参赛队员姓名：陈天弈，何乃成

学校：南京外国语学校

指导教师：柏涛

省份：江苏省

国家/地区：中国

论文题目:儿童自闭症预测模型构建及其价值研究

儿童自闭症预测模型构建及其价值研究

摘要

自闭症是一种广泛存在的神经系统发育障碍，常起病于儿童时期，主要表现为社交障碍、兴趣狭窄、重复刻板行为。现有的自闭症诊断很大程度上依赖于诊断量表，但同时大部分量表的结果只是由简单累加得出的数值或者区间。自闭症谱系障碍疾病是复杂的神经系统疾病，早期的干预和治疗有利于患者的康复，但这需要建立在早而准确的诊断上。本文借助机器学习的手段，我们通过卡方检验的方式归纳出了量表中的主要特征量，并分别运用了 logistic 回归分析与 SMO(Sequential Minimal Optimization)+SVM(Support Vector Machine)思想构建了两种预测模型，并取得了相当不错的准确率。预测模型的本身及过程中取得的一些结论都对自闭症的预测、检查普及、早干预治疗的提醒有着实用意义。本研究的创新点在于将先进的机器学习手段融入到了疾病预测中，有利于智能医疗的发展与后期个性化精准医疗方案的设计实施。

关键词：自闭症预测，神经发育障碍，诊断量表，机器学习，支持向量机

Abstract

Autism is a widespread disorder in neurological development, often onset in childhood, mainly manifested as disorders in social interactions, narrowness of interests, and repeated, mechanical actions or body movements. Existing diagnoses of autism largely rely on scales, but the indicators used in most scales are simply numerical summations of values or intervals. Autism Spectrum Disorders (ASD) are complex neurological diseases, where early intervention and treatment proved to be conducive to the recovery of patients, which is not possible without early and accurate diagnosis. Through chi-square test, this research employs machine learning to quantify the major characteristic values of the diagnostic scales. Based on logistic regression analysis and the idea of SMO (Sequential Minimal Optimization) + SVM (Support Vector Machine), our research designed two prediction models that exhibit statistically significant accuracy. The prediction models and some conclusions obtained in the process have practical implications regarding the prediction of autism, the popularization of diagnosis, and early intervention and treatment. This research proves innovative in that it incorporates advanced machine learning methods into disease prediction, which contributes to the development of intelligent healthcare and facilitates the design and implementation of precise, personalized health schemes.

Key Words: Autism prediction, Neurodevelopmental disorders, Diagnostic scale, Machine learning, SVM

本参赛团队声明所提交的论文是在指导老师下进行的研究工作和取得的研究成果。尽本团队所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人或本团队已经发表或撰写过的研究成果。若有不实之处，本人愿意承担一切相关责任。

参赛团队签名：陈天弈 何乃成

指导老师： 柏涛

日 期：2020.9.15

目录

一、 研究背景

自闭症（Autism）是自闭谱系障碍（Autism Spectrum Disorder）的简称，通常起病于儿童时期，是以社交互动障碍、言语或非言语表达困难、重复刻板行为和兴趣狭窄为主要特征的神经发育障碍[1][2]。从 WHO 公布的数据来看，世界上每 160 个孩子中就约有 1 位自闭症患者，且在近半个世纪来患病率呈上升趋势[3]。美国疾病预防控制中心 2016 年的资料显示 8 岁年龄儿童中有 1.85%的比例被诊断为自闭[4]。全球范围内，自闭症给患者及其家庭带来巨大的经济负担与情绪上的打击[5-6]。由于没有器质性病变或者具有指向性的生化指标，自闭症的诊断很大程度上依赖于基于量表的行为评估。目前没有治愈自闭症的治疗方式或药物干预，但有些行为疗法可以有效恢复患者的部分症状。因为在儿童发育的早期大脑可塑性较强，随着自闭症干预技术的发展，已经有研究认为在如果在自闭症早期诊断、干预，康复等效果会更加显著[7]。因此，更加高效与准确的诊断、预测手段对自闭症治疗具有更大程度上缓和症状、有利于设计个性化治疗方案等重大意义[8]。

自闭症症状复杂，因而涉及的诊断量表数量多、类目复杂。本文选取了其中认可度高、应用较普遍的三种，分别为国际疾病分类（International Classification of Diseases， ICD）中收录的孤独症诊断量表 ICD-10，儿童孤独症评定量表 CARS（Children Autism Rating Scale）和感觉统合能力测评。这些量表的评估结果通常是由简单累加得出的一个数值或区间，以达到在人流量大的医院、评估机构控制检验效率与成本的目的。最终诊断由专业的医生结合量表结果、患者主述或家人代述、对患者行为和状态的观察等因素得出。

现如今，对于症状类型复杂、传统诊断耗时久的精神疾病，人工智能手段在医学的不同领域得到越来越广泛的运用[9]。如借助统计模型、机器学习算法等的协助，已经有学者搭建出了较为精确的抑郁症自杀预测模型[10]，客观且高效。自闭谱系障碍的概念来源于经过临床研究与观察，很多患者在传统的社交障碍、语言交流障碍、重复刻板行为三个方面分别的缺损程度相差迥异；此外，在神经系统缺陷的核心症状之外，一些自闭症患者还会有一些出现在免疫、消化等系统的外围症状。自闭症异质性与多类型症状的本质决定了它牵扯变量的丰富程度，也凸显了多因素分析对于诊断甚至探究内在规律的必要性[11]。此类大数据算法下分析与学习手段的应用能统合数量庞大、种类繁多的数据，洞察内在规律，更高效地预测发病和行为，有效地综合自闭症量表中繁复的各类指标。因此，机器学习在自闭症预测模型中的应用有利于尽早诊断与干预。

