）作者取区间中值作为选项替代，

构造离散映射 $T_{budget}: a_{17} \to u_6$ 如式(3.3)：

$$T_{person}(c_{17}^{(j)}) = \begin{cases} 0, \text{if } c_{17}^{(j)} = \text{‘0’} \\ 2.5, \text{if } c_{17}^{(j)} = \text{‘0-5’} \\ 7.5, \text{if } c_{17}^{(j)} = \text{‘5-10’} \\ 15, \text{if } c_{17}^{(j)} = \text{‘10+’} \end{cases}, \quad \forall j \in \{1, \cdots, 4\}; \tag{3.3}$$

根据映射后的像集构建随机变量 u_6 =“家庭负债” 作为一个离散化特征。

针对字段 a_{18}（家庭人均月收入字段），选项空间 $\mathcal{C}_{18} = \{c_{18}^{(1)}, \cdots c_{18}^{(8)}\} = \{$‘600 以下’,‘600-1200’,‘1200-2000’,‘2000-3000’,‘3000-4000’,‘4000-5000’, ‘5000-6000’,‘6000+’$\}$ 对应着 8 个单选项。由于该字段的选项也属于分箱型，因此不妨取区间中值作为替代构造离散特征；对于截断上界选项 ‘6000+’，不妨取用 8500 作为替代。不妨构造离散化映射 $T_{income}: a_{18} \to u_7$ 如公式(3.4)所示：

$$T_{person}(c_{18}^{(k)}) = \begin{cases} 300, \text{if } c_{18}^{(k)} = \text{‘600 以下’} \\ 900, \text{if } c_{18}^{(k)} = \text{‘600-1200’} \\ 1600, \text{if } c_{18}^{(k)} = \text{‘1200-2000’} \\ \quad \cdots \\ 5500, \text{if } c_{18}^{(k)} = \text{‘5000-6000’} \\ 8500, \text{if } c_{18}^{(k)} = \text{‘6000+’} \end{cases}, \quad \forall k \in \{1, 2, \cdots, 8\}; \tag{3.4}$$

同样地，根据映射后的像集构建随机变量 u_7 =$\{$“人均月收入”$\}$ 作为一个离散化特征。

3.3 特征分析与探索 (EDA)

针对特征空间 $\mathcal{U}$ 进行多元统计分析 [2]，获得数据间的关联模式，更好地指导建模。

3.3.1 相关性分析

Spearman 秩相关系数是一种用于分析变量之间相关性的统计量，只要两个变量具有严格单调的函数关系，那么它们就是完全 Spearman 相关的，而 Pearson 相关系数仅在变量存在较强线性相关性时才更具有参考价值。Spearman 秩相关系数计算公式如下式(3.5)所示 [4]：

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(\mathcal{R}_i - \mathcal{Q}_i)^2}{n(n^2-1)}, \tag{3.5}$$

其中 $\mathcal{R}_i$ 代表 $u_{\mathcal{R}}^{(i)}$ 的秩次，$\mathcal{Q}_i$ 代表 $u_{\mathcal{Q}}^{(i)}$ 的秩次，$\mathcal{R}_i - \mathcal{Q}_i$ 代表 $u_{\mathcal{R}}^{(i)}$ 与 $u_{\mathcal{Q}}^{(i)}$ 的秩次之差。Spearman 秩相关系数 r_s 越接近 1，比较变量 $u_{\mathcal{R}}$ 与 $u_{\mathcal{Q}}$ 的相关性越强。计算特征与标签的 Spearman 相关系数矩阵如图1所示。

