文章编号:1672-3961(2024)04-0076-10 DOI:10.6040/j.issn.1672-3961.0.2024.057

基于异常点检测的心理健康辅助诊断方法

乔慧妍[1],段学龙[1],解驰皓[2],赵冬慧[1],马玉玲[1*]
(1.山东建筑大学计算机科学与技术学院, 山东 济南 250101; 2.聆心云(山东)智能科技有限公司, 山东 济南 250013)

摘要:采用异常点检测算法研究心理健康辅助诊断任务,提出并设计一种基于异常点检测的心理健康辅助诊断方法,有效识别心理沙盘数据中的异常样本。在构建心理健康辅助诊断模型过程中,分析数据特性,提取与用户心理健康状况高度相关的特征,构建虚拟心理沙盘数据集;使用 4 种传统异常点检测算法,识别沙盘数据集中异常样本,设计融合策略,集成不同算法检测结果,提高异常样本检测精准性和效率,辅助人类专家进行精确诊断;对模型预测性能和结果进行详细分析,结合基线模型进行对比评价。试验结果表明,基于异常点检测的心理健康辅助诊断方法在沙具使用相似度、距离度量、聚类性能等 3 项指标上获得较好性能。

关键词:心理健康辅助诊断;虚拟心理沙盘;机器学习;异常点检测;心理健康
中图分类号:TP391 **文献标志码**:A
引用格式:乔慧妍,段学龙,解驰皓,等. 基于异常点检测的心理健康辅助诊断方法[J]. 山东大学学报(工学版),2024,54(4):76-85.
QIAO Huiyan, DUAN Xuelong, XIE Chihao, et al. Approach of assisted diagnosis for mental health based on outlier detection[J]. Journal of Shandong University (Engineering Science), 2024, 54(4):76-85.

Approach of assisted diagnosis for mental health based on outlier detection

QIAO Huiyan[1], DUAN Xuelong[1], XIE Chihao[2], ZHAO Donghui[1], MA Yuling[1*]
(1. School of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, Shandong, China; 2. Lingxinyun (Shandong) Intelligent Technology Co., Ltd., Jinan 250013, Shandong, China)

Abstract: Towarding the assisted diagnosis for mental health research problem, outlier detection was employed to propose and design an outlier detection-based approach for mental health assisted diagnosis, which could effectively identify outliers from large amounts of virtual mental sandbox samples. During constructing the assisted diagnostic model for mental health, the data characteristics were analyzed, and meanwhile the features highly correlated with the user′s mental health status were extracted to establish a virtual psychological sandplay dataset. Four traditional outlier detection algorithms were utilized to identify abnormal samples in the sandplay dataset. A fusion strategy was designed, and the detection results of different algorithms were integrated to enhance the accuracy and efficiency of anomaly sample detection, to assist human experts in making precise diagnoses. The performance and result of the proposed approach were analyzed in details through comparing with the baseline models. The experimental results showed that compared with other models, the assisted diagnosis for mental health based on outlier detection approach had better performance in three indexes such as similarity of sand tool usage, distance of mean vector and clustering performance.

Keywords: assisted diagnosis for mental health; virtual mental sandbox; machine learning; outlier detection; mental health

收稿日期:2024-03-18
基金项目:国家自然科学基金资助项目(62177031,62077033);山东省自然科学基金资助项目(ZR2021MF044);山东省教育教学研究课题资助项目(2021JXY012);教育部产学合作协同育人项目(202102423045);2023 年度教育部人文社会科学研究专项任务资助项目(高校辅导员研究)(2023JDSZ3174);2023 年度济南市市校融合发展战略工程资助项目(JNSX2023064)
第一作者简介:乔慧妍(1998—),女,河南焦作人,硕士研究生,主要研究方向为教育数据挖掘。E-mail:hyqiao0205@163.com
*通信作者简介:马玉玲(1979—),女,河南濮阳人,副教授,硕士生导师,博士,主要研究方向为机器学习与教育大数据挖掘。
E-mail:mayuling20@sdjzu.edu.cn

0　引言

随着国家现代化进程推进以及生活节奏加快,人们在工作、就业、生活等方面压力逐渐变大,具有心理健康问题人数与日俱增。据世界卫生组织统计,2019 年,全世界约 9.7 亿人面临着焦虑、抑郁等精神健康问题,约四分之一的人会在人生某个阶段经历心理问题或精神障碍[1]。我国国民心理健康问题也日益严峻。据统计,我国有 1 亿以上人存在心理健康问题,其中高达 5 400 万人患有抑郁症,占到我国总人口数的 4.2%[1]。心理健康问题不仅影响到患者个人工作、生活,乃至生命安全,还严重影响国民经济高质量发展,已经引起国家社会高度重视。

对于心理健康问题,早期发现和诊断是最为关键的一个环节,发现越早,治疗效果越好。目前,心理健康早期诊断技术主要有以下 2 种:(1)专家诊断法。专家与咨询者面对面或采用远程心理服务方式对用户心理状况进行诊断,诊断结果往往比较精准。尤其是在新型冠状病毒肺炎爆发期间,远程心理服务提高了心理服务效率[2]。不足之处在于,该方法成本较高,耗时耗力,外加专家人数非常有限,难以在大范围人群中展开。(2)问卷和量表。目前心理健康诊断多是基于调查问卷及心理健康自测量表,例如症状自评量表[3]、抑郁症自评量表、焦虑自评量表等。该方法比较简单,成本较低,在大规模心理健康普查工作中得到较广泛应用。这类方式主要依据受访者自我评估报告,所得结论易受到受访个体主观意识影响[4]。随着人们对心理健康问题越来越重视,具有心理服务需求的人数与日俱增,心理咨询机构或部门积累了海量用户心理健康数据。如何利用机器学习、大数据等相关技术,从海量心理健康相关数据中挖掘出有价值信息,辅助人类专家进行精确诊断,成为“人工智能+心理健康”背景下重要研究课题。