二、 研究目的

通过对自闭症量表条目的细分和统合，借以信息手段构建有一定灵敏度的预测模型，为做出高效、精确的诊断提供帮助。同时探究机器学习中得出的规律，完善对自闭症疾病表现的了解，同时也可以为后期患者提供个性化治疗方案提供参考。

三、 研究方法和结果

3.1 数据整理

选取 2020 年 5-7 月在南京天佑儿童医院就诊患者的检查数据，符合条件的共 542 组。原始表格形式如下：

1. 感觉统合评测

表 1 感觉统合评测表格

测试结果	
项目	结果评价
前庭和双脑分化程度	正常
脑神经生理抑制状态	轻度
触觉防御和脾气敏感状况	轻度
发育期运动和日常操作运用	边缘
空间形态与视知觉	边缘
本体感（重力不安全症）	正常
学习、情绪与自我形象感	正常
心理承受压力及行为表现	正常
总分与总评价	轻度

感觉统合失调干预建议：

1. 出现 2 个以上重度、3 个以上中度，建议 10 个周期以上的感觉统合训练（每个周期为 12 次）
2. 出现 1 个重度、3 个以下中度，建议 7 个周期以上感觉统合训练。
3. 出现 2 个中度、3 个以上轻度，建议 5 个周期以上感觉统合训练。
4. 出现 1 个中度、3 个以下轻度，建议 3 个周期以上感觉统合训练。
5. 出现 2 个轻度、3 个以上边缘，建议 2 个周期以上感觉统合训练。
6. 出现 1 个轻度、3 个以下边缘，视具体情况安排感觉统合训练。
7. 此报告仅供此次参考

2. 孤独症 CARS 测评

表 2 孤独症 CARS 测评表格

序号	项目名称	得分
1	人际关系	3
2	模仿(词和动作)	3
3	情感反应	2
4	躯体运用能力	2
5	与非生命物体的关系	3
6	对环境变化的适当	2
7	视觉反应	2
8	听觉反应	2
9	近处感觉反应	2
10	焦虑反应	2
11	语言交流	3
12	非语言交流	2
13	活动水平	2
14	智力功能	2
初步结果		32 分

3. ICD-10 孤独症

表 3 ICD-10 孤独症测评表格

序号	测试内容							结果
1	社会交往、语言交流和兴趣爱好有明显缺陷							至少 6 条
项目	序号	结果	项目	序号	结果	项目	序号	结果
言语交流行为缺陷	(1)	√	言语交流质的缺陷	(1)	√	刻板、重复、局限的行为、兴趣与活动	(1)	√
	(2)	√		(2)			(2)	
	(3)	√		(3)	√		(3)	
	(4)			(4)	√		(4)	√
2	功能异常或延迟，且出现在 3 岁以前						(1)	√
							(2)	√
							(3)	√
3	并非瑞特综合症或儿童瓦解性精神障碍							

医院原始数据嵌套了四层文件夹，最后一层文件内容为某一天某患者来做的所有检查报告，大约每个患者会做 5-8 个检查不等，这几个检查数据是分开存储的。假设有 500 个患者来检查，都只做了 5 个检查，就会有 2500 份文档中的数据需要提取，所以在前期如果手动提取记录数据，会造成人为错误且工作量巨大。所以为了使后续工作更加可靠，同时也能使得一整套检查更加智能、高效，用 python 撰写了数据处理代码，能在短时间内准确地提取出想要的信息量。实现步骤如下：

因为 python 不能处理 doc 文档，先把全部 doc 文档转为 docx(注：会覆盖原来的 doc 文档)，代码见附录 A。此处我们用了 docx 包和 win32com 中的 client 包，打开.doc 文件后在文件名后加‘x’，并另存为.docx 文件。

文档转换完成后，对每个表进行遍历读取数据，以读取全部“孤独症 CARS”表格为例。

首先根据“孤独症 CARS.docx”文档中要提取的关键信息来设计数据存入的表格头部，如在“孤独症 CARS.docx”中，我们想得到如下信息：姓名，编号，人际关系，模仿(词和动作)，情感反应，躯体运用能力，与非生命物体的关系， 对环境变化的适当，视觉反应，听觉反应，近处感觉反应，焦虑反应，语言交流，非语言交流，活动水平，智力功能，就将这些作为表头，在 python 中操作表格需导入 openpyxl 包中的 Workbook。接着遍历全部文件夹中的“孤独症 CARS.docx”文档，读取文件需提前导入 docx 包中的 Document，再将文档中对应表头的内容按行列提取出来，存入 excel 表格对应位置，遍历完成后通过指令 workbook.save('$$path\\孤独症 CARS.xlsx')将表格进行存储。

现分析主要基于“孤独症 CARS”、“ICD-10 孤独症”、“感觉统合测评报告单”三个表格，再将三个表格整合到一起，得到总表，总表部分数据如图：

		感觉统合测评									孤独症CARS								
自闭情况	编号	和双脑化	经生理抑制	脚和脾气	动和日常	形态与视	（重力不安	情绪与自我	受压力及行	总与总评价	人际关系	与（词和动作	情感反应	体运用能	生命物体的	境变化的	视觉反应	听觉反应	处
0	2005010028	2	0	1	0	0	0	0	0	2	2	1	2	3	2	2	2	2	
0	2005010036	0	2	0	2	0	1	1	0	2	3	3	2	3	2	2	2	1	
1	2005010041	1	2	1	2	0	0	1	0	2	3	3	2	2	3	3	2	3	
1	2005010043	1	2	0	0	0	0	1	0	2	2	3	2	2	3	2	2	2	

