

运用基因表达式编程的自陈量表数据建模*

钱锦昕 余嘉元

(南京师范大学心理学院, 南京 210097)

摘要 探讨基因表达式编程对自陈量表测量数据的建模方法。运用威廉斯创造力测验和认知需求量表获得 400 位中学生的测量分数, 通过数据清洗, 保留 383 个被试的分数作为建模的数据集。运用哈曼单因素检验方法没有发现共同方法偏差。采用均匀设计方法对基因表达式编程中的 5 个参数进行优化配置, 在测试拟合度最大的试验条件下, 找到了测试误差最小的模型。比较基因表达式编程和 BP 神经网络、支持向量回归机、多元线性回归、二次多项式回归所建模型的预测精度。研究表明, 基因表达式编程能用于自陈量表测量数据的建模, 该模型比传统方法所建的模型具有更高的预测精度, 而且模型是稳健的。

关键词 自陈量表; 基因表达式编程; 建模; 创造力; 均匀设计

分类号 B841.7

1 前言

在科学研究中, 数学模型是人类对于客观世界的一种重要表达形式。“所谓数学模型, 是关于部分现实世界为一定目的而作的抽象、简化的数学结构”(赵静, 但琦, 2003)。

在科学的各个领域, 数学模型都是学科的重要基石。在经典力学中, 牛顿用运动三定律表达了宏观世界物体运动的规律; 在电磁学中, 麦克斯韦方程组完整地描述了电磁场的基本规律; 在心理物理学中, 韦伯—费希纳定律反映了在中等强度刺激条件下心理量和物理量之间的关系; 在心理测量中, 逻辑斯谛模型表示了被试潜在特质、项目参数和被试正答率之间的概率关系。事实表明, 数学模型的建立都极大地推动了相应学科的发展。

近年来, 数学模型在心理学中得到了越来越多的重视, 相当多的心理学家根据实证研究的数据建立了相应的数学模型。长期以来, 人们都无法定量地描述吸毒和人格之间的动态关系, Caselles, Mico 和 Amigo (2010)对于可卡因成瘾和人格间的关系进行了系统研究, 并采用一组微分方程来建立模

型。根据这一模型, 吸毒者的某种人格特质是会受到毒品影响的, 其短期作用称为“急性效应”, 其长期作用就是“毒品成瘾”, 这一研究成果对于吸毒的预防和治疗是有积极作用的。早年与父亲或母亲分居的经历, 对于孩子成年后是否会产生不良心理影响, 这是发展心理学家感兴趣的问题。Lakshmi 和 Alamelu (2010)运用混合模型方法分析的结果显示, 孩子们的这种早期分居, 会影响其成年后的生理心理特质, 尤其是在 2 岁至 7 岁期间和父母分居的儿童, 其成年后产生的不良影响最显著, 这一研究成果对于儿童的早期教育是很有价值的。在自由操作实验中, 操作性反应率的变化是与被试类型、实验条件有关的, Bittar, Del-Claro, Bittar 和 da Silva (2012)在前人研究的基础上, 提出了动机动力学模型, 在将唤醒水平变量归一化后, 操作性反应率的变化可以用非常简洁的数学模型表示出来, 这对于揭示心理活动的规律是很有意义的。显然心理学家和其他领域的科学家一样, 对于数学模型表现出越来越浓厚的兴趣。

鉴于数学模型对于科学发展的重要性, 人们提出了多种数学建模的方法, 包括微分方程、差分方

收稿日期: 2012-11-19

* 国家社会科学基金教育学课题(BBA080050)、国家自然科学基金项目(71071065、71131004)、江苏省一级重点学科“心理学”资助成果。

通讯作者: 余嘉元, E-mail: yujiayuanwx@163.com

程、组合数学、回归分析、插值和拟合、计算机模拟、神经网络等。目前在心理学中应用较多的是回归分析,人们通常运用回归方程建模来获得各变量之间的关系,但它只能建立线性的或某些特殊函数关系的模型,对于更加复杂的关系就无能为力了。近年来,神经网络、支持向量回归机等方法也常被用于模型的建立,但是它们所获得的输出变量和输入变量之间的关系是隐含的,是无法用显性的数学表达式来描述的。那么是否可以找到某种建模方法,它既适应于变量间复杂的非线性关系,又可以用数学公式显性地表示出来呢?基因表达式编程(Gene Expression Programming, GEP)提供了这种可能性(Ferreira, 2001),它采用了和统计方法完全不同的思路,通过模拟生物进化过程中染色体的变化来寻找与数据最为拟合的数学表达式,这是一种智能计算的方法,该方法已经在科学和工程的多个领域中获得了成功的应用(Mollahasani, Alavi, & Gandomi, 2011; Mousavi, Aminian, Gandomi, Alavi, & Bolandi, 2012)。

但迄今为止, GEP 用于建模的数据大多数是通过客观方法获得的,这和心理学中通过自陈量表所得到的数据是有很大区别的。自陈量表由于操作简单而得到了广泛的应用,但在实施该方法时,被试的反应往往会受到社会称许性、宽大效应、默认偏向、积极或消极的心境、短暂的情绪状态等主观因素的影响,还受到项目的模糊性、量表的格式和锚定、负性词语、项目启动效应、项目嵌套、语境诱发的心情、量表长度、测量的时间、地点和环境等外部因素的干扰(Podsakoff, Mackenzie, Lee, & Podsakoff, 2003)。因此,它和通过仪器设备测量所得数据以及各种能力测验数据相比显然是有所不同的。

那么基因表达式编程(GEP)是否可以用于这类数据的建模呢?所建模型的误差是否在能够接受的范围内呢?它和其他智能计算方法(神经网络、支持向量回归机等)、统计方法(多元线性回归、多项式回归等)相比较,是否具有一定的优越性呢?这就是本研究要探讨的问题。

2 基因表达式编程

基因表达式编程(GEP)是葡萄牙科学家 Ferreira (2001)提出的一种新颖的随机搜索和优化的算法,它借鉴了生物遗传的基因表达规律,结合了遗传编程(GP)和遗传算法(GA)的优点,能够用