相关研究表明,在心理健康方面存在问题的人往往存在不同于正常人群的表现[5]。受此启发,本研究尝试基于心理健康相关数据,利用异常点检测方法挖掘出异常样本,辅助人类专家进行诊断。为提高异常样本识别性能,提出一种新的异常点检测方法(multiple outlier detection models, Multi-ODM),该方法融合多个模型输出作为最终判别结果。

1　相关工作

本章将从心理健康辅助诊断方式和预测模型等 2 个方面对心理健康辅助诊断进行简单介绍。

1.1　心理健康辅助诊断方式

目前,心理健康辅助诊断主要有 3 种方式:量表法、专家诊断以及心理沙盘。

(1)量表法。以症状自评量表测量心理健康是当前较为常用的方式,包括抑郁自评量表、焦虑自评量表、症状自评量表等。文献[6]使用经临床验证的焦虑量表和抑郁量表对博士生和硕士生进行综合调查,发现焦虑和抑郁在具有心理健康问题人员中占有非常高的比例。文献[7]使用一般情况调查表、自编新冠肺炎知识认知行为表、广泛性焦虑量表和抑郁症状群量表探索疫情应激状态下大学生心理健康影响因素,为开展心理健康教育提供依据。文献[8]利用症状自评量表进行研究,发现无明显抑郁症状的学生更擅长社交。文献[9]采用问卷调查方式以 18 岁及以上中国公民为调查对象,发现我国成年公众心理健康素养总体处于中偏低水平。量表法简单易行,成本较低,容易在大范围内开展心理健康问题检测工作。此类方式存在一定主观性和欺骗性,例如,一些人不愿意透露真实心理状况故意填写错误信息,所得结论易受个体主观意识影响[10]。

(2)专家诊断。顾名思义,专家利用线上远程服务或线下面对面方式对具有心理服务需求用户进行诊断。远程心理服务利用互联网、视频会议等现代技术提供心理健康服务,尤其在新冠疫情爆发期间发展迅速并得到广泛应用。文献[11]介绍了在新冠肺炎大流行期间使用远程心理服务的益处,制定提供服务和监督的最佳实践方案。文献[12]指出远程心理服务是心理诊疗的一个创新,在因危机、冲突或自然灾害而无法获得精神卫生保健时,远程心理服务为用户得到及时诊断提供机会。专家诊断虽然精准但耗时耗力,外加专家人数有限,难以在大范围内展开。远程心理服务缺乏现场感、真实感、亲密感,缺乏相关行业应用规范及专业培训体系,不适用于所有心理问题[13]。

(3)心理沙盘。心理沙盘是一种心理健康诊断与治疗辅助工具。文献[14]使用小组心理沙盘游戏有效提高游戏参与者应对压力的能力。文献[15]指出沙盘游戏可以为创伤应激、残疾或有语言障碍的患者开展较为有效的治疗。文献[16]通过

沙盘游戏疗法提高学生心理健康状况。近年来,部分研发人员利用传统实体沙盘理念开发出软件系统(虚拟沙盘)。虚拟沙盘的使用不受时间、地点等因素影响,逐渐成为专家和用户较为青睐的心理健康辅助诊断方式之一。文献[17]发现虚拟沙盘与实体沙盘在初始心理评估方面具有同样效果,在象征、想象等方面,虚拟沙盘效果较好,更强调虚拟心理沙盘与实体沙盘功效。文献[18]设计虚拟沙盘自检系统应用于广泛型焦虑患者前期自我检查,重点在虚拟心理沙盘设计方面。现有文献多关注虚拟心理沙盘功能与应用,鲜有研究对系统后台产生的海量数据进行挖掘分析等工作。

1.2 心理健康状况预测模型

近年来,机器学习、深度学习等逐渐应用于心理健康领域,带来心理健康测评和诊断方法革新。文献[19]使用决策树算法构建精神疾病风险因素预测模型并揭示与精神疾病风险最相关变量。文献[20]使用决策树算法建立强迫症状预测模型并挖掘影响大学生心理健康问题因素。文献[21]利用社交数据构建基于情绪词典决策树模型预测抑郁倾向。文献[22]使用决策树算法建立大学生心理危机预警方法。文献[23]利用大学生心理健康管理系统导出调查表,基于 k 均值算法(k-means)挖掘学生潜在心理信息。文献[24]将心理云数据与心理量表结合使用,根据改良 k 均值算法(k-means ++)对数据集进行细粒度划分并构建心理健康因子预测模型。文献[25]对学生问卷得分、观看心理视频时长、浏览心理方面内容次数等进行研究,运用 k-means 算法发现可能存在心理问题的学生并建立心理健康数据反馈体系模型。文献[26]利用关联分析算法挖掘大学生人格特征、心理健康状况和成绩之间关联规则。文献[27]采用改进关联规则算法对心理因子间关系展开挖掘。文献[28]利用关联规则算法发现心理健康数据隐藏规律与有价值信息。部分研究使用孤立森林算法进行异常行为检测[29]。

很多工作利用机器学习方法开展心理健康数据挖掘、分析,虚拟心理沙盘诊断方式较为新颖,目前鲜有研究结合虚拟心理沙盘数据和机器学习技术开展心理健康问题辅助诊断。

2 数据收集与分析

本研究收集心理健康服务平台——聆心云系统后台产生的用户操作日志数据,整理出心理沙盘数据集,该心理健康服务平台虚拟心理沙盘主体由沙子、对沙具操作以及可使用沙具组成,分别对应对沙子操作(挖沙子、堆沙子、平整地形)、对沙具操作(创建、移动、缩放、旋转、删除、缩放)和所使用沙具(玫瑰、老奶奶等)。