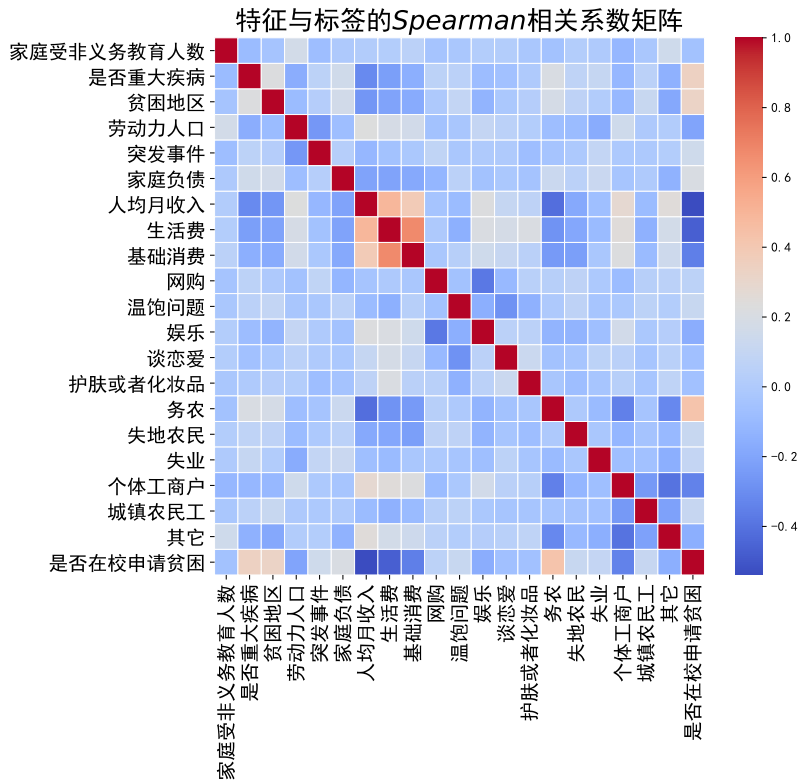

图 1: 特征与标签的 Spearman 相关系数热力图

简单分析图1可知“人均月收入”、“生活费”、“基础消费”三个变量之间存在一定的正相关性，且三者与“是否在校申请贫困”存在一定的负相关性。“是否重大疾病”、“贫困地区”、“务农”与“是否在校申请贫困”存在一定的正相关性。但是根据热力图也可以发现变量冗余严重，很多变量与“是否在校申请贫困”并不存在很强的相关性，因此取部分变量绘制部分特征与标签的 Spearman 相关系数矩阵如图2所示。