		孤独症CARS									ICD-10孤独症								
自闭情况	编号	视觉反应	听觉反应	处感觉反	焦虑反应	语言交流	非语言交流	活动水平	智力功能	总分	1	2	3	4	1	2	3	4	1
0	2005010028	2	2	2	1	3	2	2	2	28	1	1	0	0	1	0	1	0	1
0	2005010036	2	1	2	1	3	2	3	2	31	1	1	1	1	1	0	1	1	1
1	2005010041	2	3	2	2	3	2	2	2	34	1	1	1	1	1	0	1	1	1
1	2005010043	2	2	1	2	3	2	2	2	30	1	0	1	0	1	0	1	1	1
0	2005010030	1	2	2	1	3	2	2	2	24	0	0	0	0	1	0	1	1	1

图 1 总表部分数据

在进行机器学习前，先把所需数据手动转为 txt 文件，txt 文件包含的内容：“感觉统合测评报告单”、“孤独症 CARS”、“ICD-10 孤独症”三表全部数据作为特征量，再加上自闭情况作为最后一列(0 表示没有自闭，1 表示自闭)。txt 部分数据如图：

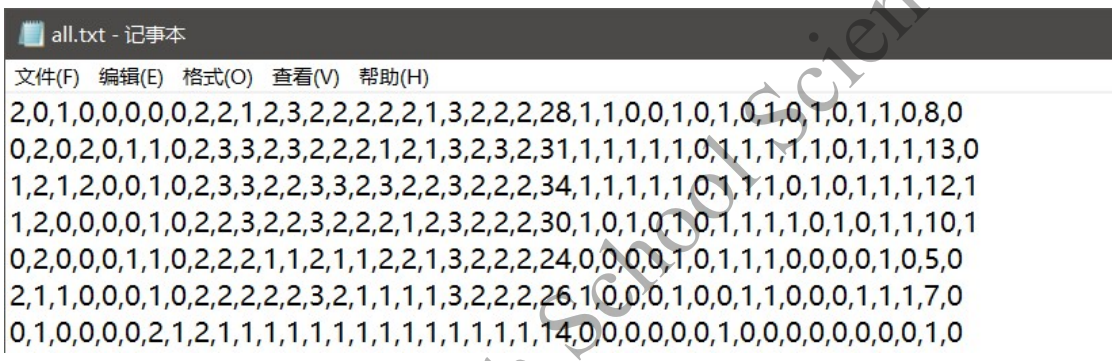

图 2 特征量部分数据

总共 542 行，40 列数据，每一行代表一组数据，每一行中的前 9 列表示感觉统合测评的各项数据，中间 15 列表示孤独症 CARS 的各项数据，后 15 列表示 ICD——10 孤独症的各项数据，最后一列代表诊断结果(1 表示诊断为自闭症，0 表示诊断为非自闭症)。

3.2 机器学习

3.2.1 均值归一化处理

不同的指标往往具有不同的量纲和单位，这样的不均一会影响到数据分析的结果，为了消除不同指标之间的量纲影响，需要进行数据标准化处理，使数据之间具有可比性，标准化数据是进行数据挖掘的预备工作，原始数据经过数据标准化处理后，适合进行综合对比评价。

特征量里有评测的总分项，这一列数据和其他数据相差有点大，所以做了归一化处理。

我们采用的方法是均值归一化，首先要选取某列数据，得到该列数据的最大值 $x_{\max}$ 和最小值 $x_{\min}$，然后将该列每个数据进行归一化($x_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$)处理，使每个 x_i 都在(0,1)范围内，从而实现特征缩放，使梯度下降法能够更快的收敛。

3.2.2 特征选择

当数据处理完成后，我们需要选择有意义的特征量和模型进行训练。现一共有 40 个特征量，如果特征与目标相关性很高，那么特征就被优先选择，而相关性较低的特征量就可以抛弃不用。对于回归和分类问题可以采用卡方检验等方式对特征进行测试，测得自变量对因变量的相关性。

卡方检验的思路是通过观察实际值和理论值的偏差来确定原假设是否成立。首先假设两个变量是独立的（此为原假设），然后观察实际值和理论值之间的偏差程度，若偏差足够小，则认为偏差是很自然的样本误差，接受原假设。若偏差大到一定程度，则否定原假设。例：检验吸烟与否与健康是否独立，原假设为吸烟与健康独立，如果卡方检验测得 p 值为 0，就是拒绝吸烟与健康独立无关，说明吸烟与健康是相关的，而 p 值越小表示相关性越强。

三个表卡方检验得到的 p 值如下表所示：

表 4 感觉统合测评特征量相关性

感觉统合测评特征量	p 值
前庭和双脑化程度	0.25008036
脑神经生理抑制状态	0.39117843
触觉防御和脾气敏感状况	0.04221487
发育期运动和日常操作运用	0.06650252
空间形态与视知觉	0.13543413
本体感(重力不安全症)	0.15007178
学习、情绪与自我形象感	0.07120045
心理承受压力及行为表现	0.10306098
总分	0.31163739

表 5 孤独症 CARS 特征量相关性

孤独症 CARS 特征量	p 值
人际关系	0.00172878
模仿(词和动作)	0.02058875
情感反应	0.00224271
躯体运用能力	0.01818358
与非生命物体的关系	0.03735618
对环境变化的适当	0.00399149
视觉反应	0.06769808
听觉反应	0.01841675
焦虑反应	0.00089975
语言交流	0.03174428
非语言交流	0.30700938
近处感觉反应	0.68931038
活动水平	0.20431492
智力功能	0.15601021
总分	0.02099851

表 6 ICD-10 孤独症特征量相关性

ICD-10 孤独症特征量	编号	p 值
言语交流行为缺陷	1	1.20750784e-02
	2	1.29875093e-05
	3	1.44507467e-03
	4	5.08721794e-06
言语交流质的缺陷	1	8.50329841e-01
	2	5.47869906e-01
	3	3.95254880e-01
	4	5.31564751e-03
刻板、重复、局限的行为、兴趣与活动	1	4.17188566e-05
	2	7.63923534e-06
	3	3.69259673e-04