2.1 数据收集

所用心理沙盘数据为系统后台产生的操作日志文件,每个日志中都包含用户所有操作(例如挖沙子、创建沙具、删除沙具、移动沙具等)和操作对象(各类沙具,例如石头、玫瑰花、儿童等)。经匿名化和特征提取,构建心理沙盘数据集。该数据集共包含 5 697 份沙盘样本,经特征提取后得到 1 595 维数据,每维数据都是用户对心理沙盘进行操作。在删除操作数量较少沙盘数据后(操作总数少于 20),共得到 4 960 份心理沙盘样本。对 4 960 份样本进行处理,得到虚拟沙盘系统中定义的不同操作和沙具集合。操作集合包括挖沙子、堆沙子、平整地形、创建沙具、移动沙具、缩放沙具、旋转沙具、删除沙具、调整沙具深度等 9 种操作类型,沙具集合即为样本中所包含沙具。每个沙盘样本数据由一系列操作及对应沙具组成:

$$U_i=\{(h_j,s_j)\mid 1<j\leqslant n,\ h_j\in H,\ s_j\in S\},\quad (1)$$

式中:U_i 表示第 i 个用户生成的沙盘样本;H 和 S 分别表示虚拟沙盘系统中定义的不同操作和沙具集合;该样本由 n 个二元组构成,h_j 表示该用户第 j 个操作,s_j 为对应操作对象(沙具)。例如,某心理沙盘样本数据为:“"[\"创建沙具(弥勒佛)\",\"移动沙具(弥勒佛)\",\"移动沙具(弥勒佛)\",\"移动沙具(弥勒佛)\",\"创建沙具(老奶奶)\",\"移动沙具(老奶奶)\"]"”,表明该用户创建沙盘流程为创建沙具(弥勒佛),对弥勒佛位置进行 3 次调整,又创建第 2 个沙具(老奶奶),并移动沙具(老奶奶)1 次。

2.2 数据分析

本节从 3 个角度对数据进行简单分析,包括操作类型角度、沙具使用情况角度以及沙盘信息量角度。

(1) 操作类型

从操作类型角度对数据进行分析,提取出每个沙盘样本中所包含操作信息,进行分类汇总,统计出不同类型操作使用情况。图 1 给出每一类操作及其对应使用次数。虚拟沙盘数据共涉及到 9 种不同操作类型,包括挖沙子、堆沙子、平整地形、创建沙

具、移动沙具、缩放沙具、旋转沙具、删除沙具以及调整沙具深度等。在 4 960 个沙盘样本中,对沙具操作较多,而对沙子和地形操作较少。移动沙具、挖沙子、创建沙具为使用次数较多的 3 个操作,分别为 143 870、98 560、64 966 次。平整地形和删除沙具为使用次数较少操作,分别为 5 149 和 4 655次。

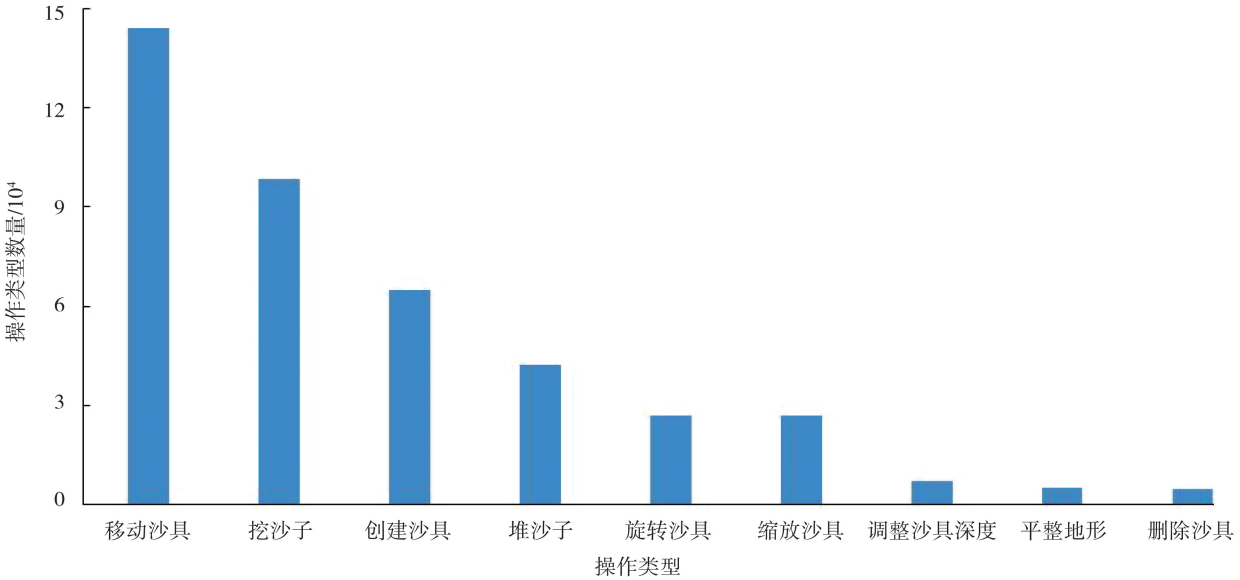

图 1　数据集所包含操作类型及其数量

Fig.1　The types and number of operations contained in the dataset

(2) 沙具使用情况

图 2 展示全体样本中沙具使用情况词云图。较多使用“草丛、椰子树、苹果树、竹、莲花、草坪”等植物类沙具,其次是“篱笆、山、石头、草房”等。心理学相关文献[30]表明,在沙具的使用上,不同人群体现出不同偏好。例如:外向人群多使用鲜花和现实人物,主题场景以社会场景为主;内向人群少使用鲜花和现实人物,主题场景以自然场景为主。沙具使用情况在一定程度上能够反映用户真实心理状态和性格倾向。

图 2　数据集中所使用沙具词云图

Fig.2　Word cloud map of the sand play used in the dataset

(3) 沙盘信息量

表 1 给出操作数量排名在前 100 的沙盘情况。4 960 份心理沙盘样本包含的信息量各不相同,部分沙盘所含操作数量高达上千,有的沙盘包含操作数量只有几百个,在操作数量上存在较大差异。

表 1　不同沙盘样本操作总数(前 100 个)

Table 1　Total number of operations for different sand table samples

操作数量	沙盘样本数量/个
>2 000	2
>1 000 且≤2 000	18
>500 且≤1 000	54
>100 且≤500	26

3　方法

3.1　符号说明

用小写字母表示标量,大写字母表示集合或者模型,黑体小写字母表示向量。为方便理解,表 2 给出所使用符号及其相关含义。

表 2　符号说明

Table 2　Symbol description

符号	含义
i,j	下标或索引变量
D	心理沙盘数据集
$\boldsymbol{x}_i$	数据集 D 中第 i 个样本
m	数据集 D 中样本的数量
t	异常点检测模型的个数
A	异常点检测基分类器算法
θ	$0<\theta\leqslant t$ 为阈值
O	异常沙盘样本集合,$O\subset D$