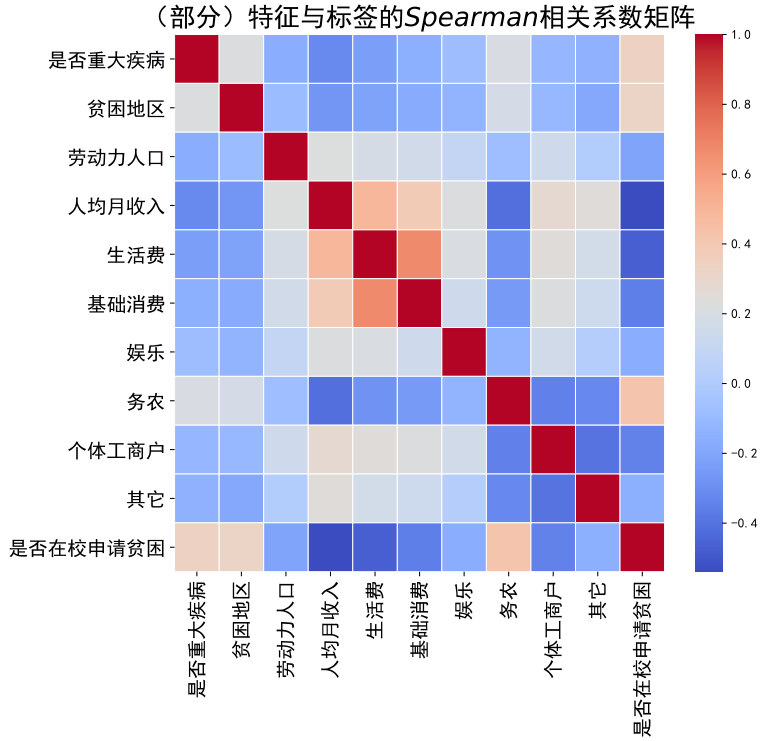

图 2: 部分特征与标签的 Spearman 相关系数热力图

以秩相关系数的绝对值为阈值取定 10 个特征后，由图2可知：“是否重大疾病”、“贫困地区”、“务农” 与 “是否在校申请贫困” 存在一定的正相关性，“劳动力人口”、“人均月收入”、“生活费”、“基础消费”、“娱乐”、“个体工商户”、“其他” 与 “是否在校申请贫困” 存在一定的负相关性。这为模型的降维提供了一定思路。

3.3.2 随机森林特征重要性

随机森林算法可以在有监督学习训练结束后输出特征的重要性排名，作为学习的副产物，它可以计算出优化目标函数过程中最有效率、最有指导意义的特征，可以用于表征特征的贡献率，用于指导建模、特征工程、降维，提高模型的准确率。根据随机森林算法学习获得的特征重要性排名如下图3和4所示：

