3.4 重复值处理

对数据进行检查后发现第 1322 和第 3518 个样本完全一致，部分特征如表 4 所示。为了最大化程度减小过拟合风险，对后面重复的样本进行删除处理，删除后数据集中共计 15999 个样本。

表 4 重复值检验

#	一天去两家医院的天数	就诊的月数	月统筹金额_MAX	个人账户金额_SUM	ALL_SUM	可用账户报销金额_SUM	治疗费用在总金额占比	是否挂号	RES
1322	0	1	248.9	27.66	276.56	27.66	0.061469	0	0
3518	0	1	248.9	27.66	276.56	27.66	0.061469	0	0

3.5 归一化处理

分析原始数据集特征可以发现，不同类型的特征变量取值范围相差很大，例如就诊天数的取值范围一般不会超过 100，比例甚至不会超过 1，但是各类就诊费用却高达几千甚至上万，如果直接使用这些特征数据进行建模会导致模型偏好数值较高的这些特征，从而造成结果的误差。因此为了减小变量取值范围相差较大的影响，需要对特征变量进行无量纲化处理。

常用的无量纲化方法主要有 Min-Max 归一化和 Z-score 标准化：

（1）Min-Max 归一化

该方法是对原始的特征变量进行归一化，将所有的数值映射到[0,1]之间，具体转化方法如公式 1 所示。

$$x^* = \frac{x - \min(x)}{\max(x) - \min(x)}$$

（2）Z-score 标准化

该方法主要是对数据进行标准化，使得数据服从标准正态分布，具体转化方法如公式 2 所示。

$$x^* = \frac{x - \mu}{\sigma}$$

在本文中，由于分类变量进行独热编码后全部为 0 或 1，故选用第一种方法对除了医院编码_NN、出院诊断病种名称_NN、BZ_民政救助、BZ_城乡优抚、是否挂号等分类变量和 RES 标签之外的所有**数值型数据**进行归一化处理，使得所有的医保变量数值都处于[0,1]之间。并且由中心极限定理得，当样本数据量足够大时，独立随机变量的均值趋于正态分布。

归一化之后的部分数据如表 5 所示