

无监督学习基础上的抑郁症分级模型研究^{*}

彭乔立 李汉民 幺冬爱 肖 辉

(武汉大学中南医院信息中心 武汉 430071)

〔摘要〕 采用无监督学习算法对 105 例抑郁症患者的多个实验室检验项目进行分组实验, 将分组模型应用在 88 例测试数据集上, 使用统计学方法检验分组结果, 最终结果表明大部分组间差异有统计学意义, 精神类疾病会对实验室检验项目产生一定影响。

〔关键词〕 无监督学习; 抑郁症; 数据挖掘

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2022.01.007

Study on the Depression Classification Model Based on Unsupervised Learning PENG Qiaoli, LI Hanmin, YAO Dongai, XIAO Hui, Information Center, Zhongnan Hospital of Wuhan University, Wuhan 430071, China

〔Abstract〕 Unsupervised learning algorithm is used to group the laboratory test items of 105 patients with depression, the grouping model is applied to 88 test data sets and statistical methods are used to verify the grouping results. The final results show that the differences between most groups are statistically significant, and mental diseases have a certain impact on laboratory test items.

〔Keywords〕 unsupervised learning; depression; data mining

1 引言

抑郁症是一种严重危害公众健康的慢性非传染性精神类疾病, 其发病原因复杂, 临床表现差异较大, 治疗成本较高, 具有难预防、难诊断、难治愈的特点。抑郁症患者不仅将面临健康损失, 同时社会性歧视可能导致抑郁症患者身心健康与个人权益遭受严重伤害^[1]。如今随着生活节奏加快与压力增长, 抑郁症发病率逐年攀升^[2]。

〔修回日期〕 2021-05-26

〔作者简介〕 彭乔立, 硕士, 中级工程师, 发表论文 5 篇; 通讯作者: 肖辉, 高级工程师。

〔基金项目〕 湖北省卫健委开放基金项目“湖北省健康医疗大数据中心与政务信息数据交换标准研究”(项目编号: ZWXX2021-01)。

2 相关概念

2.1 抑郁症分级

据 WHO 统计全球每年因为医疗资源浪费而造成损失达千亿美元。医疗资源浪费往往是由病情评估不合理导致, 由于慢性非传染性疾病治疗成本过高, 病情分级成为慢病诊疗中的重要一环, 根据定量数据划分病情等级为常用方法, 可作为重要诊疗依据, 指导医疗资源配置以提高诊疗效率。抑郁症属于精神类疾病, 由于缺乏器质性病变的临床证据, 无法通过实验室数据或其他检查数据来进行量化病情分级, 目前该疾病主要通过抑郁症自评量表(Self-rating Depression Scale, SDS)进行量化^[3], 该表由大量主观问答组成, 使用时可能产生较高主观倚倚风险。相较于量化病情分级方法, 基于主观问答的抑郁症病情分级方法精确性与灵敏度较低。

2.2 无监督学习

2.2.1 定义 无监督学习是指在缺乏先验知识的场景下,对难以区分类别、定义性质的数据进行标注、分组与特征化等处理。在数据挖掘过程中,数据集缺乏标注、分组信息的情况时有发生,可能导致人工智能、深度学习等许多智能算法无法实施。因此数据标注、分组工作是许多数据挖掘算法的前提。对于缺乏先验知识的数据而言,人工标注是最可靠的方法,但需要耗费大量人力与时间成本,甚至经常超过训练模型所需要时间^[4]。无监督学习算法可以挖掘数据内在特征并根据特征值进行相关运算,实现数据维度转换的功能。

2.2.2 分类 根据应用方式不同主要分为聚类 (Cluster)、降维 (Reduction)、人工神经网络 (Artificial Neural Network, ANN) 等技术。聚类分析可以挖掘并分析数据特征,计算聚类中心,根据特征

值不同聚类分析可分为以下几类:均值聚类 (Means)、层次聚类 (Hierarchy)、概率聚类 (Probability) 等^[5];数据降维通过分析数据维度共同特征,将其合并成更少、更精简的新维度空间,根据合并依据的不同降维分析可以分为:主成分分析 (Principal Component Analysis, PCA),因子分析 (Factor Analysis, FA) 以及线性判别 (Linear Discriminant Analysis, LDA) 等;人工神经网络是当前最热门的数据分类算法,该算法通过构建输入层-隐藏层-激活函数-输出层的类生物神经系统结构,实现数据分类功能^[6]。本研究使用聚类分析中的 K 均值 (K-Means) 算法,以抑郁症患者实验室数据为对象,对其进行定量分析以获取分级模型,再用该分级模型对测试数据集进行分组,最后使用统计学方法进行评价,验证分级模型有效性,见图 1。