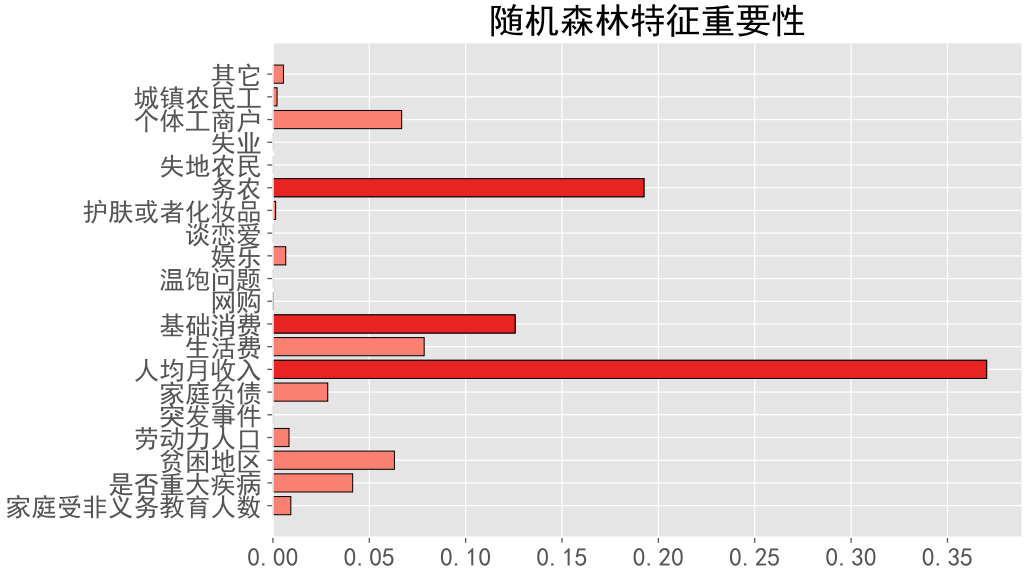

图 3: 随机森林算法输出的特征重要性

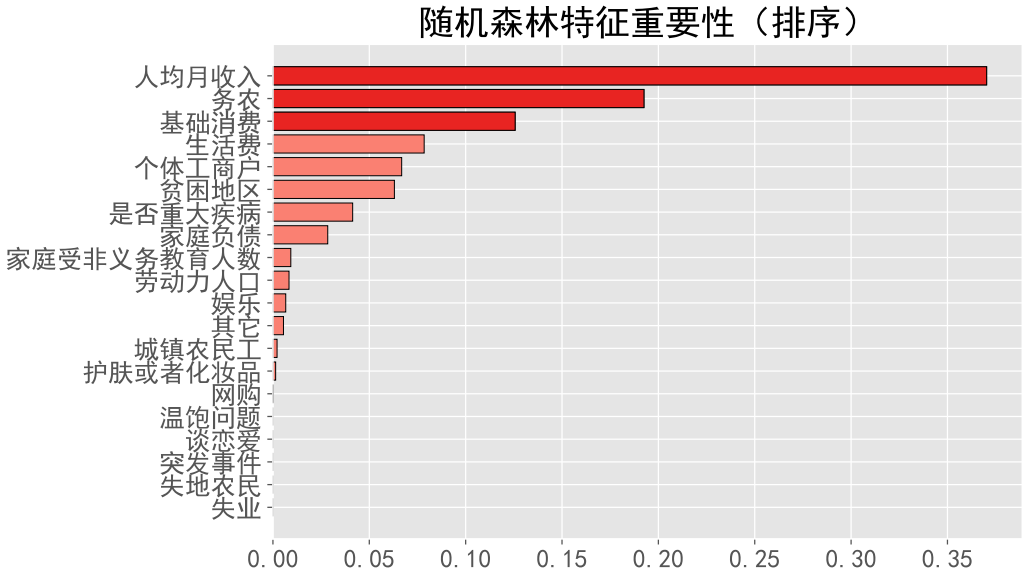

图 4: 随机森林特征重要性排名

由图3和4可知，排在前 5 位的重要特征包括“人均月收入”、“务农”、“基础消费”、“生活费”、“个体工商户”，结合 Spearman 秩相关系数可知，它们对模型的重要性最大；而排在后面的特征，包括“网购”、“温饱问题”、“谈恋爱”、“突发事件”、“失地农民”、“失业”对随机森林回归的重要性几乎为 0，它们对模型的贡献最小，这与相关性分析的结果相近。绘制随机森林特征重要性帕累托分布图如图5所示：

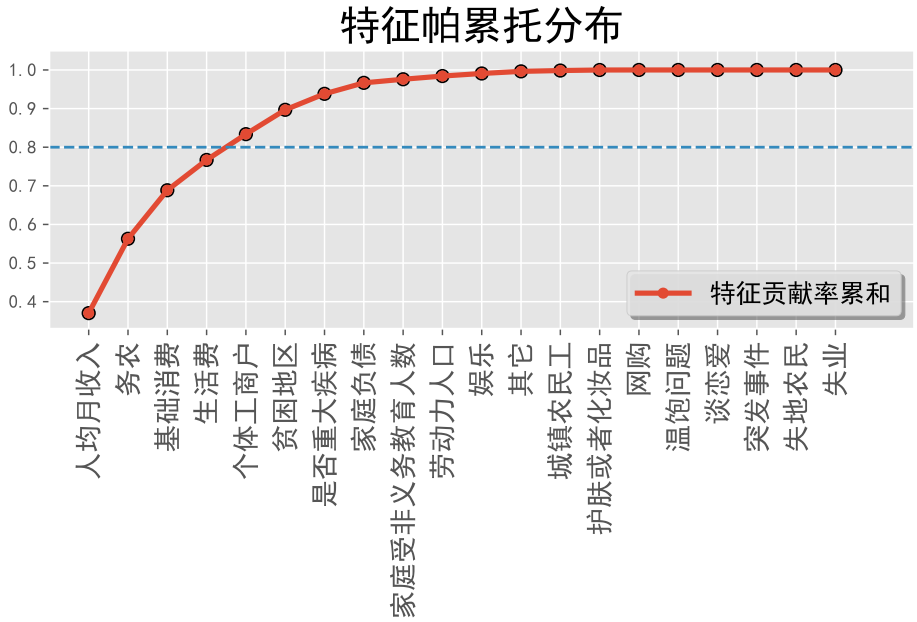

图 5: 随机森林特征贡献帕累托分布

帕累托分布表现了特征对随机森林分类器贡献的累积和，著名的“二八定律”表明分类器的特征往往由 80% 的特征贡献决定，剩下的 20% 可视为冗余特征，它们对分类器的贡献并不如头部的重要特征那样重要。在大样本学习的情形下，减少冗余特征能显著提高学习速度。80% 是一个经验阈值，可以根据实际情况自行确定。由图5可知，贡献排名靠前的 5 个特征其贡献累积和已经超过 80%，前 6 个特征的贡献累积和达到 90%，这对于数据降维具有重要意义。