	4	4.38980731e-03
功能异常或延迟，且出现在 3 岁以前	1	2.53877948e-06
	2	4.40109005e-01
	3	1.23744700e-04

经过测试，最终选取以下列作为特征量输入：

1,2,3,4,5,6,7,8,9,10,11,12,15,18,24,26,27,28,33,34,35,37,39

这些特征量中 1~9 为感觉统合测评中 p 值<0.4 的特征值，分别为前庭和双脑化程度、脑神经生理抑制状态、触觉防御和脾气敏感状况、育期运动和日常操作运用、空间形态与视知觉、本体感(重力不安全症)、学习、情绪与自我形象感、心理承受压力及行为表现。

特征量中 10、11、12、15、18、24 表示孤独症 CARS 表中 p 值<0.01 的特征值，分别为人际关系、模仿(词和动作)、情感反应、对环境变化的适当、焦虑反应及智力功能。

特征量中 26、27、28、33、34、35、37、39 表示 ICD-10 孤独症中表中 p 值<0.002 的特征值。其中 26、27、28 代表“言语交流行为缺陷”的后三项指标，33、34、35 代表了“刻板、重复、局限的行为、兴趣与活动”的前三项指标，37、39 代表了“功能异常或延迟，且出现在 3 岁以前”的第一项和第三项指标。

选择部分特征量可有效提高训练的速度，在选择特征量时，一开始的想法是将所有特征量合并到一起进行排序，选 p 值较小的，但是这样选择后，所得的训练正确率较低，训练出的模型无实用价值。将三个表里的特征量各自取 p 值较低的进行训练后，和全部特征量都选取时，在小样本上的训练正确率是一致的，且训练速度快。按照上面的原则选取的特征量，与全特征量参与训练的正确率几乎完全一致，可以在速度与性能之间取得一个较好的平衡。

3.2.3 logistic 回归分析

线性回归是最基本的回归模型，它使用线性函数描述变量之间的关系，将连续或离散的自变量映射到连续的实数域。

预测函数： $\mathrm{h} = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$

要求得预测函数，就必须把参数 θ 求出来，引入误差平方代价函数：

$$J(\theta) = \frac{1}{2\mathrm{m}} \sum_{i=1}^{m} (h_\theta(x^{(i)}), y^{(i)})^2$$

我们的目标便是通过学习算法，最小化 $J(\theta)$，从而求出理想的模型参数 θ。流程图如下：

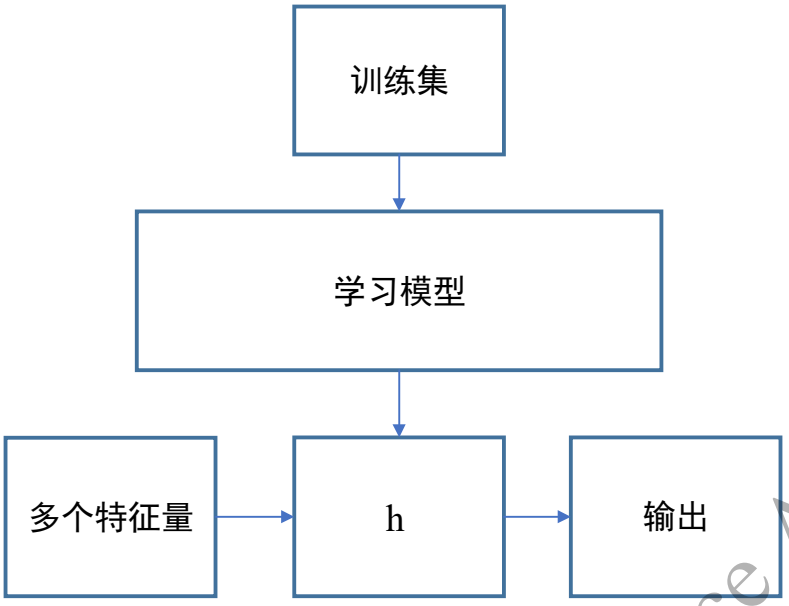

图 3 线性回归流程图

logistic 回归的流程与线性回归相似，不同的是：线性回归输出值可能大于 1 或小于 0，而对于一个二分类问题，输出值永远在 0 到 1 之间，因此要嫁给输出值映射到 0-1 之间，预测函数变为：$h_\theta(x)=g(\theta^T x)=\frac{1}{1+e^{-\theta^T X}}$，代价函数也因此会有变化。

Logistic 回归的目的是从特征学习出一个 0/1 分类模型，而这个模型是将特性的线性组合作为自变量，由于自变量的取值范围是负无穷到正无穷。因此，logistic 函数（或称作 sigmoid 函数）$h_\theta(x)=g(\theta^T x)=\frac{1}{1+e^{-\theta^T X}}$ 的估计值表示了输入值 x 的条件下得到 y=1 的概率，即该样例为正类的概率，$h_\theta(x)=P(y=1|x;\theta)$。从而，当我们要判别一个新来的特征属于哪个类时，只需求 $h_\theta(x)$ 即可，若 $h_\theta(x)$ 大于 0.5 就是 y=1 的类，反之属于 y=0 类。

预测函数：$h_\theta(x)=g(\theta^T x)=\frac{1}{1+e^{-\theta^T X}}$，其中 $\theta^T x=\theta_0 x_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_n x_n$。

代价函数(cost function)：

$$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}),y^{(i)})^2=-\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log h_\theta(x^{(i)})+(1-y^{(i)}+)\log(1-h_\theta(x^{(i)}))]$$