3.2 异常沙盘检测方法

给定心理沙盘数据集 $D=\{x_1,x_2,\cdots,x_m\}$，其中，m 为样本个数；$\boldsymbol{x}_i\in\mathbf{R}^d$ 为一个 d 维向量，是对第 i 个用户 U_i 的沙盘数据的特征描述。沙盘数据异常点检测任务的目的是找出数据集 D 中明显异于全局分布的异常点样本集合 $O\subset D$。

为进一步缩小异常样本范围并提高异常点检测准确性，提出一种新的异常点检测算法 Multi-ODM，该算法构建 t 个异常点检测模型 $f_1,f_2,\cdots,f_t$，运用这些模型得到异常点集合分别记为 $O_1,O_2,\cdots,O_t$，基于融合 t 个模型的输出得到本算法异常样本集合 O。假设给定样本 $\boldsymbol{x}_i$，若判定 $\boldsymbol{x}_i$ 为异常样本的模型个数超过设定的阈值 $\theta(0<\theta\leqslant t)$，则判定其为异常样本并加入异常样本集合 O。集合 O：

$$O=\left\{\boldsymbol{x}_i\mid\left(\sum_{j=1}^{t}I(\boldsymbol{x}_i\in O_t)\right)\geqslant\theta\right\},\quad \text{s.t. } \boldsymbol{x}_i\in D,\ 0<\theta\leqslant t \tag{2}$$

式中：$I(\boldsymbol{x}_i\in O_t)$ 为指示函数，如果条件满足，取值为1，否则为0；阈值 θ 为不大于 t 的正整数，其作用是调节异常点集合 O 中样本置信度，θ 越大，则异常样本置信度就越高。

算法 1 Multi-ODM 算法

输入 心理沙盘数据集 $D=\{x_1,x_2,\cdots,x_m\}$、异常点基分类器算法 $A_i(i=1,2,\cdots,t)$、集成阈值$\theta(0<\theta\leqslant t)$

输出 异常点样本集合 O

（1）心理沙盘数据集 D 输入到异常点基分类器算法 A_i 中；

（2）在 D 上运行算法 A_i 得到模型 $f_i(i=1,2,\cdots,t)$；

（3）得到各模型 f_i 异常点集合 $O_i(i=1,2,\cdots,t)$；

（4）根据公式(2)计算得到最终预测结果；

（5）输出异常点样本集合 O。

4 试验

本章主要对试验进行简单介绍与分析，包括试验设置、评价指标、试验结果与分析、超参数对试验结果影响，对所检测出的异常样本进行数据分析及可视化。

4.1 试验设置

引入集成学习到异常点检测任务上，该学习范式通过综合考虑多个基分类器预测结果以达到提高模型性能的目的。基分类器性能很大程度上影响集成学习模型性能。本研究所用基分类器选用目前在异常点检测任务上较为流行的 4 种算法，包括孤立森林算法（isolation forest，iForest）[31]、局部异常因子算法（local outlier factor，LOF）[32]、k 近邻算法（k-nearest neighbor，kNN）[33] 以及 k-means 算法[34]。在参数设置方面，iForest 算法建立 100 棵树，最大特征数为 9，kNN 算法、LOF 算法均采用默认参数，k-means 算法中 k 为 2。

不同于传统多数投票方法，在预测结果集成阶段将阈值 θ 设为 4，即对于一个沙盘样本，只有在全部学习器都判断其为异常样本情况下，最终判断为异常样本。

4.2 评价指标

异常点是指数据中不符合一个定义明确正常行为概念的模式，其观测值往往偏离其他数据点[31]。利用沙具使用情况相似度（similarity of sand tool usage，S_{stu}）、均值向量距离（distance of mean vector，d_{mv}）和聚类性能评价指标（davies-bouldinindex，D_{bi}）等 3 个评价指标来衡量模型性能。

（1）沙具使用情况相似度 S_{stu}

如 2.1 节所述，令 $U_i=\{(h_j,s_j)\mid 1<j\leqslant n_i,\ h_j\in H,\ s_j\in S\}$，表示第 i 个用户的沙盘数据，其中：H 和 S 分别表示虚拟沙盘系统中定义的不同操作和沙具集合；h_j 表示该用户第 j 个操作，s_j 为对应操作对象（沙具），n_i 表示第 i 个用户的操作数量。给定沙盘数据集 $D=\{x_1,x_2,\cdots,x_m\}$，其中$\boldsymbol{x}_i\in\mathbf{R}^d$ 为一个 d 维向量，是对第 i 个用户 U_i 沙盘数据的特征描述。则全体沙盘数据用户集合为 $U_{\text{all}}=\{U_i\mid\boldsymbol{x}_i\in D\}$，异常点数据对应用户集合 $U_{\text{ano}}=\{U_j\mid\boldsymbol{x}_j\in O\}$，其中，$O$ 为算法输出的异常点集合。

沙具集合 $T=\{\tilde{s}_1,\tilde{s}_2,\cdots,\tilde{s}_t\}$，即共有 t 个不同的沙具，则全体沙盘数据用户所使用沙具集合 T_D 和异常点用户使用沙具集合 T_O 可分别通过计算公式(3)(4)得到。

$$T_D=\{\tilde{s}_k\mid\exists i,j,\ (h_i,s_i)\in U_j \text{ and } U_j\in U_{\text{all}}\}, \tag{3}$$

$$T_O=\{\tilde{s}_k\mid\exists i,j,\ (h_i,s_i)\in U_j \text{ and } U_j\in U_{\text{ano}}\}。 \tag{4}$$

对 T_D 和 T_O 中的沙具根据每一个沙具使用次数进行排序，令 Max_rank(T_D,r) 和 Max_rank(T_O,

r)分别表示两个集合中排名前 r 的沙具集合,则沙具使用情况相似度 S_{stu} 指标可使用集合的交并比运算得到:

$$S_{\text{stu}}=\frac{\text{Max\_rank}(T_D,r)\cap\text{Max\_rank}(T_O,r)}{\text{Max\_rank}(T_D,r)\cup\text{Max\_rank}(T_O,r)}。\tag{5}$$