简单的编码解决复杂问题,得到了国际学术界的高度重视。

2.1 GEP 的编码形式

GEP 中的基因有两种表现形式:基因型和表现型,基因型为 K 表达式,表现型为表达式树(彭昱忠, 元昌安, 麦雄发, 覃晓, 2011)。GEP 的遗传编码是等长的线性符号串,称为染色体。一个染色体可以由多个基因组成,每个基因都由头部和尾部组成,头部可以包含终结符和函数符号,而尾部只能包含终结符。基因头部长度 h 是根据具体问题而定的,而尾部的长度 t 则可以由下述公式得到:

$$t = h * (n - 1) + 1 \quad (1)$$

其中 n 表示所使用的函数集中需要变量最多的函数的变量个数。GEP 的染色体可以包含一个或多个基因,多基因染色体中的每个基因都有相同的长度,分别表示为一棵表达式树。选择某个函数作为连接函数,就可以把这些基因连接成一棵表达式树。

2.2 GEP 中的算子

GEP 所采用算子中的选择算子和复制算子和遗传算法中的相应算子是一样的,但变异算子则有所不同,尽管它也可以作用在染色体的任意位置,可是作用在不同位置所受限制是不同的,如果变异发生在头部,它就可以改变为任意的函数或终结符,如果变异发生在尾部,就只能改变为终结符。GEP 中还提出了一些和遗传算法完全不同的算子,包括插串算子、根插串算子、基因重组算子等。

2.3 GEP 的基本步骤

GEP 算法的基本步骤是首先随机产生一定数量的染色体个体,对这些染色体进行表达,根据一定的适应度计算方法,得到每个个体的适应度值,并对个体进行选择、复制、变异等遗传操作,产生具有新特性的后代。这个过程将重复若干代,直至算法发现一个优良解。

和遗传算法一样, GEP 也是运用适应度函数来评价解的优劣性以及引导整个演化过程。目前最常用的是 Ferreira (2001)提出了两种适应度函数,如下所示:

$$f = \sum_{j=1}^m \left(M - |y_j - y'_j| \right) \quad (2)$$

$$f = \sum_{j=1}^m \left(M - \left| \frac{y_j - y'_j}{y_j} \right| \right) \quad (3)$$

在上面两式中, m 是训练集中包含的数据组数, y_j 为训练集中第 j 组数据的观测值, y'_j 为运用 GEP

所建的模型对于训练集中第 j 组数据的输入而得到的输出值, M 是一个常数。显然, (2)式是基于绝对误差进行评价, (3)式是基于相对误差进行评价。

2.4 GEP 的特点

GEP 结合了遗传算法和遗传编程的优点, 它采用等长线性符号编码来表示染色体, 遗传操作很简单, 但该编码又对应于非线性的相对复杂的树结构。这种将基因型(染色体)和表现型(表达式树)既分离又相互转化的结合, 使得它能够用简单编码来解决复杂的问题。正如 Ferreira (2002)所指出的, GEP 能够解决组合爆炸和早熟的问题, 而且比遗传算法和遗传编程要快 2~4 个数量级。

3 实证研究

3.1 研究目的

为了探讨 GEP 是否可以对自陈量表所提供的数据进行建模, 本研究以中学生的创造力个性特征和认知需求分数为例(这两种数据都是通过自陈量表获取的), 探讨是否可以用基因表达式编程对这些心理测量数据进行建模, 预测误差究竟有多大; 和多元线性回归、多项式回归等统计方法相比较有哪些优点; 和神经网络、支持向量回归机等其他智能建模方法相比较有哪些特色?

对于创造力的研究始终是心理学家感兴趣的课题, 现有的研究表明创造力的高低受到多种因素的影响, 人格就是其中的重要因素之一(Eysenck, 1993), 人格中的多种特质都和创造力是有密切关系的(Mayer & Mussweiler, 2011, Lin, Hsu, Chen, & Wang, 2012), 有些人格特质还直接影响到发散性思维(Lee & Dow, 2011)。创造力个性是人格的某种特征, 它体现了被试对新事物的探索愿望, 和人们的内部动机有密切关系, 而内部动机又会影响到创造力的高低(Prabhu, Sutton, & Sauser, 2008)。认知需求是个体对于复杂认知活动所具有的内部动机, 心理学家探讨了它和人们多种行为的关系, 例如, 与学生的思考努力程度的关系(Bertrams & Dickhauser, 2010)、与员工的工作绩效的关系(Reinhard & Dickhauser, 2009)、与消费者购买行为的关系(Kuo, Horng, Lin, & Lee, 2012), 显然, 认知需求是一种能够对人的行为产生重要影响的内部动机。

本研究假设对于新事物探索愿望越是强烈的被试, 他们处理复杂认知活动的内部动机也就越是强烈, 即创造力个性和认知需求之间是存在某种关

系的, 因此试图采用基因表达式编程的方法建立起两者关系的模型。

3.2 数据来源

对于创造力的测量方法是多种多样的(Kim, 2010), 其中很重要的一种是自陈量表法(Silvia, Wigert, & Palmon, 2012), 本研究采用台湾学者林幸台和王木荣(1997)等修订的威廉斯创造力测验, 以及台湾学者高泉丰(1994)修订的认知需求量表, 这两种测量工具都是属于自陈量表的范畴。在苏州市的 4 所中学里, 采用整群随机抽样的方法, 抽取初一、初二和高一、高二的学生 427 人进行测量(蔡玲琴, 2010)。

通过威廉斯创造力测验, 得到了被试在冒险性、好奇心、想象力、挑战性这 4 个维度上的创造力个性特征分数, 通过高泉丰(1994)的认知需求量表得到了被试的认知需求分数。

3.3 数据预处理

由于数据的采集过程受到多种无关因素的影响, 可能存在某些缺值或者异常点, 因此要首先进行数据清洗(Liao, Chu, & Hsiao, 2012)。本研究中对被试的创造力个性特征和认知需求都是采用问卷法进行的, 存在着共同方法偏差的风险, 因此还要对数据进行共同方法偏差的检验(Podsakoff et al., 2003)。

