计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2022/16(07)-1543-09
doi: 10.3778/j.issn.1673-9418.2101028

面向智慧康养的数据集构建方法及其应用

张麟宇,涂志莹+,杭少石,张柏林,初佃辉
哈尔滨工业大学 计算机科学与技术学院,山东 威海 264209
+ 通信作者 E-mail: tzy_hit@hit.edu.cn

摘 要:互联网和计算机技术的快速发展,使得在人口老龄化的今天发展智慧康养服务成为可能。然而,养老领域的数据问题严重地制约着该领域的智慧化进程。真实数据的缺失、脏数据的干扰以及标准样本过少等问题层出不穷。针对数据集缺失问题,在收集了某市社区康养的小样本数据基础上,提出了一种基于机器学习的三阶段数据生成模型:第一阶段,使用基于树形结构的生成策略,按照原始数据的分布生成了数据集的基础属性;第二阶段,使用朴素贝叶斯算法生成样本的基础行为能力评估指标;第三阶段,在前两个阶段的基础上采用多元线性回归的方法生成高阶行为能力指标以及评估阶段。此外,为验证该模型生成的数据集对下游任务的有效性,在生成数据基础上,利用神经网络设计多个康复训练计划推荐模型,实现5个多分类任务和2个多标签分类任务。通过对实验结果的分析以及专家知识的注入,验证了生成数据的真实性和有效性。

关键词:智慧康养服务;小样本数据;朴素贝叶斯;多元线性回归
文献标志码:A **中图分类号:**TP301.6

Data Set Construction Method for Intelligent Health Care and Its Application

ZHANG Linyu, TU Zhiying+, HANG Shaoshi, ZHANG Bolin, CHU Dianhui
School of Computer Science and Technology, Harbin Institute of Technology, Weihai, Shandong 264209, China

Abstract: The rapid development of Internet and computer technology makes it possible to improve smart health care services in today's aging population. However, there are some data problems that seriously restrict the process of intelligence in the field of elderly care, such as the lack of real data, the interference of dirty data, and too few standard samples. To solve the problem of lacking data set, this paper proposes a three-stage data set construction method based on machine learning on the basis of small sample data which are collected from the community health care in a city. In the first stage, this paper designs a tree structure-based generation strategy to generate the basic attributes of the data set according to the distribution of the original data. In the second stage, this paper obtains the basic behavioral ability evaluation index of the samples with naive Bayesian algorithm. In the third stage, this paper constructs a variety of multiple linear regression equations to get high-order behavioral ability index and evaluation stage on the basis of the first two stages. In order to verify the effectiveness of the data set generated by the model for downstream tasks, this paper designs multiple rehabilitation training plan recommendation models based on the generated data with neural network, and achieves 5 multi-classification tasks and 2 multi-label classification tasks. This paper verifies the authenticity and validity of generated data through analysis of experimental results and expert knowledge.

Key words: smart health care service; small sample data; naive Bayes; multiple linear regression

基金项目:国家重点研发计划(2018YFB1004800);国家自然科学基金(61772159);山东省自然科学基金(ZR2017MF026)。
This work was supported by the National Key Research and Development Program of China (2018YFB1004800), the National Natural Science Foundation of China (61772159) and the Natural Science Foundation of Shandong Province (ZR2017MF026).
收稿日期:2021-01-07 **修回日期:**2021-05-17

真实、有效、完备的数据集意味着机器学习模型将有较好的输入，模型通过学习发现规律，挖掘并分析当中的关联规则与信息，可以很好地为现实中社会生产活动提供知识决策[1-3]。另外，从提升模型的泛化能力出发，也应该相应地增大训练数据的规模。

中国作为世界上最大的发展中国家，人口老龄化程度已经比肩中高收入国家群体，并在未来30年（到2050年）将迅速攀升，超过高收入国家群体[4]。缺乏相关的信息化技术以及成熟的康养公共服务设施的辅助，康养数据的采集和获取是比较困难的。真实、有效的数据集的缺失，成为了研究相关工作的障碍。

针对这一问题，本团队从慢病康复训练指导入手，通过长期的社区公益服务采集了某市的社区康养的标准数据。在此基础上，本文提出了一种基于机器学习的三阶段数据生成模型，以采集到小样本数据集为基础，实现了大批量具有区域养老人群特征的样本数据生成。该模型在第一阶段使用基于树形结构[5]的基础属性生成策略，按照自上而下的思想，生成符合原始数据集分布的基础属性样本；接着提出了基于朴素贝叶斯[6]的基础行为能力指标生成策略，将基础行为能力指标的生成转化为分类问题进行实现；第三阶段，又提出了基于多元线性回归[7]的高阶行为能力指标生成策略，在前两个阶段的基础上，通过选定合适的自变量，拟合9个线性回归方程，完成高阶行为能力指标数据的生成。最后，通过整合三个阶段的结果，完成了康复养老数据的生成工作。

另外，本文利用了模型生成的数据集，设计了基于神经网络的分类推荐模型，在将生成的数据集反馈给康复专家验证、筛选、标注之后，经过属性特征提取，把其输入到模型当中，实现了康复训练计划推荐的任务。

