

基于均匀设计与支持向量回归的发酵配方优化

袁哲明^{1,2}, 左斌³, 谭泗桥², 谭显胜², 熊兴耀¹

(1. 湖南农业大学作物种质创新与资源利用国家重点实验室培育基地, 湖南 长沙 410128;

2. 湖南农业大学生物安全科学技术学院, 湖南 长沙 410128; 3. 湖南农业大学生物科学技术学院, 湖南 长沙 410128)

摘要: 结合均匀设计与支持向量回归, 提出了一种新的配方优化实验设计与分析方法 UD-SVR. 将其应用于优化产谷氨酸脱羧酶大肠杆菌诱变株的培养基配方与发酵条件, 在考虑 9 因素时仅通过 2 轮 28 个实验, 酶活性(吸光度 OD₆₃₀)即由初始的 1.528 高效提升至 2.303, 明显优于二次多项式偏最小二乘回归等经验风险最小参比模型. UD-SVR 为多因素多水平配方优化实验设计与分析提供了一套预测精度高、指导性强、可解释性好、优化高效的整体解决方案.

关键词: 均匀设计; 支持向量回归; 配方优化

中图分类号: Q939.9

文献标识码: A

文章编号: 1009-606X(2009)01-0148-05

1 前言

发展新的实验设计与分析方法, 通过实施少量实验, 优化配方并解释各因素效应, 对发酵工程、动物营养及其他多因素多水平实验设计意义重大^[1]. 正交设计与均匀设计(Uniform Design, UD)能有效降低实验个数, 特别是均匀设计, 在实验范围内选择具有低偏差趋于均匀分布的好格子点集来安排试点, 可大幅度降低实验个数到允许范围^[1,2]. 均匀设计后续分析常采用的多元线性回归(Multiple Linear Regression, MLR)、逐步线性回归(Stepwise Linear Regression, SLR)、偏最小二乘回归(Partial Least-squares Regression, PLR)、二次多项式回归(Quadratic Polynomial Regression, QPR)、偏最小二乘二次多项式回归(PLS-QPR)、人工神经网络等经验风险最小回归模型假定样本充分大或无限大, 但配方优化是典型的小样本, 因此其泛化推广能力在小样本前提下存在严重弊端: 由经验风险最小回归模型推测的最优配方, 在验证中往往发现不满意配方; 在配方优化中各变量间常存在多重相关或样本点未充分大于变量的情形, 使 MLR 和 QPR 的应用受限; 即使采用 SLR, PLR 和 PLS-QPR, 由于各自变量与因变量间往往存在非线性关系, 这些传统回归模型的线性或拟线性本质使其在解释各因素效应的合理性时受到质疑^[3]. 以响应面方法优化配方时, 重构函数的选择是一个关键问题, 通常采用的二次多项式对非线性函数的逼近能力有限^[4]. 人工神经网络具有很好的非线性逼近能力, 但存在模型结构难以确定、可解释性差、易于出现过度训练和训练不足、陷入局部最小等诸多缺陷^[5].

支持向量机是机器学习领域的集大成者^[6,7], 最初用于语音识别等模式识别即支持向量分类, 现已扩展到回归预测即支持向量回归(Support Vector Regression, SVR)^[7]. 支持向量机基于结构风险最小, 较好地解决了小样本、非线性、过拟合、维数灾和局极小等问题, 泛化推广能力优异, 但可解释性差. 均匀设计用于快速搜索支持向量机最优参数已有研究^[8,9], 但均匀设计与 SVR 结合用于配方优化设计很少报道.

γ -氨基丁酸(Gamma Aminobutyric Acid, GABA)是有效的抑制性神经递质, 具有降血压、保持神经安定等作用, 在医药、食品工业中有广泛应用^[10]. 谷氨酸脱羧酶(Glutamate Decarboxylase, GAD)是酶法生产 GABA 的关键酶^[11]. 近年来以大肠杆菌(*Escherichia coli*)产 GAD 进而将谷氨酸转化为 GABA 已形成产业. 前期本课题组筛选到一株大肠杆菌诱变株, 其 GAD 产量较普通大肠杆菌有明显提高. 本研究结合均匀设计与 SVR, 以该诱变菌株产 GAD 的培养基配方与发酵条件优化为例, 提出了一种新的配方优化实验设计与分析方法 UD-SVR, 具有预测精度高、指导性强、可解释性好、优化高效等优点.

2 实验

2.1 材料

产 GAD 大肠杆菌诱变菌株由本实验室保存, 菌种斜面培养基为蛋白胨 10 g/L, 牛肉浸膏 5 g/L, NaCl 3 g/L, pH 6.6(雷滋便携式 pH 计测定), 基本发酵培养基见表 1, 所用试剂为市售生化试剂(分析纯).

收稿日期: 2008-08-22, 修回日期: 2008-11-18

基金项目: 国家自然科学基金资助项目(编号: 30570351); 教育部新世纪优秀人才计划基金资助项目(编号: NCET-06-0710); 高等学校博士点基金资助项目(编号: 200805370002)

作者简介: 袁哲明(1971-), 男, 湖南省岳阳市人, 博士, 教授, 主要从事模式识别与预测等研究, Tel: 0731-4635217, E-mail: zhmyuan@sina.com.