沙具使用情况相似度 S_{stu} 衡量在沙具使用频率上,异常点沙盘与全体沙盘样本之间差异,该值越大,说明异常点检测算法得出样本与全局样本在沙具使用上分布相差越小,性能也就越差。心理学相关研究表明内向型和外向型人在沙具的使用上存在明显差异[28],例如:外向型人格较多使用鲜花、树木、真实场景中的人物,较少使用篱笆、石头等;内向型人格则呈现出相反的情况。这进一步说明通过沙具使用情况相似度指标对异常点算法性能进行度量具有一定客观性和合理性。

(2) 均值向量距离 d_{mv}

给定沙盘数据集 $D=\{x_1,x_2,\cdots,x_m\}$ 及异常点集 $O\subset D$, 均值向量相似度 d_{mv} 定义为:

$$d_{\text{mv}}=\text{dist}(\boldsymbol{u}_D,\boldsymbol{u}_O),\tag{6}$$

式中,dist(·)为距离度量函数,可以是欧氏距离、余弦距离等;$\boldsymbol{u}_D$,$\boldsymbol{u}_O$ 分别为数据集 D 和 O 的均值向量:

$$\boldsymbol{u}_D=\frac{1}{|D|}\sum_{i=1}^{|D|}\boldsymbol{x}_i,\tag{7}$$

$$\boldsymbol{u}_O=\frac{1}{|O|}\sum_{j=1}^{|O|}\boldsymbol{x}_j。\tag{8}$$

均值向量距离 d_{mv} 越小,表明异常点样本与全局样本越相似,算法性能也就越差;反之,d_{mv} 越大,异常点样本平均分布就越远离全局样本分布,算法性能也就越好。所使用距离为欧氏距离。

(3) 聚类性能评价指标 D_{bi}

聚类性能评价指标是一种用于评估聚类算法质量指标,该指标综合考虑类内样本相似度以及类间样本差异度[35]。该指标越小,表示类内样本距离越小,类内相似度越高,类间样本距离越大,类间样本相似度越低。

给定沙盘数据集 $D=\{x_1,x_2,\cdots,x_m\}$,异常点集 $O\subset D$,$D/O\subset D$。聚类性能评价指标 D_{bi} 定义为:

$$D_{\text{bi}}=\frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}p_{ij},\tag{9}$$

式中:k 为聚类后簇的个数; p_{ij} 为簇 i 与簇 j 之间的相似度,可由公式(10)得到。v_i 表示第 i 个簇中所有样本点到簇中心距离的平均值,v_j 表示第 j 个簇中所有样本点到簇中心距离的平均值,d_{ij} 表示第 i 个簇与第 j 个簇之间的距离(即两个簇中心之间的距离)。v_i、v_j、d_{ij} 均可利用公式(8)、(11)得到,其中所使用距离均为欧氏距离。

$$p_{ij}=\frac{v_i+v_j}{d_{ij}},\tag{10}$$

$$\boldsymbol{u}_{D/O}=\frac{1}{|D/O|}\sum_{j=1}^{|D/O|}\boldsymbol{x}_j。\tag{11}$$

D_{bi} 越小,表明聚类质量越高,即异常点样本簇内部紧密,与其他样本分离较远,算法性能越好;反之,D_{bi} 越大,异常点样本分布就越接近其他样本分布,算法性能越差。

4.3　试验结果与分析

4.3.1　对比试验

将 Multi-ODM 算法与 5 种传统异常点检测算法进行对比,包括:孤立森林 iForest[30]、局部异常因子算法 LOF[32]、kNN 算法[33]、k-means 算法[34],以及基于并行集成的异常检测算法(locally selective combination in parallel outlier ensembles, LSCP)[36]。其中,LSCP 算法使用 4 个不同参数的 LOF 算法进行并行,邻居数分别为 15、20、25 和 35。

(1) 孤立森林 iForest 算法[31]。该算法假定异常点往往偏离整体样本分布,体现在决策树算法上,异常样本点往往需要较少判断便可以得到决策结果,即异常点样本到根节点路径长度往往较短。孤立森林算法生成若干个完全随机树(训练集、划分属性、切分点均为随机获取),通过计算样本平均路径长度定义样本异常分数,异常分数高的样本组成异常点集合。

(2) 局部异常因子 LOF 算法[32]。局部异常因子 LOF 算法是一种基于密度的离群点检测方法。基于密度离群点检测方法有一个基本假设:非离群点对象周围密度与其邻域周围的密度类似,离群点对象周围密度显著不同于其邻域周围密度。通过给每个数据点分配一个依赖于邻域密度离群因子来判断该数据点是否为离群点。给定某个样本,该算法通过计算该样本邻域内密度,得到与密度成反比局部异常因子,该值越小,说明该样本越有可能为密集点;反之,该值越大(大于 1),则该样本越可能是异常点。

(3) k 近邻算法[33]。kNN 算法在对样本进行异常点检测时,依次计算每个样本点与它最近 k 个样本的平均距离,利用该距离与事先设定的阈值进行比较,如果大于阈值,则认为是异常点。

(4) k 均值算法(k-means)[34]。该算法初始化

k 个类中心并计算各个数据对象到聚类中心的距离,把数据对象划分至距离其最近的聚类中心所在类簇中;根据所得类簇,更新类簇中心,一直迭代,迭代终止时得到最终聚类结果。在异常点检测任务中,往往事先假定样本有两类(异常样本\正常样本),即 k 设为 2,执行聚类任务,样本数量较少簇认为是异常点样本。

(5) LSCP 算法[36]。该算法训练 k 个基础异常检测器并得到所有基学习器的输出结果,输出结果均值或最大值作为伪标签;对每个测试点生成局部空间,即近邻。在生成的局部空间中,对所有基学习器产生的伪标签进行评估,距离大的则认为是异常样本。