1 相关工作

与传统的机器学习不同，现在基于神经网络的深度学习模型通常采用多层的网络结构，其复杂程度较高，因此也需要尽可能多的数据进行训练。而训练模型所必须的海量训练数据样本难以获取已经成为阻碍深度学习技术进一步推广的一个普遍性难题。目前，学术界提出了很多解决小样本数据集上学习的方法[8-9]。一种常见的思路是把小样本的数据应用到改进后的算法中。文献[10]提出了基于卷积神经网络的小样本图像识别方法，结合了深度学习与迁移学习技术，先在卷积神经网络中对相关领域的大数据集进行预训练，提取预训练模型的权重和样本特征，应用到目标小数据集中对模型进行初始化，然后展开训练。但是该方法受到了相关领域大规模数据集的限制，无法很好地应用到缺少大规模数据集的工作当中。

小样本的模型学习问题，在不使用大数据集辅助的情况下，文献[11]提出了一种迭代提升欠采样模型（under sampling with iteratively boosting，USIB），进行疾病预测。该方法迭代地从多数类样本中进行欠采样，构建多组弱分类器，通过加权组合的方式集成一个强分类器，提高模型的学习能力。但是基于该方法更多地关注错误分类和分类置信度不高的样本去改善模型的预测能力，并没有真正实现数据生成的任务。

集成方法也是解决小样本学习的常用方法，通过融合集成技术和采样技术，充分利用了两者的优点。Liu等人提出了EasyEnsemble集成算法，结合了Bagging和欠采样技术[12]。在此基础上，Liang和Cohn提出了UBagging算法，该算法将Bagging应用到不平衡数据集训练中，不断增加负样本采样数量来训练多个分类器，集成多个分类器提高整体分类性能[13]。融合集成技术和采样技术的算法通过多次采样解决了单次采样中样本信息缺失问题，但是每次随机性地采样，忽略了分类器之间的关系，限制了模型的整体性能。

另一种常用的方法是在已有数据集的基础上，进行特定技术的处理来增加样本的数量。文献[14]提出了一种深度卷积生成对抗网络（deep convolutional generative adversarial networks，DCGANs），通过设计生成器与判别器，学习图像中物体到场景的层次化表征信息，最终生成新的图像数据集。文献[15]提出了一种基于Wasserstein GAN的小样本数据增强方法，使用训练集样本训练GAN后生成模拟样本数据，扩增训练集样本规模。虽然GAN方法的生成不用考虑样本属性间的内在联系，但是GAN在实际应用当中会存在一些问题：

（1）无法进行稳定的训练，导致生成模型生成无意义的输出，对于离散型数据的学习效果较差；

（2）生成的数据的可解释性差，有时GAN生成的样本只是对真实样本的简单改动，导致生成样本的多样性较差。

综上，已有的数据生成方法存在着随机性、盲目

性，并且有模型参数选择和复杂程度的限制。研究新的数据模型，并将其应用到相关工作当中具有重要意义。

2 数据生成模型

本次用于扩充的康复养老数据集包含140条数据，每个样本包含老年人的基础属性、行为能力评估指标以及行为能力评估阶段等信息。

在样本数据方面，通过对采集到的样本数据进行分类，可以得到基础属性、行为能力评估指标和行为能力评估阶段等类别信息，其中每个分类包含的属性如下：

（1）基础属性：姓名、年龄、家庭条件、残疾原因、残疾类型、残疾等级。

（2）基础行为能力评估指标：翻身、坐、站、转移。

（3）高阶行为能力评估指标：步行或驱动轮椅、上下台阶、进食、穿脱衣物、洗漱、入厕、交流、日常家务、社会活动。

（4）行为能力评估阶段：康复初期、恢复期、治愈期。

在数据生成的模型设计中，本文分为了三个阶段：第一阶段按照基于树形结构的规则自上而下生成样本的基础属性；第二阶段对于基础的行为能力指标的生成，使用贝叶斯模型来实现；第三阶段，使用多元线性回归生成高阶行为能力指标。具体的模型设计流程图如图1所示。

2.1 基于树形结构的基础属性生成策略

在原始数据集当中基础属性包括了性别、年龄、残疾类型、残疾原因、残疾等级共5个维度特征，且它们在数据集中都呈现出一定的分布规律。如果要同时生成各个维度的值，则会忽略它们之间的相关性；如果只是简单地用随机的方法生成各个维度的值，则生成的数据会不满足原始数据集的分布，失去有效性和真实性。因此，本节提出了基于树形结构的基础属性生成策略：考虑先根据残疾类型的分布情况，采用改进后的轮盘赌算法确定生成样本的残疾类型，然后以此为根节点，性别特征为其子节点，利用条件分布，再次使用改进后的轮盘赌算法确定性别特征；以此类推，按照树形结构的思想，不断利用条件分布，采用改进后的轮盘赌算法，自上而下地确定各个维度的值，最终实现基础属性的生成。

在确定了生成策略之后，本文对原数据进行了预处理，从原数据集中筛选出真实可用的80条数据。对这些数据的基础属性进行统计分析，样本的分布呈现出一定的规律，各个基础属性统计分布如表1所示。

表1 基础属性统计分布

Table 1 Basic attribute statistical distribution

基础属性	数据包含项	分布比例
性别	（男、女）	[48∶32]
年龄	（按照年龄段划分）[0,44],[45,59],60以上	[15∶37∶28]
残疾类型	（偏瘫、截瘫、脑瘫等）	[59∶6∶3∶3…]
残疾原因	（疾病、感染、产伤等）	[40∶10∶8∶5…]
残疾等级	（一级、二级、三级等）	[10∶10∶20∶40]