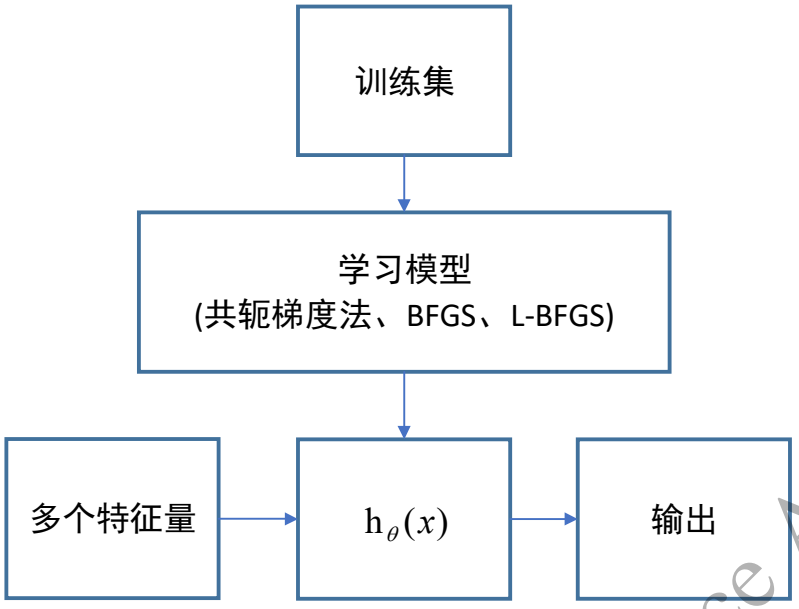

图 4 logistic 回归流程图

进行 logistic 回归分析的大概步骤如下：

首先用 pandas 读取数据，数据处理好之后用.values 转化为矩阵进行运算，再定义 sigmoid 函数、代价函数和梯度，带入更高级的算法中自动优化求解参数。

代码现在分别用三种高级算法来实现：共轭梯度法(conjugate gradient)、拟牛顿法(BFGS、L-BFGS)。

其中共轭梯度法是一个迭代方法，该系统经常用于数值求解偏微分方程，因此它适用于稀疏矩阵系统或求解无约束的最优化问题。它的优点是实现起来比较容易，而且当训练样本足够多时优化速度非常快。缺点是需要人为调整一些参数，比如学习率，收敛准则等。

拟牛顿法是最优化算法中最有效的一类算法。优点是迭代中仅需一阶导数，不需要 Hessen 阵，对于一般情形，具有超线性收敛速率，而且还具有 n 步二级收敛速率。而它的缺点是所需的存储量较大，对于大型问题会面对存储方面的困难。

表 7 三种算法正确率

逻辑回归算法	正确率
BFGS	0.7785977859778598
牛顿共轭梯度	0.7785977859778598
L-BFGS-B	0.7785977859778598

由于逻辑回归采用的是线性回归，在样本容量为 542 的情况下，回归函数只有 26 个有效参数，而三种回归算法得到的准确样本数恰好都为 422，因此正确率在三种逻辑回归算法下完全相同。

拟合数据的一个更好的方法是从每个数据点创建更多的特征，我们将这些特征映射到所有 x 的多项式上，直到第六次幂，相当于通过多项式核函数将低维向量映射到高维空间，因此可能出现过拟合，这时在代价函数中加入 L2 正则化条件可以弥补过拟合带来的影响，正确率如下表所示：

表 8 加入多项式核函数后三种算法正确率

逻辑回归算法	正确率
BFGS	0.8062730627306273
牛顿共轭梯度	0.8007380073800738
L-BFGS-B	0.7952029520295203

从表中可以看出，在引入多项式核函数之后，回归函数的参数个数大大提高，正确率也得到一定的提升。因为加入了正则化条件，也不会因为参数的个数过多而发生过拟合。在这种情况下 BFGS、牛顿共轭梯度以及 L-BFGS 的样本正确个数分别为 437、434、431，拟牛顿法(BFGS)表现最好，正确率达到 80.63%。

3.2.4 logistic 回归与 SVM 的联系

尝试把 logistic 回归做个变形。首先，将使用的结果标签 y = 0 和 y = 1 替换为 y = -1,y = 1，然后将 $\theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$（$x_0 = 1$）中的 θ_0 替换为 b，最后将后面的 $\theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$ 替换为 $w^T x$。如此，则有了 $\theta^T x = w^T x + b$。也就是说除了 y 由 y=0 变为 y=-1 外，线性分类函数跟 logistic 回归的形式化表示 $h_\theta(x) = g_\theta(x) = g(w^T x + b)$ 没区别。

3.2.5 SVM

支持向量机（Support Vector Machine，SVM）是一种用于分类和回归的线性监督机器学习方法，主要用于分类，而且不会造成过拟合的问题[12]。由于核函数的存在，SVM 算法可以通过将数据转换为更高的空间来解决非线性可分问题[13]，通过选择合适的核函数，达到良好的分类效果[14]。

SVM 要拟合的不是一个超平面，而是两个超平面，且这两个超平面之间的距离要尽可能大。主要关注距离分类边界近的边缘数据点（支撑向量 support vector），通过凸优化方法使得正例和反例之间的间隔最大，从而使边界超平面附近的数尽可能少，避免边界附近样本点的误判，提高正确率。

用 $\mathrm{label} \bullet (w^T x + \mathrm{b})$ 来表示点到分割面的函数间隔，$\mathrm{label} \bullet (w^T x + \mathrm{b}) \bullet (1/\| \mathrm{w} \|)$ 来表示点到分割面的几何间隔。那么, 当下的研究目标就是找出分类器定义中的 w 和 x，即找出其支持向量。一旦找到具有最小间隔的数就要对该间隔最大化， 函数可以写作：

$$\arg\max_{\mathrm{w,b}}\{\min_{\mathrm{n}}(\mathrm{label} \bullet (w^T x + \mathrm{b})) \bullet \frac{1}{\| \mathrm{w} \|}\} \tag{1}$$