3.3.1 数据清洗

数据清洗的目的是使得数据干净和整齐, 以免对建模产生不利影响, 其主要工作包括对缺值的处理、噪声点(异常点)的清除等。

(1)缺值的处理

缺值是指某些被试在某些属性上缺失取值, 大多数算法对于缺值是无法进行处理的。常用的缺值处理方法包括直接删除法、新值法、统计填充法和预测估计法, 在一般情况下, 如果含有缺值的实例数相对于整个数据集的实例数比较少, 就可以采用直接删除法(胡可云, 田凤占, 黄厚宽, 2008)。在本研究中, 被试人数为 427 人, 存在缺值的被试为 27 人, 相对整体来说, 人数不算太多, 因此就采用的直接删除法, 进行缺值处理后得到有效被试 400 人。

(2)单维数据异常点的检测

对于单维数据的异常点, 通常采用基于均值和标准差的方法进行检测。根据拉依达准则, 在均值左右 3 个标准差区间外的数据可以看作是异常点, 应该将其删去(洪伟, 吴承祯, 2004)。令被试在冒险性、好奇心、想象力和挑战性这 4 个维度上的分数

分别为 X_1 、 X_2 、 X_3 和 X_4 , 认知需求的分数为 y 。分别计算这 5 个变量的平均数和标准差, 根据拉依达准则, 需要删去的被试有 8 个: 8, 64, 88, 98, 116, 135, 244 和 330, 保留下的 392 个被试的数据继续进行处理。

由于在删去部分被试以后, 各个变量的数据分布发生了变化, 因此, 继续根据拉依达准则, 删除了被试 25, 56, 79, 199, 233, 253 和 372, 保留下的 385 个被试的数据。根据同样的道理, 对保留下来的数据还要根据拉依达准则, 继续删除被试 182 和 248 的数据, 对于保留下来的 383 个被试的数据, 经拉依达准则分析没有发现新的异常数据, 因此它们被用于继续分析。

(3) 多维数据异常点的检测

虽然对于每个维度的异常点分别进行了上述的检测, 但对于多维数据而言, 某些数据即使在其各个维度内部不是异常点, 但综合起来仍然可能是异常点(胡可云等, 2008), 因此要进行多维数据异常点的检测。常用的检测方法有统计法、聚类法、距离法和密度法等, 本研究采用聚类法, 该方法的基本思想是, 如果所得到的聚类大小为 1 或者非常小, 那么就认为该聚类中的点为异常点。

本研究采用 k 均值聚类方法, 首先对数据进行归一化变换, 由于聚类的数目对于异常点的检测是有影响的, 较小的聚类数目会掩盖异常点, 较大的聚类数目则可能出现过多的异常点, 因此, 本研究取 k 等于 2~11 类, 进行了 10 次试验, 结果表明, 没有出现某个类别人数很少的情况, 即没有发现多维异常数据。

3.3.2 共同方法偏差检验

在本研究中, 创造力个性特征分数和认知需求分数都是采用自陈量表对于同一批被试获取的, 就有可能存在共同方法偏差。为了减少这种偏差, 本研究采用了程序控制和统计控制的方法(Podsakoff et al., 2003)。程序控制的措施包括对两组变量的施测实行时间上的分离; 保护受测者的匿名性, 降低其对评价的担心; 对量表中各项目的语句和用词进行润色完善, 尽量使其简单明了, 避免模糊性。在统计控制中, 采用哈曼单因素检验方法, 这是共同方法偏差检验中的常用方法(Aulakh & Gencurk, 2000)。该方法认为, 如果对量表的所有项目得分进行因素分析, 结果只是提取出一个因子, 或者一个因子就能解释大部分的方差, 则可能存在共同方法偏差。

在本研究中, 首先对所有数据进行 KMO 检验和 $Bartlett$ 球度检验, 得到 $KMO=0.83$, $Bartlett$ 值为 423.40, $p<0.001$, 因此, 该数据合适进行因素分析。分析后发现特征值大于 1 的公因子有 4 个, 如表 1 所示。

表 1 因素分析结果

因子	特征值	方差贡献率%	累积方差贡献率%
1	3.534	28.64	28.64
2	2.871	23.27	51.91
3	2.043	16.56	68.47
4	1.578	12.79	81.25

分析结果表明, 在上述的 4 个公因子中, 第一个因子只解释了方差的 28.64%, 由此可见, 并不存在共同方法偏差的问题。

3.3.3 训练集和测试集的划分

在对心理测量数据建模时, 通常将原始数据划分为训练集和测试集两部分, 训练集的功能是对模型进行建构, 测试集的功能是对模型的质量进行评估, 并把在测试集上的精度作为该模型的精度估计。

对于非时序数据集, 常用的划分方法有单一训练集和单一测试集法、交叉验证法和自助抽样法。交叉验证法是将原始数据集随机平均分成 k 部分, 每次使用 $k-1$ 部分作为训练集, 其余部分作为测试集, 重复进行 k 次建模, 最后总的误差是这 k 次运行误差的平均值。由于要进行 k 次建模和误差分析, 其计算代价相当高。自助抽样法是从原数据集中进行随机有放回的抽取与原数据集实例数相等的新数据集作为训练集, 其余部分作为测试集, 该方法要独立进行 k 次, k 的值是根据经验确定的, 一般认为 $k=200$ 是较为合理的, 因此这种方法的计算代价也是很高的。而且, 这两种方法都比较适用于较小的数据集, 本研究中的原始数据集较大, 就没有必要采用这两种方法了。而且采用单一训练集和单一测试集法的计算代价也比较小, 于是本研究采用了单一训练集和单一测试集法, 即训练集的被试占原始数据集的三分之二左右, 其余的被试组成测试集。至此, 就完成了对数据的预处理。

3.4 运用基因表达式编程(GEP)的建模

在运用 GEP 建模时, 涉及到多个参数的设置, 因此需要采用试验设计的方法来寻找优化的参数取值组合。由于每个参数都有多个值可供选择, 也

就是各个因素的水平数目较多,对于这种情况采用均匀设计是比较合适的(Song, Song, & Sun, 2012),具体步骤如下:

3.4.1 确定因素及其水平

参考关于 GEP 研究的文献(Özcan, 2012; Azamathulla, 2102),确定了以下 5 种因素,由于各个因素的取值范围都比较大,因此取 9 个水平。具体情况如下:

A: 头部长度,分别取 4、6、8、10、12、14、16、18、20。

B: 基因数目:分别取 3、4、5、6、7、8、9、10、11。

C: 适应度函数:分别取 *RESR*、*AESR*、*R-square*、*MSE*、*RMSE*、*MAE*、*RSE*、*RRSE*、*RAE*。

D: 染色体个数:分别取 20、30、40、50、60、70、80、90、100。

E: 变异概率 *Pm*:分别取 0.04、0.08、0.12、0.16、0.20、0.24、0.28、0.32、0.36。

模型中选择的函数包括:加法、减法、乘法、除法、平方根、指数、自然对数、正弦函数、余弦函数。

3.4.2 建立试验因素水平表

对于上述 5 个因素及其 9 个水平,建立试验因素水平表,如表 2 所示。

表 2 试验因素水平表

水平	头部长度 A	基因数目 B	适应度函数 C	染色体个数 D	变异概率 E
1	A1=4	B1=3	C1= <i>RESR</i>	D1=20	E1=0.04
2	A2=6	B2=4	C2= <i>AESR</i>	D2=30	E2=0.08
3	A3=8	B3=5	C3= <i>R-square</i>	D3=40	E3=0.12
4	A4=10	B4=6	C4= <i>MSE</i>	D4=50	E4=0.16
5	A5=12	B5=7	C5= <i>RMSE</i>	D5=60	E5=0.20
6	A6=14	B6=8	C6= <i>MAE</i>	D6=70	E6=0.24
7	A7=16	B7=9	C7= <i>RSE</i>	D7=80	E7=0.28
8	A8=18	B8=10	C8= <i>RRSE</i>	D8=90	E8=0.32
9	A9=20	B9=11	C9= <i>RAE</i>	D9=100	E9=0.36

3.4.3 选用均匀设计表和表头设计

这是一个 9 水平 5 因素的试验,因此选择 $U_9(9^5)$ 均匀设计表。该表最多可以安排 6 个因素,本研究将 5 个因素放在表的前面 5 列,由此建立了如下的试验方案表表 3。

其他参数设置为:进化代数:1000;倒位概率:0.1;IS 变换概率:0.1;RIS 变换概率:0.1;单点重

组概率:0.3;两点重组概率:0.3;基因重组概率:0.1;基因变化概率:0.1;连接函数:+。

表 3 试验方案表

方案	头部长度 A	基因数目 B	适应度函数 C	染色体个数 D	变异概率 E
1	1(4)	2(4)	4(<i>MSE</i>)	5(60)	7(0.28)
2	2(6)	4(6)	8(<i>RRSE</i>)	1(20)	5(0.20)
3	3(8)	6(8)	3(<i>R-square</i>)	6(70)	3(0.12)
4	4(10)	8(10)	7(<i>RSE</i>)	2(30)	1(0.04)
5	5(12)	1(3)	2(<i>AESR</i>)	7(80)	8(0.32)
6	6(14)	3(5)	6(<i>MAE</i>)	3(40)	6(0.24)
7	7(16)	5(7)	1(<i>RESR</i>)	8(90)	4(0.16)
8	8(18)	7(9)	5(<i>RMSE</i>)	4(50)	2(0.08)
9	9(20)	9(11)	9(<i>RAE</i>)	9(100)	9(0.36)

3.4.4 进行试验

由于 GEP 运行时染色体的初始数值是随机产生的,因此对于上述每一种试验条件进行了 10 次试验,记录下每次试验的“训练拟合度最大值”和“测试拟合度最大值”,计算 10 次试验“测试拟合度最大值”的平均数,发现在试验条件 5 的情况下,测试集数据和模型的拟合度最好,因此将该条件下得到的表达式树转化为数学模型。该条件下的拟合度最大值如表 4 所示。

表 4 试验条件 5 时的拟合度最大值

试验次数	训练拟合度最大值	测试拟合度最大值
1	995.518	989.772
2	994.916	988.533
3	995.674	990.613
4	995.332	990.157
5	996.060	990.373
6	994.740	989.640
7	994.937	989.321
8	995.502	990.418
9	995.122	989.891
10	995.865	990.351

3.4.5 建立数学模型

基因表达式编程的运行结果是表达式树,为了计算模型的精度,还需要将其转化为数学表达式,得到显性的数学模型。根据上述的运行结果,可以将 10 次运行所得到的 10 棵表达式树转化为 10 个数学模型。为了节省篇幅,在此只列出了其中“测试拟合度最大值”大于 990 的数学模型。

模型 3:

$$X_4 \left[\cos(X_2 - 4.11) + (\sqrt{X_2} - \sin(-8.984)) \right] + (X_1^2 X_2 - X_4) (X_2^2 (X_2 - X_4)) + \sin \left[\sqrt{\frac{\sin(\sqrt{\sqrt{1.236} X_1})}{5.056}} \right]$$

$$= X_4 \left[\cos(X_2 - 4.11) + (\sqrt{X_2} + 0.4266) \right] + (X_1^2 X_2 - X_4) (X_2^2 (X_2 - X_4)) + \sin \left[\sqrt{\frac{\sin(\sqrt{\sqrt{1.236} X_1})}{5.056}} \right]$$

模型 4:

$$\cos \left(\frac{\sin X_4}{-21.045} - X_1 \right) + \cos(4.028) + 0.034 \left(\frac{X_1}{X_4} \right) + \sin(X_4 X_1) = \cos \left(\frac{\sin X_4}{-21.045} - X_1 \right) - 0.6322 + 0.034 \left(\frac{X_1}{X_4} \right) + \sin(X_1 X_4)$$

模型 5:

$$\sin \left[\sin X_2 (X_4 - 0.423) - (1.335 + X_1) \right] X_4 + \cos \left[\cos(\sqrt{X_4} + (X_4 - X_1)(X_1 - X_2)) \right] + \left[X_4 - (X_2^2 - 0.412) 0.021 \right] X_4$$

模型 8:

$$\cos(\cos(\sin(X_1 X_2))) + \cos(\sqrt{\cos X_4}) + \cos(-2.295) = \cos(\cos(\sin(X_1 X_2))) + \cos(\sqrt{\cos X_4}) - 0.6625$$