常见的轮盘赌算法通常需要先计算适应度比例，即对于数量为 N 的养老康复样本，给每个个体 x_i 一个适应度值 $f(x_i)$，则每个特征值的选择概率为：

$$p(x_i)=\frac{f(x_i)}{\sum_{j=1}^{N} f(x_j)} \tag{1}$$

然后计算每个个体的累计概率，即每个个体之

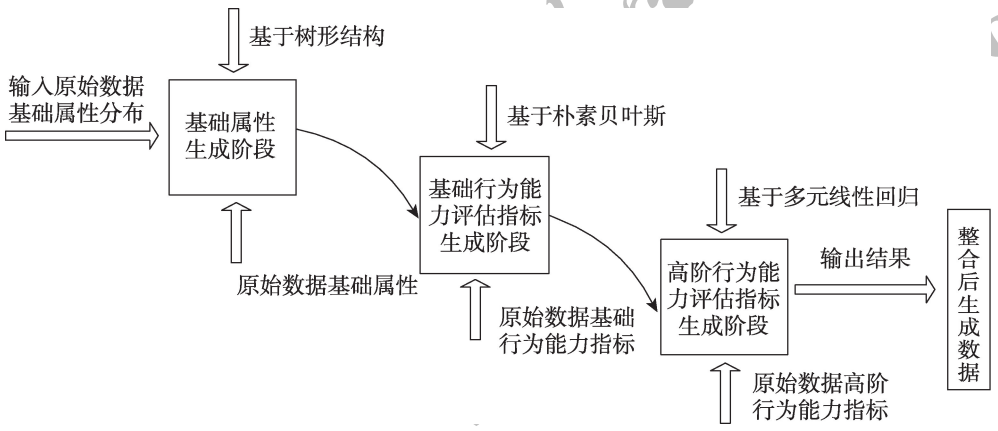

图1 数据生成模型

Fig.1 Data generation model

前所有个体的选择概率之和：

$$q_i=\sum_{j=1}^{i}f(x_j) \tag{2}$$

在确定了累计概率之后，随机生成一个数组 ra，数组的长度为 N，元素值的范围属于[0,1]，然后有序排列，用于确定个体是否能够被选择。若累计概率 q_i 大于随机生成数组中的 $ra[i]$，则 x_i 被选择，将继续比较 $ra[i]$，若不大于，则不选择，比较下一个个体 x_i+1，以此类推，从而确定生成属性。

在样本的数据集中，要生成的基础属性中的特征值只有一个，在轮盘赌算法中即每次需要选择出来的个体只有一个，为了能够方便地控制选择的个体的数量，本文改进了轮盘赌算法，在函数参数中增加了一个控制生成个体数量的参数 n。具体的算法如下所示。

算法 1 基于条件分布的轮盘赌算法

输入：残疾类型分布数组 T，性别条件分布数组 S，年龄条件分布数组 A，残疾原因条件分布数组 R，残疾等级条件分布数组 L。

1. 根据式(1)计算每个个体的选择概率，得到选择概率数组 C 及其长度 m；

2. 根据式(2)计算累计概率，得到累计概率数组 P；

3. 生成随机数组 ra；

4. 比较选择，返回结果

```
for i = 1,2,⋯,n
    for j = 1,2,⋯,m
      //比较，并选择个体
        If ra[i]<=P[0]:
            Return 0
        If P[j]<ra[i]<=P[j+1]:
            Return j+1
      End for
End for
```

通过对原始数据集的处理、分析、统计，按照树形结构的思想，采用改进后的轮盘赌算法，实现了基础属性的生成。

2.2 基于朴素贝叶斯基础行为能力指标生成策略

贝叶斯方法是以贝叶斯原理为基础，使用概率统计的知识对样本数据集进行分类，因此有着较好的统计和数学基础，分类的准确率较高。该方法通过使用数据集中统计出的先验概率和后验概率，既避免了只使用先验知识的主观偏见，也避免了单独使用样本信息的过拟合现象。

朴素贝叶斯分类，以贝叶斯定理为基础，并且使用条件独立性假设的方法，先通过已给定的训练集，以特征属性之间独立作为前提假设，学习从输入到输出的联合概率分布，再基于学习到的模型，输入 X 求出使得后验概率最大的输出 Y。