直接求解会比较棘手，这是一个给定了约束条件后求最优解的问题，通过引入拉格朗日乘子，就可以基于约束条件，将超平面写成数据点的形式，优化目标函数也可以写成：

$$\max_{\alpha}[\sum_{i=1}^{m}\alpha - \frac{1}{2}\sum_{i,j=1}^{m}\mathrm{label}^{(i)} \bullet \mathrm{label}^{(j)} \bullet \alpha_i \bullet \alpha_j \bullet < x^{(i)}, x^{(j)} >] \tag{2}$$

约束条件为：

$$C \geq \alpha \geq 0 \sum_{i=1}^{m}\alpha_i \mathrm{label}^{(j)} = 0 \tag{3}$$

其中，常数 C 用于控制“最大化间隔”和“保障大部分点的函数间隔小于 1.0”这 2 个目标的权重，称之为松弛变量。在高维特征空间中，不需要完全按照上述变换进行特征变换，只需进行内积运算就可以实现，而且内积运算可以利用原空间中的函数，不需要知道变换的形式。根据泛函分析的有关理论，如果某一选定核函数 $\mathrm{k}(x, x_T)$ 满足 Mercer 规定的相关条件，它就会对应某一个高维特征变换空间的内积。由此可知，只要能在最优分类面中找到适当的内积函数，就可以实现非线性变换后的线性分类，而且不会增加计算的复杂度[15]。

核函数的引入为有效地解决线性不可分问题提供了方法。预判评价对象的复杂程度，然后根据评价对象的复杂性确定核函数和最优参数值。这是建立支持向量机模型的核心内容，关系到模型的有效性。实践中，通常会根据样本数据选择核函数和参数值。常见的核函数类型见表[16]。

表 9 常见核函数类型

核函数类型	表达式	参数取值
线性核	$k(x_i, y_i) = x_i^T x_j$	
多项式核	$k(x_i, y_i) = (x_i x_j)^d$	$\mathrm{d} \geq 1$
拉普拉斯核	$k(x_i, y_i) = \exp(-\frac{\| x_i - x_j \|^2}{2\sigma^2})$	$\sigma > 0$
高斯核	$k(x_i, y_i) = \exp(-\frac{\| x_i - x_j \|}{\sigma})$	$\sigma > 0$
Sigmoid 核	$k(x_i, y_i) = \tanh(\beta x_i^T x_j + \theta)$	$\beta > 0, \theta < 0$

当特征数量多、训练集数量较少时，一般选用逻辑回归或者不带核函数的 SVM（线性核函数）。

最后，SVM 会得到得到两个距离最远的超平面 B、C，如图 5 所示，然后选择超平面 A 作为决策超平面因为使用超平面 A 进行划分对训练样本局部扰动的“容忍”度最好，分类的鲁棒性最强。例如，由于训练集的局限性或噪声的干扰，训练集外的样本可能比图 5 中的训练样本更接近两个类目前的分隔界，在分类决策的时候就会出现错误，而超平面 A 受影响最小，也就是说超平面 A 所产生的分类结果是最鲁棒性的、是最可信的，对未见样本的泛化能力最强。超平面 A 即是最后 SVM 得到的分类超平面。

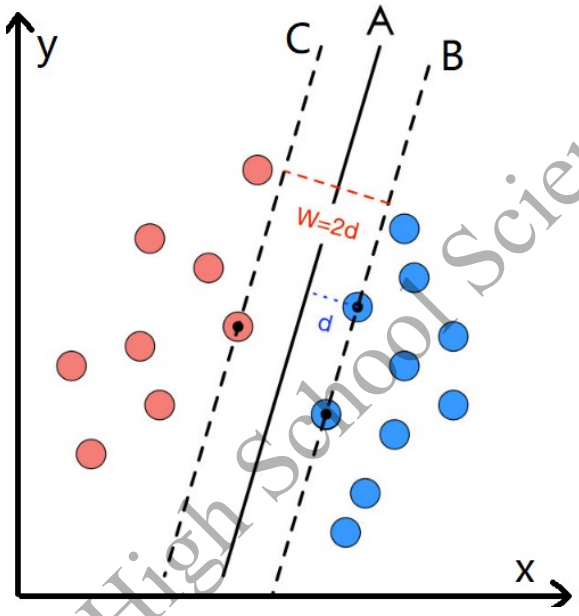

图 5 SVM 超平面示意图

3.2.6 SMO

由 Platt 开发的 SMO (Sequential Minimal Optimization， 序列最小优化)是 SVM 训练阶段最有效的解决方案之一，能高效解决 SVM 中的对偶问题。 它是通过求解一系列二次规划(QP)问题来导出的，在每个迭代器中，在工作集中选择两个变量 alpha(α1，α2)进行优化处理，如果 α1 确定，α2 也随即可得[16]。这些被逐个分解的 QP 往往更容易计算，因此一般情况下 SMO 有着更低的时间和空间复杂度。 为了判断是否收敛，可以检查 KKT 条件,由于 KKT 条件本身是比较苛刻的，所以需要设定一个容忍值 tol，一般来说这个参数的值在 0.001 到 0.01 之间,即所有样本在容忍值范围内满足 KKT 条件则认为训练可以结束，实现步骤即重复以下过程直到收敛：

（1）选择两个拉格朗日乘子 α_i 和 α_j；

（2）固定其他拉格朗日乘子 α_k (k 不等于 i 和 j)，只对 α_i 和 α_j 优化;