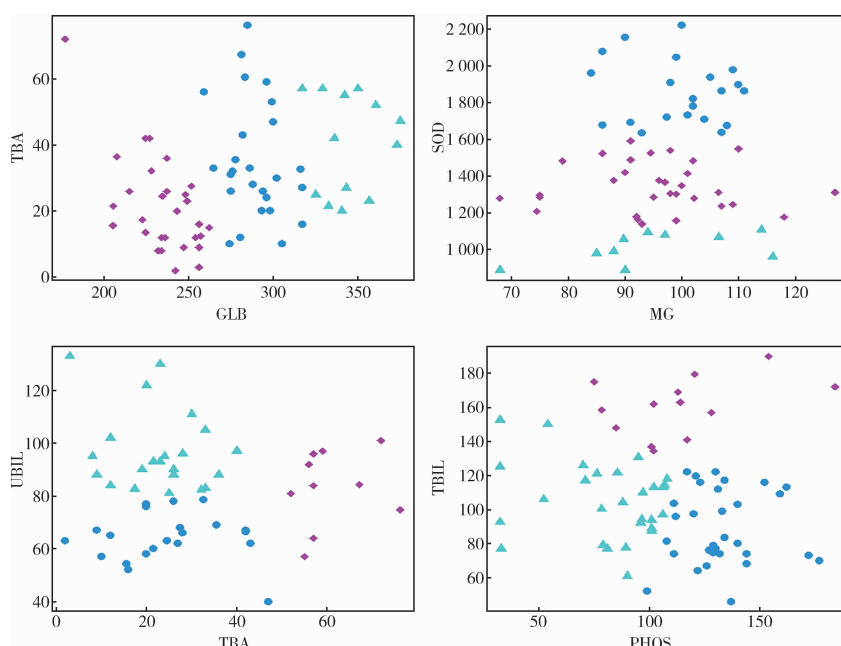


图 1 分组结果散点 (应用于部分维度)

3 方法

3.1 数据集准备

使用健康信息系统 (Health Information System, HIS) 查询符合要求的患者数据,以主诊断编码类

目为“F32”“F33”(抑郁症对应主诊断编码)为选入标准获取患者索引号,以该索引关联检验信息系统 (Laboratory Information System, LIS) 以获取具体检验项目数据,对所得数据进行脱敏处理,去除姓名、身份证号、住址等患者隐私信息,以患者主索引作为主键与单个样本数据进行关联,对数据

集中的 NA 值采用平均化处理方案，包含 0 值或 NA 值多于半数的列视为无效并对其进行剔除。完成数据脱敏与完整性处理后，数据集最终包含 193 个样本行，其中 105 行样本作为训练集，另外 88 行作为测试集。

3.2 使用无监督学习算法获取分级模型

3.2.1 概述 本研究将探讨抑郁症量化分级方案，采用距离聚类中的 K - Means 算法，使用该算法训练数据，将其分为 3 组并获取聚类中心，尝试使用聚类中心为测试数据集进行分组，最后使用统计学方法验证分组结果效能。

3.2.2 去除异常值 数据集中零散分布的异常值会影响数据分布特征并使聚类结果产生偏倚，因此在计算聚类中心前需要去除异常值。利用百分位数（Percentile）过滤异常值，只保留第 1（Q1）和第 3（Q3）分位数之间的数据，此范围外的数据将被视为异常值而不纳入计算。

3.2.3 计算聚类中心 运用 K - Means 聚类算法

获取聚类中心，具体算法如下：根据数据分布情况随机选取 3 个初始中心点；分别计算每个样本到初始中心点的欧式距离；计算误差值，根据误差重新计算聚类中心；重复上两个步骤，直到达到迭代数上限或者误差小到满足条件时停止，得到分级模型。K - Means 算法公式如下：

$$\text{label}_i = \min \parallel x_i - \frac{1}{|c_j|} \sum_{t \in c_j} x_t \parallel$$

3.3 统计学评价

将分级模型应用在测试数据集上并分为 3 组，分别计算各组统计数据，针对数据集分布特性使用不同统计学算法验证分组结果，使用 Shapiro 方法验证正态性，采用 Barlett 方法验证方差齐性，对于满足正态性与方差齐性的数据列，使用 ANOVA 方法分析组内与组外差异，对于组内差异使用 LSD - T 方法进行两两比较，对不满足正态性或方差齐性的数据列使用 Kruskal 方法两两比较，然后通过 Wilcoxon - Mann - Whitney 方法分析其组内差异，见表 1。

表 1 训练数据分组模型

项目名称	TBA	GLB	SOD	MG	UBIL	TBIL	PHOS
组 1	36. 38	280. 63	1 169. 23	96. 41	59. 54	74. 21	171. 58
组 2	57. 92	261. 06	1 128. 18	78. 26	105. 29	148. 87	102. 72
组 3	46. 88	289. 84	1 457. 46	101. 46	95. 77	119. 26	105. 38