设有样本数据集 $S=\{s_1,s_2,\cdots,s_n\}$，对应样本数据的特征属性集 $X=\{x_1,x_2,\cdots,x_d\}$，且类变量为 $Y=\{y_1,y_2,\cdots,y_m\}$，即 S 可以分为 y_m 个类别。其中 $x_1,x_2,\cdots,x_d$ 相互独立且随机，则 Y 的先验概率 $P_{\text{prior}}=P(Y)$，Y 的后验概率 $P_{\text{post}}=P(Y|X)$。由朴素贝叶斯算法可得，后验概率可以由先验概率、证据 $P(X)$、类条件概率 $P(X|Y)$ 以及在给定样本类别 y 时计算得出公式如下：

$$P(Y|X)=\frac{P(Y)P(X|Y)}{P(X)} \tag{3}$$

$$P(X|Y=y)=\prod_{i=1}^{d}P(x_i|Y=y) \tag{4}$$

由以上两公式可以得出后验概率为：

$$P_{\text{post}}=P(Y|X)=\frac{P(Y)\prod_{i=1}^{d}P(x_i|Y)}{P(X)} \tag{5}$$

由于在每次的计算过程中 $P(X)$ 的大小是一样的，在比较后验概率的时候，只比较上式的分子部分即可。最终可以得到一个样本数据属于类别 y_i 的朴素贝叶斯计算公式：

$$P(y_i|x_1,x_2,\cdots,x_d)=\frac{P(y_i)\prod_{j=1}^{d}P(x_j|y_i)}{\prod_{j=1}^{d}P(x_j)} \tag{6}$$

在样本的康复养老数据集中，样本的基础行为能力评估指标包含翻身、坐、站、转移共四项，每项指标的评估分为 0、1、2、3 四个等级，评估得分越高表明该项指标的能力越强。因此，对于每项基础行为能力评估指标的生成，可以看作一个分类问题。通过计算原数据集中样本的基础属性（性别、年龄、残疾类型、残疾原因、残疾等级）之间的相关性，如表 2 所示的基础属性的 Spearman 相关性系数矩阵，可以发现它们之间有较低的关联程度，即使用朴素贝叶斯算法，考虑每个特征之间的独立性假设是合理的。

在分析了样本基础属性间的基础属性后，通过预先设定好的数值化规则进行基础属性约束，设原始数据集的基础属性的数值化矩阵 $\boldsymbol{M}1$，原始数据集的基础行为能力评估指标的数值化标签数组 $L1$，阶段一中生成的基础属性的数值化矩阵 $\boldsymbol{M}2$；将 $\boldsymbol{M}1$、$\boldsymbol{M}2$ 以及 $L1$ 输入到朴素贝叶斯模型中，最终得到模型

表2 样本基础属性Spearman相关性系数矩阵

Table 2 Spearman correlation coefficient matrix of sample basic attributes

基础属性	性别	年龄	类型	原因	等级
性别	1.000	0.127	−0.013	−0.073	0.073
年龄	0.127	1.000	−0.087	0.050	−0.029
类型	−0.013	−0.087	1.000	0.545	0.105
原因	−0.073	0.050	0.545	1.000	0.157
等级	0.073	−0.029	0.105	0.157	1.000

预测出的基础行为能力指标。

数据生成的第二阶段，本文使用朴素贝叶斯算法，通过原数据集的训练，分别得到翻身、坐、站、转移四个基础行为能力评估指标的生成模型，再利用阶段一中已生成的基础属性，最终得到每个生成样本的基础行为能力评估指标。

2.3 基于多元线性回归高阶行为能力指标生成策略

在回归分析中，如果有两个或两个以上的自变量，就称为多元回归。在现实中，一个现象结果的出现往往是与多个因素相联系的，由多个自变量的最优组合共同来预测或估计因变量，比只用一个自变量进行预测或估计更有效，更符合实际，因此多元线性回归的应用场合常常更为广泛。多元线性回归模型如下：

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m+e \tag{7}$$

式中，β_0 为常数项，$\beta_i(i=1,2,\cdots,m)$ 表示在其他变量保持不变时，X_i 增加或减少一个单位时 Y 的平均变化量，被称为偏回归系数。同样，e 被称为残差，表示去除 m 个自变量对 Y 影响后的随机误差。通常，多元线性回归模型的应用需要满足如下条件：

（1）Y 与 $X_1,X_2,\cdots,X_n$ 之间具有线性关系；

（2）各个样本的观察值 $Y_j(j=1,2,\cdots,n)$ 相互独立；

（3）残差 e 服从均值为0，方差为 σ_2 的正态分布，等价于对任意一组自变量 $X_1,X_2,\cdots,X_m$ 值，因变量 Y 具有相同的方差，并且服从正态分布。

使用最小二乘法，根据样本数据求得模型参数估计值：

$$\hat{Y}=b_0+b_1X_1+b_2X_2+\cdots+b_mX_m \tag{8}$$

通过建立多元线性回归方程求解：

$$\begin{cases} l_{11}b_1+l_{12}b_2+\cdots+l_{1m}b_m=l_{1Y} \\ l_{21}b_1+l_{22}b_2+\cdots+l_{2m}b_m=l_{2Y} \\ \vdots \\ l_{m1}b_1+l_{m2}b_2+\cdots+l_{mm}b_m=l_{mY} \end{cases} \tag{9}$$

$$b_0=\bar{Y}-(b_1\bar{X}_1+b_2\bar{X}_2+\cdots+b_m\bar{X}_m) \tag{10}$$

最后确定 $b_0,b_1,\cdots,b_m$ 的值，得到最终的多元线性回归方程。

在康复养老的数据集中，高阶行为能力评估指标包含步行或驱动轮椅、上下台阶、进食、穿脱衣物、洗漱、入厕、交流、日常家务、社会活动等九项指标，每项指标分为0、1、2、3四个等级，得分越高表示该项能力越强。通过统计原数据各项能力指标的相关性，本文发现基础行为能力指标与高阶行为能力指标之间有较强的相关性。当指标得分被看作连续性数值时，两者具有一定的线性关系。基于上述分析，在数据生成的第三阶段，本文采用多元线性回归算法，通过对原始数据集的训练，针对不同的高阶行为能力指标，分别构建了对应的回归方程。