4.3.2 试验结果与分析

不同方法之间性能对比如表 3 所示,用粗体表示在某个指标上的最优结果,加下划线表示次优结果,从表 3 可以看出,相比传统异常点检测方法,Multi-ODM 算法在 3 项指标上均取得较好性能,S_{stu} 比次优方法(k-means)降低 2.07%,d_{mv} 与次优方法相比,d_{mv} 高出 3.75%,在 D_{bi} 指标上,与最优方法(k-means)达到相近的效果。这意味着该方法检测出异常样本与全局样本存在更大差距且聚类效果更好,异常样本可信程度更高。

表 3 不同方法的性能比较

Table 3 Performance comparison of different methods

方法	S_{stu}	d_{mv}	D_{bi}
iForest[31]	0.754 4	1.742 0	0.896 6
LOF[32]	0.694 9	1.070 6	1.851 2
kNN[33]	0.754 4	1.499 3	0.836 3
k-means[34]	0.449 3	1.276 3	**0.310 4**
LSCP[36]	0.754 4	1.089 2	1.770 2
Multi-ODM	**0.428 6**	**1.779 5**	0.317 7

4.4 超参数 θ 的影响

试验对比异常点检测模型阈值 θ 对试验结果的影响。图 3 列出不同阈值 θ 的性能对比,阈值 $\theta=4$ 时,孤立点数量(个)、S_{stu} 和 D_{bi} 均取得最优。

4.5 异常样本数据分析及可视化

从检测出异常沙盘样本中随机选出 3 个样本,对其所包含操作类型进行可视化。如图 4 所示。由图 4 所知,这 3 个异常沙盘样本操作数量最多操作类型是挖沙子、堆沙子,很少关注沙具。