模型 10:

$$\sin \left(\frac{4.388 X_1^2}{e^{3.476}} - \cos X_4 \right) + \cos \left(\frac{X_2 \sin X_3}{X_3} - X_1 \right) + \sin(\cos(4.997))$$

$$= \sin(0.1357 X_1^2 - \cos X_4) + \cos \left(\frac{X_2 \sin X_3}{X_3} - X_1 \right) + 0.2771$$

3.4.6 对于模型精度的误差分析

对于模型精度的分析通常是将测试集中自变量的数据代入模型, 计算出其估计值 y_2 , 然后运用下式计算其和实际值 y_1 的平均误差 AE ,

$$AE = \frac{1}{n} \sum abs(y_2 - y_1)$$

式中的 n 为测试集中实例的数目。

3.5 运用 BP 神经网络的建模

神经网络是模拟人脑功能的一种智能技术, 它也可以对测量数据进行建模, 人们常常把根据相同数据建立的 GEP 模型与神经网络模型进行比较 (Colbourn et al., 2011)。MATLAB 提供了运用神经网络对数据进行分析的平台, 在这个平台上, 可以通过 GUI 工具方便地建构神经网络, 并用实证数据对网络进行训练和建模, 所得到的模型具有数据拟合和预测等功能(朱凯, 王正林, 2010)。

本研究采用了 3 层 BP 神经网络, 网络的输入变量是创造力个性特征分数, 有 4 个维度, 因此网络的输入层包含 4 个神经元, 网络的输出是单维的认知需求分数, 即输出层的神经元的个数是 1, 根据 Kolmogorov 定理, 隐含层的神经元个数为

$2 \times 4 + 1 = 9$, 因此该 BP 网络的结构为 4-9-1。

在本研究的 383 个被试的中, 有 70%(即 268 个)被试的数据用于训练 BP 网络, 30%(即 115 个)被试的数据用于测试。对于经过训练的神经网络, 将测试集创造力个性特征分数作为输入, 网络的输出值就是认知需求分数的估计值。计算该估计值和实际值的平均误差, 得到 $AE=2.76\%$ 。

3.6 运用支持向量回归机(Support Vector Regression, SVR)的建模

支持向量机是一种机器学习新方法(Vapnik, 1995), 它以统计学理论为基础, 以结构风险最小化为原则。它借助于最优化方法, 非常成功地处理了回归问题(Farquard, Ravi, & Raju, 2010), 发展了支持向量回归机(SVR)的技术, 它在解决小样本、非线性及高维数据回归方面中表现出特别的优势, 在交通、金融、能源需求预测等领域得到了很好的应用(Castro-Neto, Jeong, & Jeong, 2009)。

支持向量回归机是将输入变量映射到一个高维特征空间, 根据统计学习理论, 将原来在低维空间的非线性模型转换为高维特征空间的线性模型, 对该模型中的需辨识数据进行处理, 由于其等价于

求解一个优化问题,该问题就可以转化为对偶问题的求解,因此在定义核函数、经过等价交换后就可以使原来的问题得到解决。

在本研究中,将创造力个性特征分数作为SVR的输入,将认知需求分数作为SVR的输出,利用训练集数据进行建模。在此过程中,确定了SVR的正规化参数和核函数参数。然后,运用测试集对该模型进行验证,计算该模型的输出值和实际值的平均误差,得到 $AE=2.31\%$ 。

3.7 运用多元线性回归方法的建模

采用和上述相同的数据运用多元线性回归方法进行建模,得到回归方程为:

$$y = 0.318 + 0.047X_1 + 0.156X_2 - 0.091X_3 + 0.079X_4$$

对方程进行显著性检验,得到 $F=4.032$, $p=0.003<0.01$,表明该回归方程是高度显著的。将测试集中创造力个性特征分数输入到该多元回归方程,其输出值就是认知需求分数的估计值。计算该估计值和实际值的平均误差,得到 $AE=3.86\%$ 。

3.8 运用多项式回归的建模

采用和上述相同的数据运用二次多项式回归方法进行建模,得到回归方程为:

$$\begin{aligned} y = & 0.405 - 0.755X_1 - 0.121X_2 - 0.054X_3 + 0.993X_4 \\ & - 0.459X_1^2 - 0.232X_2^2 - 0.453X_3^2 - 0.365X_4^2 \\ & + 1.209X_1X_2 + 1.477X_1X_3 - 0.303X_1X_4 \\ & - 0.301X_2X_3 - 0.057X_2X_4 - 0.482X_3X_4 \end{aligned}$$

对方程进行显著性检验,得到 $F=2.853$, $p=0.000<0.01$,表明该多项式回归方程是高度显著的。将测试集中创造力个性特征分数输入到上述多项式回归方程中,其输出值就是认知需求分数的估计值。计算该估计值和实际值的平均误差,得到 $AE=3.21\%$ 。

4 结果和讨论

4.1 结果

(1)对于基因表达式编程建立的10个模型,分别计算其平均误差 AE ,将其作为模型精度的指标,结果如表5所示。

10个模型的平均精度为1.32%。

(2)根据上述结果可以看出,模型10的精度最高,为1.28%。该模型的表达式树如图1所示。

图1中的 $d0$ 、 $d1$ 、 $d2$ 和 $d3$ 分别为变量 $X1$ 、 $X2$ 、 $X3$ 和 $X4$, $Sub-ET1$ 至 $Sub-ET3$ 表示3个染色体, $Sub-ET1$ 中的 $c4=3.476$, $c9=4.388$, $Sub-ET3$ 中的

$c4=4.997$ 。

(3)对于这10个模型各自涉及的自变量考察后,发现 $X1$ 和 $X4$ 对于因变量的预测是最重要的,但其他自变量也有一定的贡献,模型10由于包含了所有的4个自变量,因此精度最高。

表5 10个模型的精度

模型	$AE \times 100\%$
1	1.329
2	1.351
3	1.310
4	1.304
5	1.315
6	1.334
7	1.342
8	1.301
9	1.323
10	1.278