在自变量的选择过程中，本文采取了逐步回归法进行筛选。该方法将前进法和后退法相结合，首先使用前进法挑选变量，然后将已入选的自变量使用后退法进行剔除，在整个过程中，通过观察实验中设定的相关检验标准，选择和剔除合适的自变量，最后建立较优的回归方程。

通过统计和实验发现，当同时引入翻身、坐、站、转移四项作为自变量放入方程当中，多元线性回归模型的效果最好。在完成模型的训练之后，将第二阶段生成的基础行为能力评估指标数据作为输入，可以完成高阶行为能力评估指标的生成。然后根据整个行为能力评估指标可以得到评估阶段；最后整合每个生成样本的基础属性、行为能力评估指标以及评估阶段可以得到一个完整的生成数据集。

3 康复数据的应用

在现实中，在得到了一个样本的基础属性、行为能力评估指标以及行为能力评估阶段后，康复专家就可以根据这些特征进行一些康复训练计划的推荐，用于辅助患者的康复治疗。因此，将模型生成的数据集反馈给康复专家，经过专家的评审、筛选和标注，最后可以形成一批标注后的完整数据集。在此基础上，可以设计一个模型，用于康复计划的推荐。

在得到了样本的数据信息后，统计需要推荐的项目包含运动康复目标、生活自理能力康复目标、生活适应能力康复目标、康复训练项目、康复疗法、康复训练强度、康复训练组数共7项。

通过对样本数据的整理、统计后，得到的具体推荐数据如表3所示。

表3 推荐模型样本数据展示

Table 3 Sample data presentation of recommended model

样本属性	样本属性名称	解释
基本信息	性别	男、女
	年龄	样本的年龄
	残疾类型	偏瘫、截瘫等
	残疾原因	疾病、产伤等
	残疾等级	中国残疾人等级划分
	行为能力评估指标	翻身、坐等13项
	行为能力评估阶段	分为三个阶段
康复计划	运动康复目标	如能进行身体锻炼
	生活自理能力康复	如能独立吃饭
	生活适应能力康复	如能独立参加活动
	康复训练项目	如翻身、坐、站等
	康复疗法	如理疗、运动疗法
	康复训练强度	如每次训练15 min
	康复训练组数	如每周做3组

其中，虽然每项计划的内容为文本数据，但是内容的划分是分类别的。因此考虑构建基于神经网络的分类模型，用于实现康复训练计划的推荐。

具体的推荐任务划分为多分类任务和多标签分类任务。根据康复训练计划数据的特点，其中运动康复目标、生活自理能力康复目标、生活适应能力康复目标、康复训练强度、康复训练组数五项推荐属于多分类任务，康复训练项目与康复疗法的推荐属于多标签分类任务。

对于每个样本而言，它都包含性别、年龄、残疾类型、残疾原因、残疾等级、行为能力评估指标以及行为能力评估阶段七项基本信息，在分类推荐模型设计之前，需要先确定样本的特征属性和模型的输入。在现实中，专业的康复医护人员根据残疾人的各项生理特征以及康复过程信息进行康复计划的制定，在与康复专家沟通之后，通过分析残疾人的特征属性和影响康复计划推荐的主要因素，本文设计了一个基于神经网络的样本特征提取模型。具体的特征向量提取模型设计如图2所示。

首先，经过数据的预处理，将样本的基本信息中的属性数字化，然后将数字化的特征通过嵌入层的映射变换为16维或32维的低维特征。使用 $X=\{x_1, x_2,\cdots,x_n\}$ 表示样本基础信息中的各个特征项，通过激活函数ReLU的非线性变换得到低维特征，接着通过全连接层将各个低维特征拼接融合，得到样本特征，放入到隐藏层当中，最终得到400维的高阶融合特征向量 $\boldsymbol{v}$。

在网络结构的设计中，本文将特征提取模型中得到的向量 $\boldsymbol{v}$ 作为分类网络的输入层，通过隐藏层的加工抽象，选择合适的损失函数完成输出。

在网络的输出层中，对于不同的分类任务，模型使用了不同的损失函数。针对多分类任务，模型使用了Softmax交叉熵损失函数，针对多标签分类任务，数据中类别标签是独立且不互斥的，因此可以将其视为多个二分类任务，使用Sigmoid交叉熵损失函数。

4 实验结果与分析

4.1 基于朴素贝叶斯的基础行为能力指标生成策略实验

在现实中，康复师对患者进行行为能力评估时存在较多的不确定性和主观因素，有时评估指标得分在0~4之间并没有严格的区分度。例如当一个样本中“站”这项行为能力指标的真实值为2，其作为测试数据放入模型当中，被预测出的结果为0或1或3时，都应该给其一定的正确权重，而不是直接判错。即在分析阶段二的生成模型的准确率时不能完全按