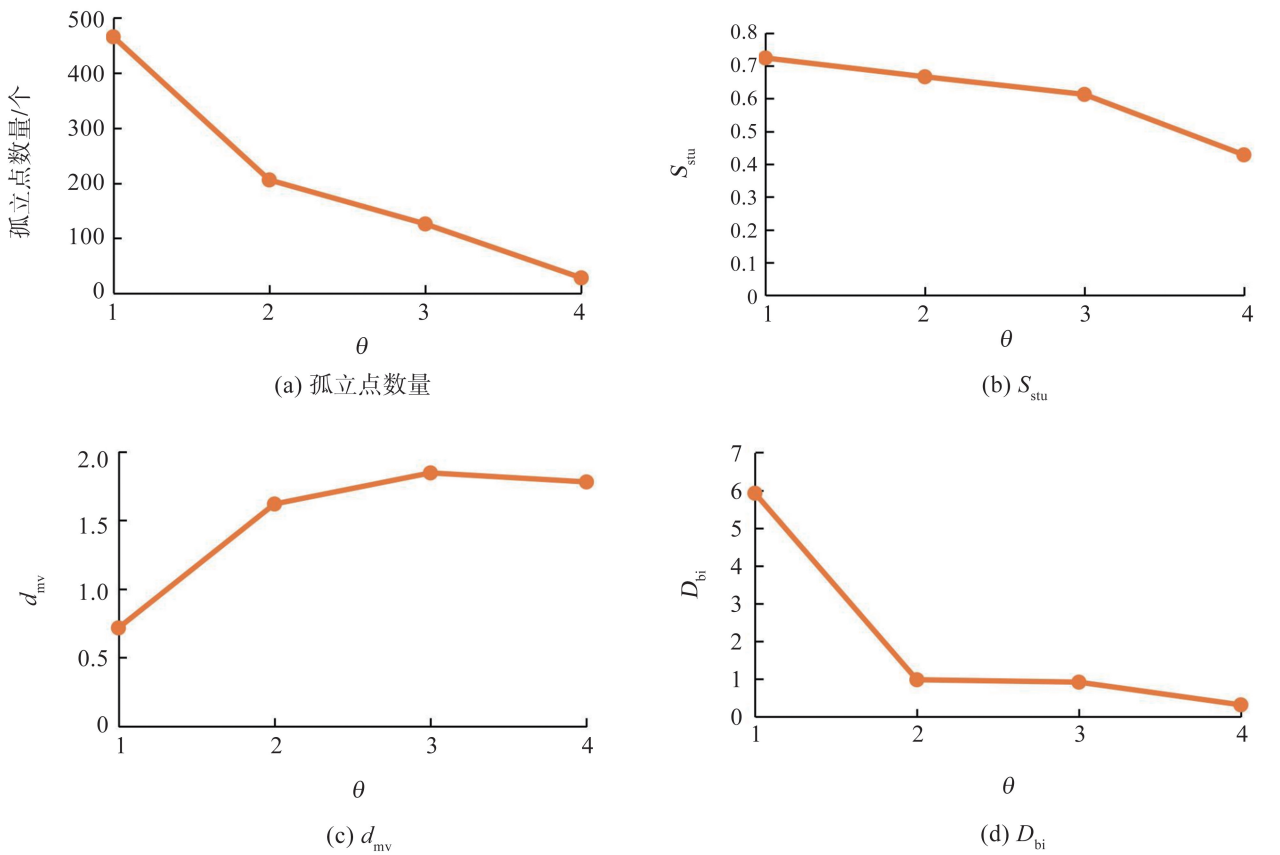

图 3 阈值 θ 对算法性能影响

Fig.3 The impact on algorithm performance of threshold θ

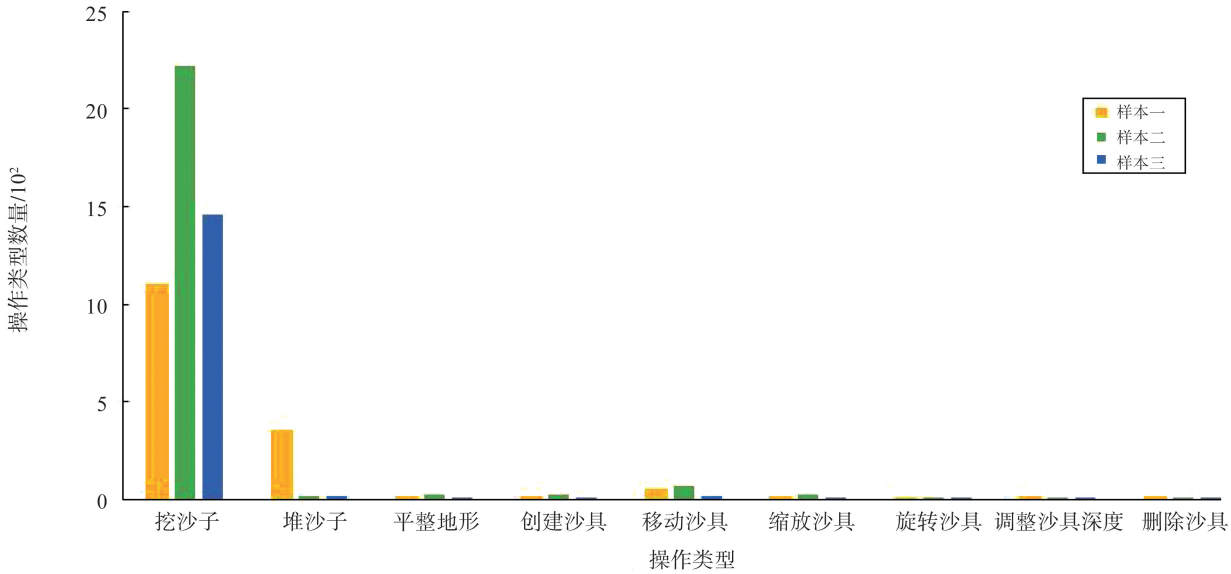

图 4　部分异常样本数据可视化

Fig.4　Partial abnormal sample data visualization

5　结论

本研究基于虚拟心理沙盘数据,提出利用异常点检测进行心理健康辅助诊断方法。针对心理沙盘数据,相比传统异常点检测方法,所提方法得到异常样本的可信程度更高,检测范围更小,进一步提高人类专家诊断效率,为虚拟心理沙盘技术在大规模心理筛查应用场景中发挥作用提供技术支撑。

遗憾的是,由于数据隐私问题,每个沙盘样本的真实标记未知(即:是否存在心理健康问题),导致文中对算法性能只能直观上粗略估计和评价。在后续研究中,计划邀请人类专家对检测出的异常样本进行标注,获得高质量数据样本,结合半监督学习或小样本学习提高异常样本识别的查准率和查全率,促进机器学习、大数据技术在心理健康辅助诊断中落地应用。

参考文献:

[1] World Health Organization. Depression and other common mental disorders: global health estimates[R]. Geneva, Switzerland: World Health Organi-zation, 2017.

[2] 靳宇倡, 张政, 郑佩璇, 等. 远程心理健康服务:应用、优势及挑战[J]. 心理科学进展, 2022, 30(1):141-156.
JIN Yuchang, ZHANG Zheng, ZHENG Peixuan, et al. Tele-mental health services: applications, benefits and challenges[J]. Advances in Psychological Science, 2022, 30(1):141-156.

[3] 刘媛媛. 基于 SCL-90 量表的中国人群心理健康现状及 30 年变化特征分析[D]. 西安: 中国人民解放军空军军医大学, 2018.
LIU Yuanyuan. Analysis of the current situation and 30-year change characteristics of the mental health of the Chinese population based on the SCL-90 scale[D]. Xi′an: Chinese People′s Liberation Army Air Force Military Medical University, 2018.

[4] 林仲贤, 丁锦红. 心理测验的含义及其应用[J]. 中国临床康复, 2004(3):522-523.
LIN Zhongxian, DING Jinhong. The meaning and application of mental test[J]. Clinical Rehabilitation in China, 2004(3):522-523.

[5] 张雯, 张日昇, 姜智玲. 强迫症状大学生的箱庭作品特征研究[J]. 中国临床心理学杂志, 2011, 19(4):553-557.
ZHANG Wen, ZHANG Risheng, JIANG Zhiling. A study on the characteristics of obsessions in college students with obsessive symptoms[J]. Chinese Journal of Clinical Psychology, 2011, 19(4):553-557.

[6] EVANS T M, BIRA L, GASTELUM J B, et al. Evidence for a mental health crisis in graduate education[J]. Nature Biotechnology, 2018, 36(3): 282-284.

[7] 昌敬惠, 袁愈新, 王冬. 新型冠状病毒肺炎疫情下大学生心理健康状况及影响因素分析[J]. 南方医科大学学报, 2020, 40(2):171-176.
CHANG Jinghui, YUAN Yuxin, WANG Dong. Analysis of mental health status and influencing factors of college students under novel coronavirus pneumonia epidemic[J]. Journal of Southern Medical University, 2020, 40(2): 171-176.

[8] 聂敏. 高校学生行为分析及应用研究[D]. 成都: 电子

科技大学，2020.
NIE Min. College student behavior analysis and application research [D]. Chengdu: University of Electronic Science and Technology, 2020.

[9] 江光荣，李丹阳，任志洪，等. 中国国民心理健康素养的现状与特点[J]. 心理学报，2021，53(2):182-201.
JIANG Guangrong, LI Danyang, REN Zhihong, et al. The current situation and characteristics of Chinese national mental health literacy [J]. Acta Psychologica, 2021, 53(2):182-201.

[10] NAYAK M, NARAYAN K A. Strengths and weaknesses of online surveys[J]. Technology, 2019, 6(7): 0837-2405053138.

[11] HAMES J L, BELL D J, PEREZ-LIMA L M, et al. Navigating uncharted waters: considerations for training clinics in the rapid transition to telepsychology and telesupervision during COVID-19 [J]. Journal of Psychotherapy Integration, 2020, 30(2): 348.

[12] INCHAUSTI F, MACBETH A, HASSON-OHAYON I, et al. Telepsychotherapy in the age of COVID-19: a commentary[J]. 2020, 30(2):394-405.

[13] 陈红，汪卫华，袁水平，等. 远程心理咨询与面对面咨询的对比研究[J]. 精神医学杂志，2010，23(2): 128-129.
CHEN Hong, WANG Weihua, YUAN Shuiping, et al. A comparative study of remote psychological counseling and face-to-face counseling[J]. Military Medical Journal, 2010, 23(2):128-129.