由于本模型使用的样本集数据量较小，属于小样本学习，考虑到减少特征数可能造成数据过拟合，不一定提高模型的准确率，故在本样本集中未对特征空间 $\mathcal{U}$ 采取降维处理。

3.4 机器学习与深度学习建模

特征工程完成后，作者采用支持向量机、随机森林、XGBoost、神经网络 4 种弱分类器进行二分类求解。样本空间 $\mathcal{X} \in \mathcal{U}$，作者取样本空间

$$\mathcal{X} = \mathcal{U} \backslash \{u_8\}; \mathcal{Y} = \{u_8\};$$

其中 $\{u_8\}$ =“是否在校申请贫困”}。记第 i 个样本对应的标签为 $y^{(i)}$，模型输出的样本标签为 $y_{pre}^{(i)}$，则有待优化损失函数 $\mathcal{F}_{loss}$ 如(3.6)所示：

$$\min \mathcal{F}_{loss} = \sum_{j=1}^{\mathcal{N}_{clf}} f_{loss}^{(j)}, \tag{3.6}$$

其中 $\mathcal{N}_{clf}$ 为分类器个数，此处 $\mathcal{N}_{clf} = 4$；$f_{loss}^{(j)}$ 为每个弱分类器的损失。4 个分类器均采用软分类，也就是获得的输出为 $p^{(i)} = P(y_{pre}^{(i)} = 1) \in [0, 1]$，待判定样本属于正样本的概率；相应的 $P(y_{pre}^{(i)} = 0) = 1 - P(y_{pre}^{(i)} = 1)$。记训练样本集大小为 $|\hat{\Omega}_{train}|$，$f_{loss}^{(j)}$ 采用二分类交叉熵损失 (Cross Entropy Error)，如式(3.7)所示：

$$\begin{aligned} f_{loss}^{(j)} = \frac{1}{|\hat{\Omega}_{train}|} \sum_{i} -[y^{(i)} \cdot log(p^{(i)}) + (1 - y^{(i)}) \cdot log(1 - p^{(i)})]; \\ j = 1, 2, \cdots, \mathcal{N}_{clf}; i = 1, 2, \cdots, |\hat{\Omega}_{train}|. \end{aligned} \tag{3.7}$$

4 个弱分类器通过算法迭代，不断最小化目标函数(3.6)，以最小化判别损失和。

根据学习器输出，判定正负样本的判别式 δ 如式(3.8)所示：

$$\delta(p^{(i)}) = \begin{cases} \text{Positive Sample, if } p^{(i)} \geqslant 0.5 \\ \text{Negative Sample, if } p^{(i)} < 0.5 \end{cases}, \quad \forall i \in \{1, \cdots, |\hat{\Omega}| - |\hat{\Omega}_{train}|\}. \tag{3.8}$$

在融合模型的意义下，模型输出的正样本概率 $p^{(i)}$ 是由 $\mathcal{N}_{clf}$ 个弱学习器加权而成的，其计算方法如式(3.9)所示：

$$p^{(i)} = \sum_{1}^{\mathcal{N}_{clf}} \lambda_j p_j^{(i)}, \tag{3.9}$$

其中权系数满足归一化条件：

$$\sum_{1}^{\mathcal{N}_{clf}} \lambda_j = 1, \lambda_j \geqslant 0; \forall j \in [1, \mathcal{N}_{clf}], j \in \mathbb{N}.$$