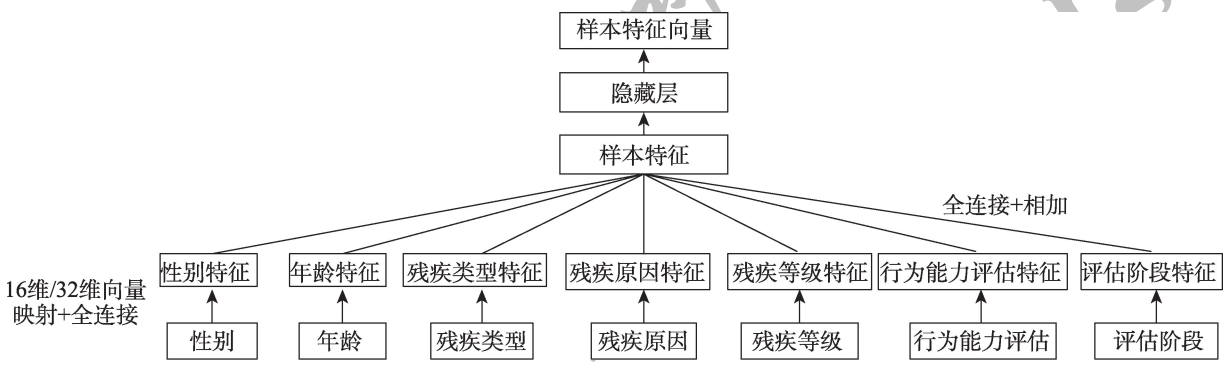

图2 样本特征提取模型

Fig.2 Sample feature extraction model

照分类问题的准确率的计算来定义。

通过与康复养老方面的专家沟通，本文定义了正确程度评价矩阵，用来合理地计算模型的准确性。

对于数据规模为 n 的测试数据集 T，T_{ij} 表示第 i 个测试样本的第 j 个评估指标的真实值，P_{ij} 表示第 i 个测试样本的第 j 个评估指标的预测值，则第 i 个测试样本第 j 个评估指标被预测的正确程度 D_{ij} 为：

$$D_{ij}=1-\frac{|T_{ij}-P_{ij}|}{4} \tag{11}$$

那么，第 j 个指标预测模型的准确率计算为：

$$Accuracy=\frac{\sum_{i=1}^{n}D_{ij}}{n} \tag{12}$$

利用上述公式，得到正确程度评价矩阵如表4所示。

表4 正确程度评价矩阵

Table 4 Matrix of correctness degree evaluation

指标得分	0	1	2	3
0	1.00	0.75	0.50	0.25
1	0.75	1.00	0.75	0.50
2	0.50	0.75	1.00	0.75
3	0.25	0.50	0.75	1.00

在设计好了正确程度矩阵，改进了评价的标准之后，对模型的准确率进行了实验验证，设定实验迭代次数为10，分别计算了未使用正确程度评价矩阵和改进后的评价标准，绘制出了“翻身”基础行为能力指标的生成模型准确率折线图，如图3所示。

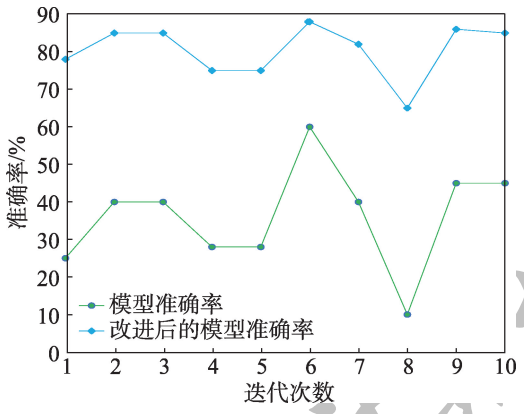

图3 翻身行为能力指标生成实验结果

Fig.3 Experimental results of ability index of turning over

从图3中可以发现，在使用了定义的正确程度评价矩阵衡量之后，模型的准确率可以达到80%。在改进了阶段二实验的评价算法之后得到了较高的模型准确率，说明阶段二的生成是可解释的。

4.2 基于多元线性回归的高阶行为能力指标生成策略实验

在第三阶段生成高阶行为能力指标时，本文总共构建了9个回归方程，因变量整体对方程的解释使用了 R^2 和 F 值，各个变量对方程的显著性影响使用了 p 值，并且统计列出了各个方程自变量的系数。其中 X_1、X_2、X_3、X_4 分别为自变量翻身、坐、站、转移，各个方程的 R^2 和 F 值如表5所示。

表5 各个回归方程 R^2 与 F 值统计表

Table 5 Each regression equation R^2 and F value

回归方程	自变量	R^2	F 值
1	X_1, X_2, X_3, X_4	0.579	29.190
2	X_1, X_2, X_3, X_4	0.612	32.140
3	X_1, X_2, X_3, X_4	0.376	1.820
4	X_1, X_2, X_3, X_4	0.078	1.788
5	X_1, X_2, X_3, X_4	0.280	8.260
6	X_1, X_2, X_3, X_4	0.301	9.136
7	X_1, X_2, X_3, X_4	0.344	11.130
8	X_1, X_2, X_3, X_4	0.456	17.780
9	X_1, X_2, X_3, X_4	0.347	11.300