（3）根据优化后的α_i和α_j，更新截距 b 的值[15]；

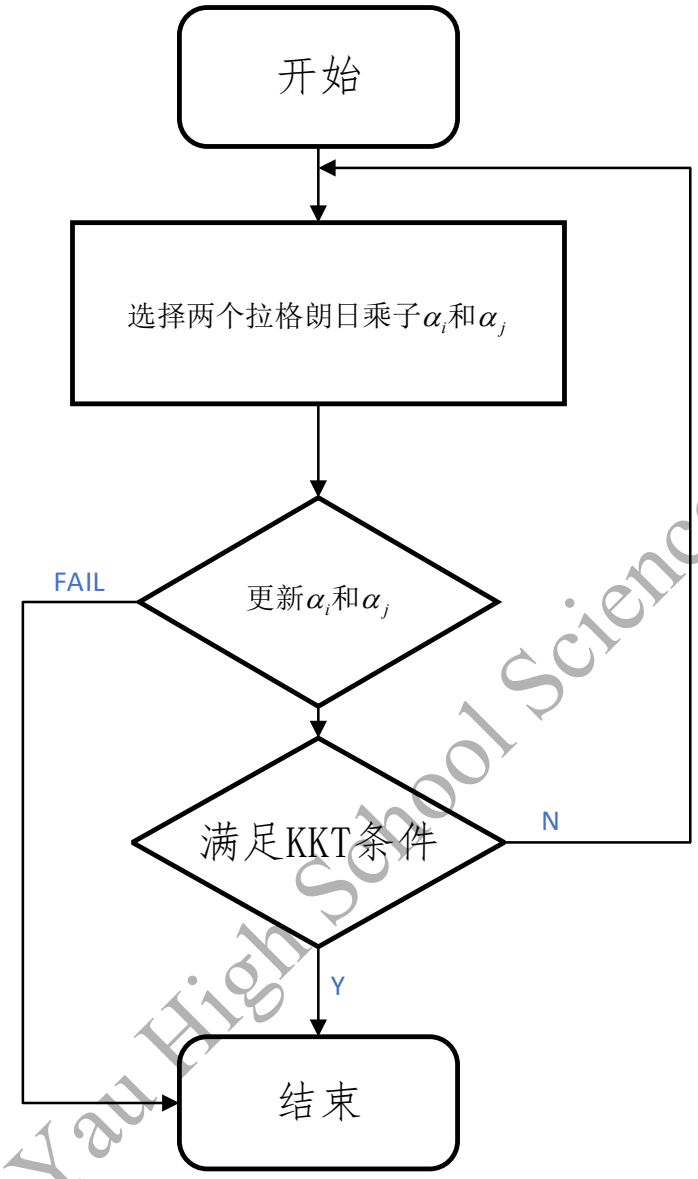

图 6 SMO 算法流程图

3.3 模型有效性检验

在构建 SMO+SVM 预测模型时，关键在于确定惩罚参数 C、SVM 停止训练的误差精度 toler、迭代次数和核参数。如果这几个主要参数设置不合理、参数值过大或过小，将直接导致训练过程中产生过拟合或者拟合不足的问题，进而影响评价模型的精准度。

现有样本总数为 542 个，随机选取 500 个样本作为训练集，另外的 42 个作为测试集，通过导入 python 中 sklearn 算法包，用 SVM 训练数据，然后测试得到的最高准确率为 87.805%，如下图所示：

图 7 部分代码及训练后准确率

四、 分析与讨论

此刻，全球自闭症的患病率已近 1%， 对自闭症的尽早预防与诊断对无数的家庭乃至整个社会具有重大意义。在自闭症的诊断过程中，需要用到孤独症诊断量表 ICD-10、儿童孤独症评定量表 CARS 与感觉统合能力测评量表。现有的诊断量表依靠简单累加得出结果以达到控制成本、保证效率的目的，然而不变的标准很难适应随着时代变化疾病整体特征的变迁以及尚未被发现的特征的内在联系。由于牵扯到症状变量的规模与复杂程度和基于患者特点随时学习调整的必要，新兴的人工智能机器学习手段是值得一试的选择。借助机器学习手段得出更快、更准的预测结果，对于自闭症及潜在群体得到尽早干预和自闭症检测的普及有很大帮助。本文基于一定体量的患者数据，分别运用了 logistic 回归分析与 SMO+SVM 思想构建了预测模型并取得了证明其有应用价值的准确率。

在特征选择的模块中，由单因素分析得出的卡方检验特征量让细分条目有了与结果相关性更直观的体现，能够反映各条目不同的重要性，而由于有一定的数据量，得出的结果有统计学意义。对于自闭症在内的很多心理疾病，现在很重要的问题是人们不太有检查、预防的意识。病发率比大多数人想象的高，而很多家长都不会想起来在孩子有些微自闭迹象的时候通过某种方式筛查。在这样的情境下，一份描述通俗易懂、条目具有代表性的量表在时间和成本上都高效，对于基本的预测很有帮助。从这一特征选择结论中筛选出的相关性高的特征量有利于编纂这样一份更简洁、且有相当准确率的量表。大大降低了医生和患者的工作难度和门槛，有利于自闭症的快速诊断和测试普及。

本研究基于量表利用两种方法构建了四种算法的预测模型。以往研究显示误判率小于 10%或 20%就有良好的应用价值[8]。本研究中的 logistic 回归模型的最高准确率是 80.63%，

而利用 SMO+SVM 思想构建的预测模型最高准确率达 87.805%，已经超越 81%准确率的传统量表使用标准最低值，是颇有意义的成果。

由 logistic 回归分析思想下的三种算法正确率虽不够高，但在改进之后也与量表结果基本持平，且主要胜在速度快。这个系统适合用于门诊的风险筛查，可以快速完成评估，节省人力物力。诚然，这种方法也有一定的缺陷，要提升速度就得牺牲受试的时间，可能会出现做评估者没完全理解和思考、过于主观的情况，结果不够准确。在这一点上需要后期对流程的完善，更好地帮助实现又快又准的评估。