(4)本研 究 所 得 到 的 10 个 模 型, 即 使 表 现 形 式 不 同, 但 它 们 的 预 测 误 差 却 是 接 近 的。 这 说 明 采 用 GEP 建 模 的 方 法 是 很 稳 健 的。 而 且, 研 究 者 可 以 从 所 得 的 一 组 模 型 中 找 到 最 便 于 解 释 的 模 型, 如 果 实 在 没 有, 还 可 以 继 续 建 立 新 的 模 型, 或 者 通 过 改 变 选 用 的 函 数, 使 得 模 型 更 加 便 于 解 释。

(5)本研究中得到了各种建模方法预测精度 $AE\%$ 分别为: GEP, 1.28; BP神经网络, 2.76; SVM, 2.31; 多元线性回归, 3.86; 多项式回归, 3.21。由此可见, 基因表达式编程(GEP)所建模型的精度要高于其他方法所建的模型。

4.2 讨论

(1)本研究运用 GEP 方法建模时, 参考了前人的工作, 选用了四则运算、平方根、指数、自然对数、正弦函数、余弦函数等运算符, 但是否存在更优的运算符子集呢? 如何选择更合适的运算符, 是需要进一步探讨的问题。

(2)在本文的 GEP 参数设置中, 直接采用均匀设计结果中最好试验条件所对应的参数, 那么是否还有更好的参数设置呢? 也是可以继续研究的。

(3)笔者运用基因表达式编程在现有的参数设置条件下, 对创造力个性与认知需求间的关系进行了建模获得了成功, 但由于GEP建模时, 它的重要参数基因头部长度的 h 是根据具体问题确定的, 而 h 的取值范围又会影响到它所取水平的数量和各个水平的值, 因此对于不同的心理测量问题, 基因头部长度的及其

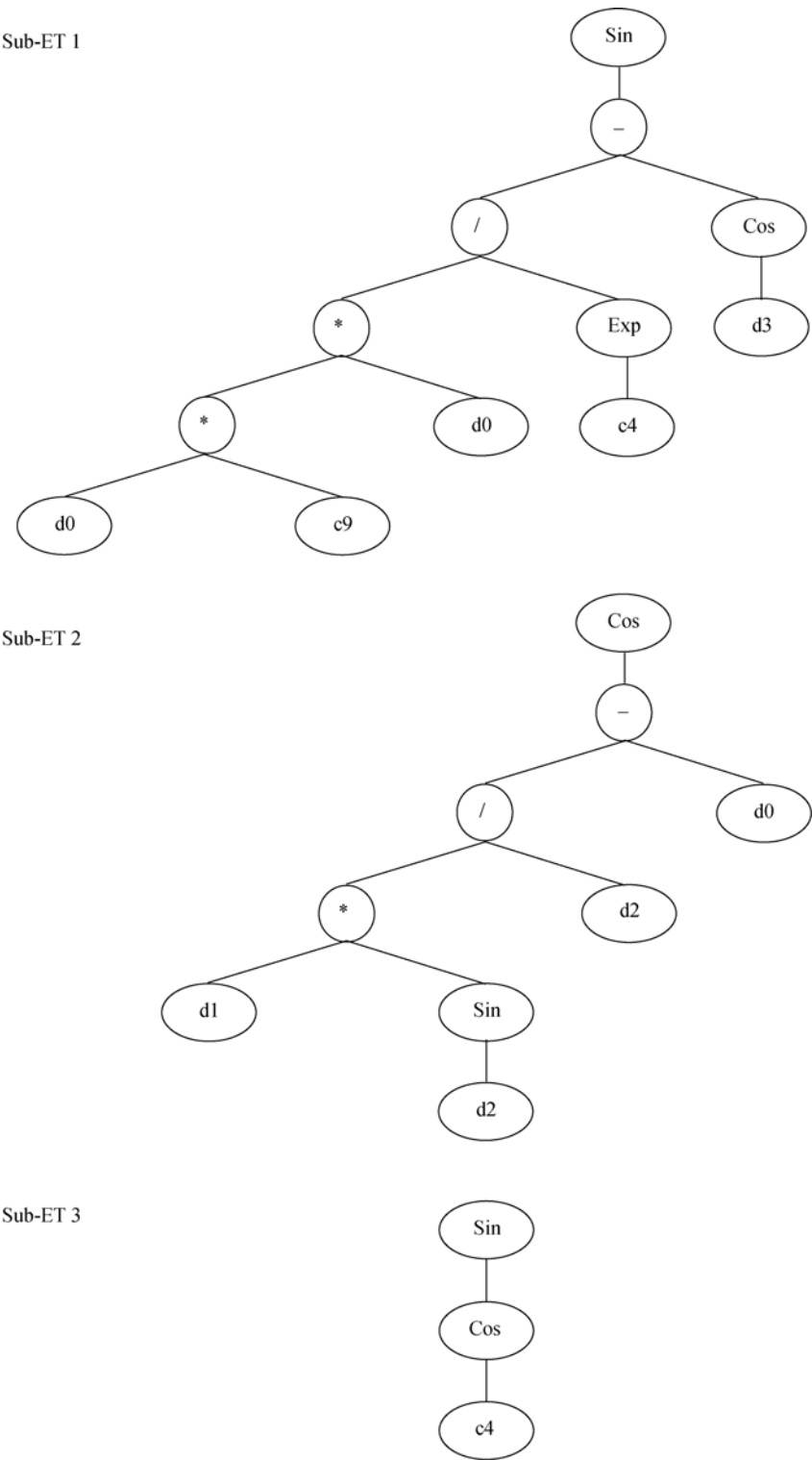


图 1 模型 10 的表达式树

相关的水平数都是要重新设置的, 由此也会引发对其他诸如基因数目、染色体个数等参数的重新设置。

(4)采用 GEP 建模方法可以获得显性的数学公式, 这对于揭示心理变量间关系的重要意义是显而易见的。但是, 我们是否可以将 GEP 随心所欲运用

于各种与心理量有关变量的建模呢? 回答是否定的。建立心理量数学模型的条件是首先要从心理学的基本原理去考察这些变量是否可能存在某种关系, 如果有可能存在某种关系, 但这种关系的具体形式尚不清楚, 则可以采用 GEP 方法寻找它们之