其中，以翻身、坐、站、转移为自变量，步行或驱动轮椅为因变量，拟合出的回归方程的系数、标准差、t 值、p 值的结果如表6所示。

表6 回归方程1的实验结果

Table 6 Experimental results of regression equation 1

参数	系数	标准差	t 值	$p>\|t\|$
常量	−0.021 4	0.150	−0.142	0.887
X_1	0.266 3	0.081	3.290	0.001
X_2	0.118 3	0.091	−1.303	0.096
X_3	0.223 7	0.102	2.191	0.031
X_4	0.510 1	0.094	5.437	0.001

从表5中可以发现，R^2 和 F 值最高达到0.612和32.140，说明选取的变量整体可以对方程进行解释。在衡量每个变量对方程影响的显著性时使用了 t 检验，其中当 p 值小于0.05时，表示拒绝原假设，即表明该自变量与因变量有一定的回归关系，且对方程有较高的显著性影响。

4.3 三阶段的生成模型实验

在生成阶段一中，实验保证了生成出的数据是符合原数据集分布的，在生成阶段二和阶段三中也选取了合适的实验评价指标，保证了其结果的可靠性。基于以上工作，还需要对整个生成模型的实验结果进行分析和评估。因此，本文设计了Spearman

相关性系数矩阵余弦相似度计算的方法进行实现。

先计算出原数据集中各个特征维度之间的Spearman相关性系数矩阵 $\boldsymbol{M}1$，然后计算生成数据集的各个特征维度之间的Spearman相关性系数矩阵 $\boldsymbol{M}2$，之后将两者做余弦相似度计算，得到 $\boldsymbol{M}1$ 和 $\boldsymbol{M}2$ 之间的相似度用来衡量生成数据的质量。

在具体的操作过程中，本文设定了不同的数据集生成的Batch Size，生成数据集的大小分别从100到1 000，控制每批次生成数据规模的大小。此外，在每次得到生成的数据集之后，计算数据集的Spearman相关系数矩阵，然后统计出了3个不同的矩阵相似度，用来与文中提出的模型的实验结果进行对比。其中 sim_1 表示生成阶段二、三都使用朴素贝叶斯的方法后矩阵 $\boldsymbol{M}1$ 和 $\boldsymbol{M}2$ 的相似度；sim_2 表示生成阶段二使用朴素贝叶斯，阶段三使用多元线性回归并将高阶行为能力指标得分四舍五入后矩阵 $\boldsymbol{M}1$ 和 $\boldsymbol{M}2$ 的相似度；sim_3 则表示生成阶段二使用朴素贝叶斯，阶段三使用多元线性回归后矩阵 $\boldsymbol{M}1$ 和 $\boldsymbol{M}2$ 的相似度。最后，在得到了每次实验的结果之后，统计并绘制了实验结果折线图如图4所示。

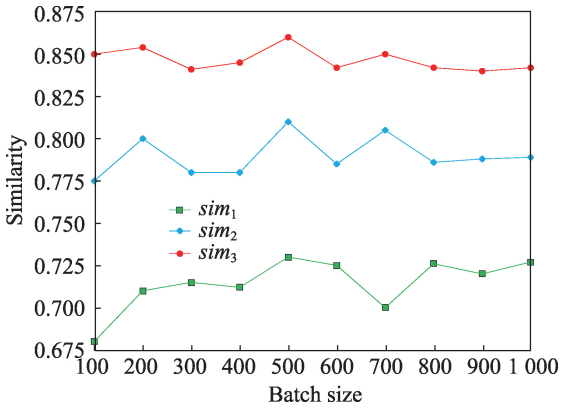

图4 矩阵相似度计算折线图

Fig.4 Result of matrix similarity calculation

从图4的数据可以看出，相似度会有极值点的出现，但随着生成数据集规模的增大，矩阵相似度趋于稳定，且具体表现为 sim_1、sim_2 以及 sim_3 分别在0.725、0.800、0.850这3个值上下波动。其中 sim_3 的值最大，说明通过本文提出的模型生成出的数据很大程度上和原数据集在各个特征维度的相关性上也保持了一致，从实验结果上验证了生成数据集的真实性和可靠性。

4.4 康复推荐模型实验

在本节的实验中，使用了三阶段生成模型并经过康复专家标注后的数据集。在具体的实验过程中，为了避免随机性对实验结果造成影响，本文做了5次实验，每次实验随机选取80%的数据作为训练数据集，剩下的20%作为测试数据集，取5次实验的平均值作为最终结果。多分类任务1~5分别为运动康复目标推荐、生活自理能力康复目标推荐、生活适应能力康复目标推荐、康复训练强度推荐、康复训练组数推荐，本文使用准确率Acc进行评估；对于多标签分类任务，本文使用AUC（area under curve）进行评估。得到的实验结果如表7、表8所示。

表7 多分类任务的实验结果

Table 7 Experimental results of multi-classification tasks %

评价指标	任务1	任务2	任务3	任务4	任务5
Acc	64	72	68	75	77

表8 多标签分类任务的实验结果

Table 8 Experimental results of multi-label classification tasks %

评价指标	康复训练项目	康复疗法
AUC	63	65

从实验结果来看，多分类任务的Acc可以达到77%，多标签分类任务的AUC可以达到65%，说明分类推荐模型有较好的效果，可以为后续的研究提供一些参考。

5 总结与展望

本文提出了一种基于机器学习的三阶段数据生成模型。实验表明，生成模型的第一个阶段保证了生成后的数据集和原数据集有相同的属性分布；在第二阶段，通过设计正确程度矩阵验证了基础行为能力指标的生成结果可以达到80%；生成阶段三提出的基于多元线性回归的高阶行为能力指标生成策略保证了生成数据集继承了原始数据集属性之间的相关性。此外，通过注入专家知识，本文有效地筛选和标注了生成数据，在此基础上，实现的多分类任务的Acc可以达到77%，多标签分类任务的AUC可以达到65%。