SMO+SVM 模型最高正确率达到了 87.805%，鉴于传统量表准确率达到 81%就可以投入使用，这样的准确率无疑十分拔群。随着后期数据量的增加，这样的预测结果会越来越准，不会因为环境因素或者人们检查意识的提高而骤然改变。这样的模型比较适合已确诊患者的后期跟踪，有利于制定跟进较久的个性化治疗方案。一个存在的问题是由于收集的数据时间、测试机构较为单一，与群体比较可能会存在偏差，但还是存在参考价值，因为如果真正投入使用其中一些偏差依旧存在。另一个问题是从一部分样本数据相差不大，但最终结果却不一致。这或是因为自闭谱系障碍每个人在各方面的病症表现都相差甚远，加之在面诊之中也会得到一些量表难以涉及的内容用以辅助诊断，但终究模型预测结果提供了便捷、可靠的参考。

五、 研究结论

机器学习方法在基于量表的儿童自闭症预测模型构建上取得了比传统量表更高的准确率，有一定的研究意义，有助于自闭症意识与筛查效率的提高、新量表编纂和后期跟进治疗，希望未来此研究能发挥它的作用。也很希望今后也有更多的预测模型在量表的基础上增加眼动追踪、功能性磁共振成像这些更为客观的因素，探究其与自闭症的关系。准确的预测自闭症是为了更早地干预疾病的发展，给患者和家庭带来力所能及的帮助，节约时间和财力，更重要的是，努力康复、拥有健康的身体。

六、 参考文献

1. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders: DSM-V. (5th edition). Washington, D.C.: American Psychiatric Publishing, 2013
2. Society of Neuroscience. 2018. *Brain Facts*. 136pp.

3. Mayada et al. Global prevalence of autism and other pervasive developmental disorders. Autism Res. 2012 Jun; 5(3): 160–179.
4. Maenner MJ, Shaw KA, Baio J, et al. Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016. MMWR Surveill Summ 2020, 69 (No. SS-4): 1-12.
5. Vargason, T., Grivas, G., Hollowood-Jones, K. L., & Hahn, J. (2020). Towards a Multivariate Biomarker-based Diagnosis of Autism Spectrum Disorder: Review and Discussion of Recent Advancements. Seminars in Pediatric Neurology, 100803.
6. 柳毅恒. 基于循环神经网络的自闭症谱系障碍预测框架[D]. 2018.
7. Aron L, Loprest P: Disability and the education system. Future Child 22:97-122, 2012
8. 孙宾宾. 2-6 岁孤独症谱系障碍儿童眼动注视特征研究及诊断预测模型的建立[D]. 2018.
9. LU H L,TANG J,HUANG Z W, et al. Development and value evaluation of a simple prediction model of suicidal behavior in depressive disorder［J］. Chinese General Practice,2020,23(26)：3247-3252.
10. Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting Risk of Suicide Attempts Over Time Through Machine Learning. Clinical Psychological Science, 5(3), 457– 469.
11. McPartland JC: Considerations in biomarker development for neuro- developmental disorders. Curr Opin Neurol 29:118-122, 2016
12. Nell Critianini，John Shawe-Tayer. 支持向量机导论[M].北京:电子工业出版社.2004:82-98.
13. 王华忠，俞金寿.核函数方法及其模型选择[J].江南大学学报(自然科学版)，2006，5(4):126 -130.
14. 周晓剑，马义中，朱嘉钢. SMO 算法的简化及其在非正定核条件下的应用［J］. 计算机研究与发展 2010，47(11) : 1962－1969.
15. 赵长春，姜晓爱，金英汉. 非线性回归支持向量机的 SMO 算法改进［J］. 北京航空航天大学学报，2014，40(1) : 125－130.
16. 李航. 统计学习方法［M］. 北京: 清华大学出版社，2012.

七、 成员介绍与致谢

团队成员陈天弈为南京外国语学校高三学生，热爱生命科学，在生物学奥林匹克中获得国家二等奖；在英国生物竞赛（BBO）中获得银奖；赴哥伦比亚大学参加了神经科学的夏

校，在 2020 年 Brain Bee 神经科学中国区比赛中获得全国二等奖，并在比赛中认识指导老师柏涛。

团队成员何乃成为南京外国语学校高二学生，热爱数学、编程，在数据分析、Python 编程方面有很高的兴趣，在杜克数学大赛（DMM2020）中国区比赛中获得个人一等奖，团队二等奖；在信息学奥林匹克竞赛中，获国家二等奖；在语言学奥林匹克竞赛国赛中，获得个人三等奖，团队二等奖；在圣路易斯大学高中（SLUH）交换半年。

团队于 2020 年暑假在天佑儿童医院参观学习，深感有神经发育障碍的小朋友和他们家庭的不易。此类疾病的原则是早诊断、早干预，可以达到更好的效果，但事实上非器质性疾病诊断不太有客观指标可以反映，加之社会上人们对此类疾病预测、预防意识淡薄，因而造成种种社会问题的出现。团队成员在其掌握的神经科学、大数据处理、机器学习等方面的知识基础上，预期将医院大量的诊断数据利用起来或许会有意想不到的效果。因而项目组在柏涛老师的指导下，开始了本论文的工作。

团队指导老师柏涛现为中国科学院脑科学与智能技术卓越创新中心的博士候选人，研究方向为创伤后应激障碍（PTSD）的神经机制和不连续事件关联学习记忆产生的机制。

陈天弈和何乃成同学共同参与完成了数据整理和处理、论文撰写的工作。项目中相关数据来自南京天佑儿童医院，获得南京天佑儿童医院伦理道德委员会的许可，用以非营利性科学研究，并隐去了患者个人相关信息。陈天弈同学完成了大部分神经科学相关信息的查找与知识储备，何乃成同学承担了数据处理、机器学习等算法的编程工作。柏涛老师在论文写作过程中与项目组成员头脑风暴，为项目组指引方向、提供思路，但重要的选择与决定都留给项目组完成，在此感谢柏涛老师所给予的无私指导！