树模型往往不需要进行标准化，而支持向量机、神经网络对数据尺度较为敏感。因此在训练支持向量机、神经网络时采用 z-score 标准化，计算式如式(3.10)所示：

$$x^* = \frac{x - \overline{x}}{\sigma}, \tag{3.10}$$

其中 $\overline{x}$ 为原始数据的均值，σ 为原始数据的标准差。

由于在小样本学习情况下 4 种判别器不具有显著差别，我们取定判定权系数 $\lambda_j = 0.25, j = 1, 2, 3, 4$；最终融合学习器在测试集上获得的判定准确率为 88.71%。

4 模型评价

1. 我们的模型在线性核支持向量机下分类效果要好于多项式核支持向量机，这说明我们的模型具有线性可分的优点，这是由于我们采用了子空间嵌入的数据处理方法。事实上，我们的模型在线性核支持向量机下的超平面是很明显的，如图6所示：

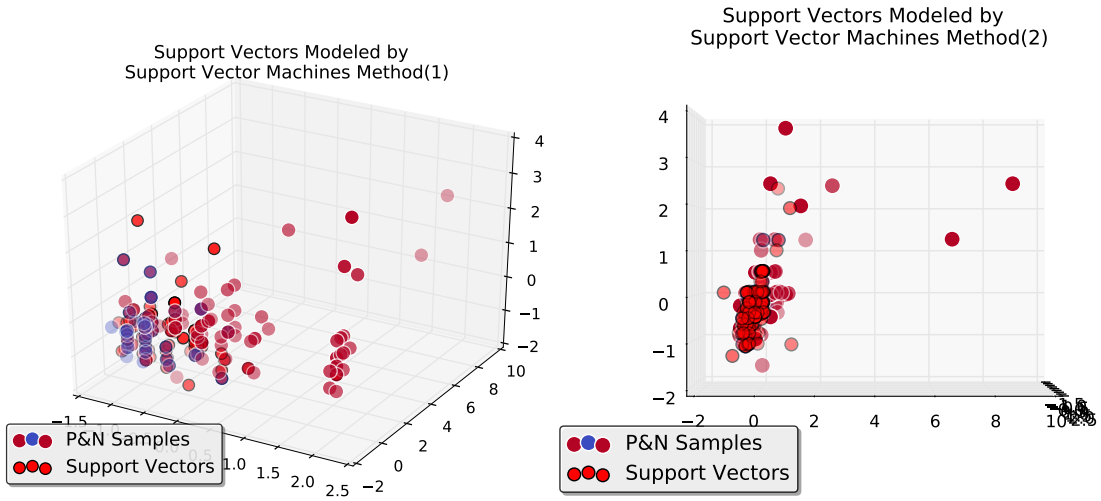

图 6: 线性核支持向量机产生的超平面（以支持向量表征）

2. 我们的模型训练了多个弱学习器，融合为一个学习器，这对于提高判定准确率是有益的，且提高了模型的泛化能力。在大数据情况下，可以采取仅训练部分学习器的方法来减少训练的计算成本，在小样本学习的情况下也可以仅训练在小样本下鲁棒的学习器，针对不同数据量模型具有处理灵活的特点。

参考文献

[1] 李步青. 基于组合 logistic 回归模型的高校贫困生认定研究 [J]. 网络安全技术与应用,2021(01):59-61.

[2] 史建梅, 孔月红. 大数据时代基于模糊层次分析法的高校贫困生认定模型构建 [J]. 文化创新比较研究,2020,4(36):17-19.

[3] 田志磊, 袁连生. 采用非收入变量认定高校家庭经济困难学生的实证研究 [J]. 北京大学教育评论,2010,8(02):145-157+192.

[4] 张良均, 谭立云, 刘名军, 江建名.Python 数据分析与挖掘实战 [M]. 北京: 机械工业出版社，2019.60-61.