尽管本文所提出的基于机器学习的三阶段生成模型可以生成一个完备有效的数据集，但是目前对生成数据集的利用有限。后续将进一步优化生成模型，并在相关的系统平台中开放相关数据集和模型接口，以便在此基础上做更多的研究工作。

参考文献：

[1] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[2] 郭丽丽, 丁世飞. 深度学习研究进展[J]. 计算机科学, 2015, 42(5): 28-33.
GUO L L, DING S F. Reaserch progress on deep learning[J]. Computer Science, 2015, 42(5): 28-33.

[3] 刘全, 翟建伟, 章宗长, 等. 深度强化学习综述[J]. 计算机学报, 2018, 41(1): 1-27.
LIU Q, ZHAI J W, ZHANG Z Z, et al. A survey on deep reinforcement learning[J]. Chinese Journal of Computers, 2018, 41(1): 1-27.

[4] 刘尚希, 赵福昌, 侯海波. 中国人口老龄化、经济增长与社会化改革[J]. 发展研究, 2020(10): 4-9.
LIU S X, ZHAO F C, HOU H B. Aging of Chinese population, economic growth and social reform[J]. Development Research, 2020(10): 4-9.

[5] LI J, YANG S, WANG X, et al. Tree-structured data regeneration with network coding in distributed storage systems [C]//Proceedings of the 17th International Workshop on Quality of Service, Charleston, Jul 13-15, 2009. Piscataway: IEEE, 2010: 1-9.

[6] HUAI M, HUANG L S, YANG W, et al. Privacy-preserving naïve Bayes classification[C]//LNCS 9403: Proceedings of the 8th International Conference on Knowledge Science, Engineering and Management, Chongqing, Oct 28-30, 2015. Cham: Springer, 2015: 627-638.

[7] ISLAM M Q, TIKU M L. Multiple linear regression model under nonnormality[J]. Communications in Statistics, 2005, 33 (10): 2443-2467.

[8] HE H B, GARCIA E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21 (9): 1263-1284.

[9] BRANCO P, TORGO L, RIBEIRO R P. A survey of predictive modeling on imbalanced domains[J]. ACM Computing Surveys, 2017, 49(2): 1-50.

[10] 段萌, 王功鹏, 牛常勇. 基于卷积神经网络的小样本图像识别方法[J]. 计算机工程与设计, 2018, 39(1): 224-229.
DUAN M, WANG G P, NIU C Y. Method of small sample size image recognition based on convolution neural network [J]. Computer Engineering and Design, 2018, 39(1): 224-229.

[11] 陈旭, 刘鹏鹤, 孙毓忠, 等. 面向不均衡医学数据集的疾病预测模型研究[J]. 计算机学报, 2019, 42(3): 596-609.
CHEN X, LIU P H, SUN Y Z, et al. Research on disease prediction models based on imbalanced medical data sets [J]. Chinese Journal of Computers, 2019, 42(3): 596-609.

[12] LIU X Y, WU J X, ZHOU Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems, Man, and Cybernetics: Part B, 2009, 39(2): 539-550.

[13] LIANG G H, COHN A G. An effective approach for imbalanced classification: unevenly balanced Bagging[C]//Proceedings of the 27th AAAI Conference on Artificial Intelligence, Bellevue, Jul 14-18, 2013. Menlo Park: AAAI, 2013: 1633-1634.

[14] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv:1511.06434, 2015.

[15] 刘宇飞, 周源, 刘欣, 等. 基于 Wasserstein GAN 的新一代人工智能小样本数据增强方法——以生物领域癌症分期数据为例[J]. 工程, 2019, 5(1): 156-163.
LIU Y F, ZHOU Y, LIU X, et al. Wasserstein GAN-based small-sample augmentation for new-generation artificial intelligence: a case study of cancer-staging data in biological [J]. Engineering, 2019, 5(1): 156-163.

张麟宇(1997—),男,硕士研究生,主要研究方向为服务计算、知识图谱。

ZHANG Linyu, born in 1997, M.S. candidate. His research interests include service computing and knowledge graph.

涂志莹(1983—),男,博士,副教授,CCF 会员,主要研究方向为服务计算、知识工程。

TU Zhiying, born in 1983, Ph.D., associate professor, member of CCF. His research interests include service computing and knowledge engineering.

杭少石(1996—),男,硕士研究生,主要研究方向为服务计算、知识图谱。

HANG Shaoshi, born in 1996, M.S. candidate. His research interests include service computing and knowledge graph.

张柏林(1997—),男,硕士研究生,主要研究方向为机器学习、服务计算。

ZHANG Bolin, born in 1997, M.S. candidate. His research interests include machine learning and service computing.

初佃辉(1970—),男,博士,教授,CCF 会员,主要研究方向为服务计算、知识工程。

CHU Dianhui, born in 1970, Ph.D., professor, member of CCF. His research interests include service computing and knowledge engineering.