间的关系;但如果从心理学的基本原理来看,这些变量之间不可能存在某种关系,例如颅骨的尺寸和职业兴趣之间不可能存在某种关系,那么就不能用 GEP 来建模。其次要考虑的是,与物理量相比,心理量更容易受到各种无关因素的影响,测量的结果也不如物理量那样精确,因此要注意对数据的预处理,要事先对数据进行清洗,使其更加便于建模。最后对于模型的解释, GEP 只能建立起各个变量间关系的数学表达式,它的实质涵义还需要进行心理学的解释,包括通过实验研究来进行解释,例如,本文中关于创造力个性特征和认知需求的关系,就是要通过进一步的实验研究才能对所建模型进行充分解释。这些都说明了只有将心理学的基本原理和 GEP 紧密结合,才能够很好发挥所建数学模型的作用。

(5)对于同一组数据运用 GEP 建模可以得到了多个数学模型,这些模型在形式上有很大的差异,那么,在实际工作中应该如何在这些模型中进行选择呢?对于这一实际问题,首先要计算这些模型的预测误差,从中找出误差较小的若干个模型;然后在其中再找出比较容易解释的模型。如果已有的模型都很难进行解释,那就要从两个方面加以考虑:其一是在建模过程中改变目前所用的算子,其二是继续进行实证研究,从实验数据中寻找证据。这样做的理由是:首先希望这是个较为精确的模型,其次还希望这个模型是便于理解的。

(6)对于 BP 神经网络和支持向量回归机的建模,理论上也存在参数设置的优化问题,本研究只是采用了最常用参数配置,如果对它们的参数也采取多种方法进行优化,是否也能提高它们的建模精度呢?在神经网络的家族中,除了 BP 网络外还有其他的一些网络,它们是否会得到较高的建模精度呢?这些也是值得探讨的。

但不管如何,即使是它们的精度也达到了 GEP 建模的水平,它们也只能得到隐性的模型,而 GEP 所建的模型则是显性的,是可以采用明确数学公式来表达的,这是基因表达式编程建模的最大优点之一。

5 结论

(1)尽管通过自陈量表获取数据往往会受到主观因素的影响,但基因表达式编程还是能够对这些数据进行建模的,关键在于首先要对 GEP 的各种参数进行优化设置。

(2)与其他智能计算方法(神经网络、支持向量回归机等)和统计方法(多元线性回归、多项式回归等)相比较,在一定的条件下 GEP 所建模型的精度更高些。

(3)基因表达式编程所建立的模型是稳健的,虽然可以建立多个表现形式不同的模型,但是它们的预测精度是非常接近的。

参 考 文 献

- Aulakh, P. S., & Gencturk, E. F. (2000). International principal-agent relationships-control, governance and performance. *Industrial Marketing Management*, 29, 521-538.
- Azamathulla, H. M. (2012). Gene expression programming for prediction of scour depth downstream of sills. *Journal of Hydrology*, 460-461, 156-159.
- Bertrams, A., & Dickhauser, O. (2010). University and school students' motivation for effortful thinking. *European Journal of Psychological Assessment*, 26(4), 263-268.
- Bittar, E. G., Del-Claro, K., Bittar, L. G., & da Silva, M. C. P. (2012). Towards a mathematical model of within-session operant responding. *Journal of Experimental Psychology: Animal Behavior Processes*, 38(3), 292-302.
- Cai, L. Q. (2010). *The relationship of middle school students' need for cognition and creativity*. Unpublished master's thesis. Nanjing Normal University.
- [蔡玲琴. (2010). 中学生家庭环境, 认知需求与创造力的关系. 硕士学位论文, 南京师范大学.]
- Caselles, A., Micó, J. C., & Amigó, S. (2010). Cocaine addiction and personality: A mathematical model. *British Journal of Mathematical and Statistical Psychology*, 63, 449-480.
- Castro-Neto, M., Jeong, Y. S., & Jeong, M. K., & Han L. D. (2009). Online-SVR for short term traffic flow prediction under typical and atypical traffic conditions. *Expert Systems with Applications*, 36(3), 6164-6173.
- Colbourn, E. A., Roskilly, S. J., Rowe, R. C., & York, P. (2011). Modeling formulations using gene expression programming: A comparative analysis with artificial neural networks. *European Journal of Pharmaceutical Sciences*, 44, 366-374.
- Eysenck, H. J. (1993). Creativity and personality: An attempt to bridge divergent traditions. *Psychological Inquiry*, 4(3), 238-246.
- Farquad, M. A. H., Ravi, V., & Raju, S. B. (2010). Support vector regression based hybrid rule extraction methods for forecasting. *Expert Systems with Applications*, 37, 5577-5589.
- Ferreira, C. (2001). Gene expression programming: A new adaptive algorithm for solving problems. *Complex Systems*, 12(2), 87-129.
- Ferreira, C. (2002). *Gene expression programming*. Portugal: Angra do Heroismo.
- Gao, Q. F. (1994). The concept and measurement of need for cognition. *Bulletin of Chinese Psychology*, 36(1), 1-20.
- [高泉丰. (1994). 认知需求的概念与测量. 中华心理学期刊, 36(1), 1-20.]
- Hong, W., & Wu, C. Z. (2004). *Experimental design and analysis: Principle operating case*. Beijing, China: China