[14] SUN P, QU Y X, WU J, et al. Improving Chinese teachers' stress coping ability through group sandplay [J]. The Spanish Journal of Psychology, 2018: 65-72.

[15] ROESLER C. Sandplay therapy: an overview of theory, applications and evidence base[J]. The Arts in Psychotherapy, 2019, 64: 84-94.

[16] GUO J, LI D. Effects of image-sandplay therapy on the mental health and subjective well-being of children with autism[J]. Iranian Journal of Public Health, 2021, 50 (10): 2046-2054.

[17] 张伯全，乔冬冬，王汝展，等. VR 沙盘与实体沙盘用于大学新生初始心理测查自身对照研究[J]. 精神医学杂志，2018，31(5):359-362.
ZHANG Boquan, QIAO Dongdong, WANG Ruzhan, et al. Self-contrast study of VR sand table and physical sand table used in the initial psychological test of college freshmen [J]. Journal of Psychiatry, 2018, 31(5): 359-362.

[18] 韦玉. 面向广泛型焦虑群体的虚拟沙盘自检系统设计研究[D]. 广州：华南理工大学，2021.
WEI Yu. Research on the design of virtual sand table self-inspection system for general anxiety groups [D]. Guangzhou: South China University of Technology, 2021.

[19] DOOSHIMA M P, CHIDOZIE E N, ADEMOLA B J, et al. A predictive model for the risk of mental illness in Nigeria using data mining [J]. International Journal of Immunology, 2018, 6(1): 5-16.

[20] 孙伟平. 决策树技术在大学生心理健康测评中的应用研究[D]. 郑州：郑州大学，2020.
SUN Weiping. Research on the application of decision tree technology in the evaluation of mental health of college students [D]. Zhengzhou: Zhengzhou University, 2020.

[21] 林靖怡，黎大坤，吴平鑫，等. 基于社交数据挖掘的心理健康预警建模与分析[J]. 电子技术与软件工程，2020(8):172-173.
LIN Jingyi, LI Dakun, WU Pingxin, et al, Modeling and analysis of mental health early warning based on social data mining[J]. Electronic Technology and Software Engineering, 2020(8):172-173.

[22] 赵丹. 基于决策树的大学生心理危机预警模型研究及应用[D]. 北京：北京林业大学，2020.
ZHAO Dan. Research and application of college students' psychological crisis early warning model based on decision tree[D]. Beijing: Beijing Forestry University, 2020.

[23] 吴婷. 基于 k-means 聚类算法的大学生心理管理系统研究[D]. 武汉：湖北工业大学，2017.
WU Ting. Research on psychological management system of college students based on k-means clustering algorithm[D]. Wuhan: Hubei University of Technology, 2017.

[24] 王震震. 心理云大数据平台中用户心理画像的研究与应用[D]. 北京：北京邮电大学，2021.
WANG Zhenzhen. Research and application of user psychological portrait in psychological cloud big data platform[D]. Beijing: Beijing University of Posts and Telecommunications, 2021.

[25] 赵向兵，白栋. 基于 Python 的学生健康数据聚类分析系统[J]. 电子技术与软件工程，2021(14):183-185.
ZHAO Xiangbing, BAI Dong. Student health data cluster analysis system based on Python[J]. Electronic Technology and Software Engineering, 2021 (14): 183-185.

[26] 梁娟，罗海据. 大数据挖掘方法在大学生心理预警系统中的应用[J]. 中国学校卫生，2018，39(12): 1821-1824.
LIANG Juan, LUO Haiju. Application of big data

mining method in college students′ psychological early warning system[J]. Chinese School Hygiene, 2018, 39(12):1821-1824.

[27] 侯震. 基于数据挖掘的高校学生心理测评与辅导系统的设计与实现[D]. 西安: 西安电子科技大学, 2020.
HOU Zhen. Design and implementation of college students′ psychological assessment and counseling system based on data mining [D]. Xi′an: Xidian University, 2020.

[28] 刘红红.基于数据挖掘的心理疾病预警分析技术研究[J].电子设计工程, 2021, 29(15):31-35.
LIU Honghong. Research on early warning analysis technology of mental illness based on data mining[J]. Electronic Design Engineering, 2021, 29(15):31-35.

[29] 祝彦森. 基于改进 iForest 的学生异常行为检测及分析系统研究[D]. 南京: 南京信息工程大学, 2019.
ZHU Yansen. Research on abnormal behavior detection and analysis system of students based on improved iforest [D]. Nanjing: Nanjing University of Information Science and Technology, 2019.

[30] 林少武, 冯春苗, 梁茵, 等. 初始沙盘特征在沙盘心理评估中的应用与发展[J]. 中国健康心理学杂志, 2019, 27(5):788-792.
LIN Shaowu, FENG Chunmiao, LIANG Yin, et al. Application and development of initial sand table characteristics in psychological evaluation of sand table [J]. Chinese Journal of Health Psychology, 2019, 27(5):788-792.

[31] LIU F T, TING K M, ZHOU Z H. Isolation forest [C]//2008 Eighth IEEE International Conference on Data Mining. Pisa, Italy: IEEE Xplore, 2008: 413-422.

[32] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: Identifying density-based local outliers[C]// Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. New York, USA: Association for Computing Machinery, 2000: 93-104.

[33] WANG X, WANG X L, MA Y, et al. A fast MST-inspired kNN-based outlier detection method[J]. Information Systems, 2015, 48: 89-112.

[34] DUAN L, XU L, LIU Y, et al. Cluster-based outlier detection[J]. Annals of Operations Research, 2009, 168: 151-168.

[35] DAVIES D L, BOULDIN D W. A cluster separation measure[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979(2): 224-227.

[36] ZHAO Y, NASRULLAH Z, HRYNIEWICKI M K, et al. LSCP: Locally selective combination in parallel outlier ensembles[C]//Proceedings of the 2019 SIAM International Conference on Data Mining. Alberta, Canda: SIAM International Conference on Data-BMining, 2019: 585-593.

(编辑:陈燕)