4 结果

根据分组结果散点可以看出 3 个分组之间分界清晰、形态稳定，聚类中心互相独立。由训练数据分组模型结果可以发现训练数据所计算得到的分组模型梯度明显，呈一定递增关系，用该模型对测试数据进行分组并用统计学方法校验分组。结果显示

大部分检验项目的组内差异具有统计学意义（ $P < 0. 05$ ， $P = 0. 001$ 为值过小故用该值替代），随后两两比较组间差异，结果表明有部分检验项目 3 组之间差异都有显著性（ P_1 、 P_2 、 P_3 分别为两两比较的 P 值），部分检验项目的组间差异不全具有统计学差异，存在少部分项目的组间差异没有统计学意义，见表 2。

表 2 测试集分组结果统计校验

项目名	均值	组 1 均值	组 2 均值	组 3 均值	P 值	P_1 值	P_2 值	P_3 值
APTT	311. 51	282. 75	294. 35	305. 54	0. 042	0. 131	0. 084	0. 016
INR	107. 34	103. 12	106. 75	104. 77	0. 606	0. 841	0. 703	0. 320
AG	133. 82	111. 87	125. 41	141. 44	0. 002	0. 001	0. 001	0. 008

续表 2

ALB	360.86	312.25	356.35	374.80	0.017	0.025	0.003	0.108
ALP	89.46	76.87	84.71	72.66	0.070	0.081	0.059	0.013
CO2	256.92	229.62	251.67	263.37	0.002	0.001	0.001	0.029
CA	219.06	219.12	217.25	221.70	0.252	0.303	0.286	0.144
CL	1 043.02	1 033.62	1 049.08	1 044.80	0.952	0.997	0.981	0.578
GLB	275.04	280.65	287.81	263.55	0.026	0.084	0.036	0.011
K	397.42	394.33	406.16	385.00	0.039	0.124	0.042	0.013
UBIL	86.28	60.66	75.46	84.84	0.018	0.011	0.007	0.093
DD	562.54	499.83	506.79	247.39	0.015	0.146	0.012	0.004
FIB	332.91	265.37	319.96	273.84	0.037	0.004	0.517	0.015
ALT	24.17	17.50	18.94	19.79	0.885	0.845	0.921	0.636
AST	27.73	15.33	23.08	22.91	0.406	0.064	0.573	0.746
ClQ	1 897.04	1 813.37	2 002.81	1 226.53	0.001	0.003	0.001	0.001
CYSC	140.82	83.11	98.14	60.28	0.001	0.124	0.002	0.001
DBIL	32.48	14.16	20.30	27.35	0.006	0.002	0.086	0.006
MG	97.12	97.50	97.40	77.61	0.001	0.981	0.001	0.001
PHOS	112.13	150.00	105.38	105.58	0.064	0.011	0.025	0.604
SA	6 238.07	6 079.12	6 513.04	4 383.72	0.001	0.457	0.001	0.001
SOD	1 472.09	1 173.50	1 461.62	1 126.13	0.001	0.027	0.006	0.001
TBA	51.63	27.50	31.30	28.82	0.976	0.994	0.982	0.870
TBIL	637.94	73.12	105.87	116.92	0.010	0.013	0.002	0.080

5 结语

本研究使用无监督学习算法对无先验知识的抑郁症患者实验室数据进行分级，通过统计学算法对分级进行验证，结果表明大部分指标的组间差异具有统计学意义，说明精神类疾病会对实验室检验项目产生一定影响。但本研究尚未解决分组数量如何确定的问题，未来将会结合已有研究结果继续探索分组数量与分组模型质量之间的关系。

参考文献

1 张钰群, 袁勇贵. 2020 年抑郁症研究进展回顾 [J]. 中华医学信息导报, 2021, 36 (4): 11-12.

2 杜婷婷, 陈颖川, 朱冠宇, 等. 基于大脑形态学特征的机器学习方法预测帕金森病患者的抑郁症状 [J]. 临床神经外科杂志, 2021, 18 (2): 121-124, 130.

3 中华医学会. 多角度深入研究揭示抑郁障碍诊疗关键 [EB/OL]. [2020-07-02]. https://www.cma.org.cn/art/2020/7/2/art_17_34835.html.

4 陈默, 蔡苗, 黄阿红, 等. 基于 K-means 聚类与支持向量机的大病患者住院费用影响因素与控制策略研究 [J]. 中国医院管理, 2019, 39 (5): 45-47, 53.

5 张昕, 缪妹妹, 戴作雷, 等. 临床数据向通用数据模型转换研究及应用实践 [J]. 中国数字医学, 2018, 13 (10): 64-67.

6 宋欣霞, 金卫. 基于医疗数据的 K-means 算法优化研究 [J]. 中国数字医学, 2016, 11 (8): 22-24.