- Forestry Press.
- [洪伟, 吴承祯. (2004). *试验设计与分析——原理 操作 案例*. 北京: 中国林业出版社.]
- Hu, K. Y., Tian, F. Z., & Huang, H. K. (2008). *Data mining theory and application*. Beijing, China: Tsinghua University Press, Beijing Jiaotong University Press.
- [胡可云, 田凤占, 黄厚宽. (2008). *数据挖掘理论与应用*. 北京: 清华大学出版社, 北京交通大学出版社.]
- Kim K. H. (2010). Measurements, causes, and effects of creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 4(3), 131–135.
- Kuo, J., Horng, D., Lin, C., & Lee, S. (2012). The causal relationship between need for cognition and advertising recall. *Social Behavior and Personality*, 40(6), 1025–1032.
- Lakshmi, S., & Alamelu, S. (2010). A mathematical model predicts HPA axis response in late adulthood due to childhood separation. *Bio-Science Research Bulletin*, 26(2), 143–149.
- Lee, S. A., & Dow, G. T. (2011). Malevolent creativity: Does personality influence malicious divergent thinking? *Creativity Research Journal*, 23(2), 73–82.
- Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications: A decade review from 2000 to 2011. *Expert Systems with Applications*, 39, 11303–11311.
- Lin, W. L., Hsu, K. Y., Chen, H. C., & Wang, J. W. (2012). The relations of gender and personality traits on different creativities: A dual-process theory account. *Psychology of Aesthetics, Creativity, and the Arts*, 6(2), 112–123.
- Lin, X. T., & Wang, M. R. (1997). *The manual of Williams creativity thinking*. Taipei, China: Psychology Press.
- [林幸台, 王木荣. (1997). *威廉斯创造性思考活动手册*. 台北: 心理出版社.]
- Mayer, J., & Mussweiler, T. (2011). Suspicious spirits, flexible minds: When distrust enhances creativity. *Journal of Personality and Social Psychology*, 101(6), 1262–1277.
- Mollahasani, A., Alavi, A. H., & Gandomi, A. H. (2011). Empirical modeling of plate load test moduli of soil via gene expression programming. *Computers and Geotechnics*, 38, 281–286.
- Mousavi, S. M., Aminian, P., Gandomi, A. H., Alavi, A. H., & Bolandi H. (2012). A new predictive model for compressive strength of HPC using gene expression programming. *Advances in Engineering Software*, 45, 105–114.
- Özcan, F. (2012). Gene expression programming based formulations for splitting tensile strength of concrete. *Construction and Building Materials*, 26, 404–410.
- Peng, Y., Yuan, C., Mai, X., & Qin, X. (2011). Survey on theoretical research of gene expression programming. *Application Research of Computers*, 28(2), 413–420.
- [彭昱忠, 元昌安, 麦雄发, 覃晓. (2011). 基因表达式编程的理论研究综述. *计算机应用研究*, 28(2), 413–420.]
- Podsakoff, P. M., Mackenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879–903.
- Prabhu, V., Sutton, C., & Sauser, W. (2008). Creativity and certain personality traits: Understanding the mediating effect of intrinsic motivation. *Creativity Research Journal*, 20(1), 53–66.
- Reinhard, M., & Dickhauser, O. (2009). Need for cognition, task difficulty, and the formation of performance expectancies. *Journal of Personality and Social Psychology*, 96(5), 1062–1076.
- Silvia, P. J., Wigert, B., Palmon, R. R., & Kaufman, J. C. (2012). Assessing creativity with self-report scales: A review and empirical evaluation. *Psychology of Aesthetics, Creativity, and the Arts*, 6(1), 19–34.
- Song, J., Song, Z. G., & Sun, R. (2012). Study of uniform experiment design method applying to civil engineering. *Procedia Engineering*, 31, 739–745.
- Vapnik, V. N. (1995). *The nature of statistical learning theory (Information Science and Statistics)*. Berlin: Springer. Link.
- Zhao, J. & Dan, Q. (2003). *Mathematical modeling and mathematical experiments* (2nd Ed.). Beijing, China: High Education Press.
- [赵静, 但琦. (2003). *数学建模与数学实验* (第 2 版). 北京: 高等教育出版社.]
- Zhu, K., & Wang, Z. L. (2010). *Master MATLAB neural networks*. Beijing, China: Publishing House of Electronics Industry.
- [朱凯, 王正林. (2010). *精通 MATLAB 神经网络*. 北京: 电子工业出版社.]

Modeling Self-reported Instrument Data with Gene Expression Programming

QIAN Jinxin; YU Jiayuan

(School of Psychology, Nanjing Normal University, Nanjing 210097, China)

Abstract

It is often difficult to represent the complex relations among psychological variables with traditional analytical models like regressions. Supposedly, neural networks and support vector regression machine can be used instead. However, the limitation is that these models are recessive. Gene expression programming (GEP) can be used to handle these models with observable variables. At present, most of the data using GEP models are

obtained with objective methods. But a lot of the psychological measurement data are obtained from self-report instruments and are affected by many subjective factors. Could these kinds of data be used in GEP models? How large is the modeling error? Is there any advantage in using the GEP modeling as compared with the multivariate linear regression or the polynomial regression modeling? Is the GEP modeling more accurate than neural networks and support vector regression machine modeling? All the above issues would be explored in this paper.

The responses of 400 middle school students were obtained with the Williams creativity assessment packet and the need for cognition scale. A total of 17 students were deleted because of the abnormality in responses and the data from 383 students were retained for modeling. Common method biases had not been found with the Harman's single-factor test. Five parameters of gene expression programming were optimized with the uniform design. These parameters were head length, gene number, fitness function, chromosome number and mutation probability. There were nine levels for each parameter, each established under different testing conditions respectively. The condition with maximum fitness was obtained through experiments. The GEP program was repeated 10 times under this condition. The accuracy of the models was calculated and the model with the minimum error was found, of which the expression tree was drawn. The models of the relations between need for cognition and creativity personality traits were established using BP neural networks, support vector regression machine, multivariate linear regression and polynomial regression respectively. These models were compared with the model using gene expression programming.

The results showed that: (a) the accuracy of model 10, with four independent variables, was the highest; (b) the expressions of these ten models were different but their predictive errors were very close, thus supporting the robustness of the GEP modeling method; and (c) the predictive errors of different models were: GEP, 1.28; BP networks 2.76; support vector regression machine 2.31; polynomial regression 3.21; multivariate linear regression 3.86 respectively.

It can be concluded that: (a) data from self-reported instruments can still be modeled with gene expression programming even though these data are affected by many subjective factors; (b) the GEP modeling is more accurate than the other intelligent computing methods (neural networks, support vector regression machine, etc.) and traditional statistical methods (multivariate linear regression, polynomial regression, etc.), and (c) the models established with GEP are robust; their predictive accuracy is similar even though their mathematical formulae are quite different.

Key words self-reported instrument; gene expression programming; modeling; creativity; uniform design