2.2 菌体培养与发酵

菌株接入斜面培养基 37 °C 培养 13 h, 挑取少许接入含 150 mL 基本发酵培养基的 250 mL 三角瓶中继续活化, 37 °C 下 140 r/min 振荡培养 13 h 后, 按 5% (ϕ) 接种量接种发酵培养。

2.3 酶活性测定

在 250 mL 三角瓶中加入 150 mL 培养液发酵。每个处理取样 3 次, 每次取 20 mL 发酵液用日立 CR22E 型离心机离心收集菌体, 加 2% L-谷氨酸 100 mL, 在 THZC 恒温振荡器中 37 °C 下 140 r/min 振荡转化 1 h。取 0.2 mL 转化液, 加 1 mL 蒸馏水、0.2 mol/L 硼酸缓冲液 (pH 9.0) 1 mL 终止转化, 以 Berthelot 显色法测定转化液中 GABA 含量^[12]。继续加入 6% 苯酚 2 mL、10% 次氯酸钠 0.8 mL 显色。用日立 UV 3000 型分光光度计测定反

应在 630 nm 处的光吸收值(OD₆₃₀), 以 3 次重复的 OD₆₃₀ 平均值表征酶活大小。

2.4 支持向量机 LIBSVM 2.8 软件包

LIBSVM 2.8 软件包简单易用, 含 4 个常用程序: Svmscale 用于对原始数据规格化, Svmtrain 用于训练, Svmpredict 用于预测, Gridregression.py 用于自动搜索核函数最优参数 c, g, p ($c \in [-1, 6], g \in [-8, 0], p \in [-8, -1]$, 步长均为 1)。各程序用法及其参数设置参见文献[13]。

3 UD-SVR 优化过程与分析

3.1 基本发酵培养基及各因素上下限

基本发酵培养基及发酵条件为生产厂家针对普通大肠杆菌的优化方案, 用于诱变菌时酶活(OD₆₃₀)为 1.528。以基本方案为中位值, 各因素预设上下限见表 1。

表 1 基本发酵培养基及各因素预设的上下限

Table 1 Base fermentation medium and the predetermination upper and lower limits of factors

Factor	Beef extract, x_1 (g/L)	Peptone, x_2 (g/L)	NaCl, x_3 (g/L)	Glu, x_4 (g/L)	Glucose, x_5 (g/L)	KH ₂ PO ₄ , x_6 (g/L)	MgSO ₄ , x_7 (g/L)	pH, x_8	Time, x_9 (h)	Activity, y (OD ₆₃₀)
Median	5	10	3	1	2	3	0.5	6.6	15	1.528
Upper limit	8	15	4	1.5	3	4	0.7	7.4	20	-
Lower limit	3	5	2	0.5	1	2	0.3	5.8	10	-

表 2 第一轮均匀设计试验及结果

Table 2 The first uniform design and the activity of GAD

Scheme	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
N_1	1.300	6.42	15.00	2.44	1.06	1.44	3.33	0.60	7.22	19.00
N_2	1.415	6.95	6.58	2.89	1.50	2.78	2.67	0.70	7.04	15.50
N_3	0.763	8.00	10.79	3.33	1.39	1.67	3.11	0.30	6.69	11.00
N_4	1.181	4.84	5.00	3.78	0.50	1.67	2.89	0.60	6.69	18.00
N_5	1.027	3.26	5.53	2.44	1.39	2.11	3.33	0.40	6.33	13.00
N_6	1.401	3.00	9.21	3.11	0.94	1.00	2.00	0.50	6.87	20.00
N_7	0.674	7.74	7.63	2.22	0.83	2.56	2.89	0.40	5.80	19.00
N_8	1.431	7.21	13.42	2.22	0.50	1.89	4.00	0.40	6.87	14.50
N_9	0.985	6.68	7.11	4.00	1.17	1.00	3.78	0.50	6.16	14.50
N_{10}	0.751	4.58	11.32	2.67	0.61	3.00	2.22	0.30	6.16	15.50
N_{11}	1.063	7.47	11.84	3.78	0.94	2.11	2.00	0.70	6.33	13.00
N_{12}	0.631	6.16	6.05	2.89	0.72	1.22	2.44	0.30	7.22	12.00
N_{13}	0.915	5.37	9.74	2.00	1.28	2.33	2.22	0.50	7.40	10.00
N_{14}	0.483	5.89	8.68	3.33	0.61	2.56	3.56	0.70	5.98	10.00
N_{15}	1.447	5.11	12.37	3.11	1.28	3.00	4.00	0.60	6.51	20.00
N_{16}	0.955	3.79	10.26	2.00	0.72	1.44	3.56	0.70	6.51	16.50
N_{17}	0.606	4.32	8.16	3.56	1.06	2.33	3.78	0.30	7.40	16.50
N_{18}	1.084	5.63	13.95	3.56	1.50	1.89	2.44	0.40	5.98	18.00
N_{19}	0.768	4.05	12.89	2.67	1.17	1.22	2.67	0.60	5.80	11.00
N_{20}	0.959	3.53	14.47	4.00	0.83	2.78	3.11	0.50	7.04	12.00

3.2 第一轮均匀设计

按 9 因子($x_1 \sim x_9$)20 水平进行均匀设计^[2], 部分因子的水平间差异太小进行归并, 为方便操作对 x_9 进行了微调。第一轮 20 次处理($N_1 \sim N_{20}$)及其实测酶活性值 y (OD₆₃₀)见表 2。

3.3 最优核函数选择与非线性变量筛选

SVR 的一个弊端是核函数的选择缺乏先验的理论指导, 是经验性的。常用的核函数包括线性核、多项式核、径向基核和 Sigmoid 核。袁哲明等^[14]、张永生等^[15]

给出了依训练集 n -fold 交叉测试(其极限是留一法)均方误差(Mean Squared Error, MSE)最小原则从 4 种常用核函数中自动选择最优核函数的算法与程序; 类似于 SLR, 还基于 SVR 给出了一种非线性变量筛选方法: 以多轮末尾淘汰法从包含全部输入变量的 SVR 模型中以留一法依 MSE 最小原则非线性逐次剔除对提高预测精度不利的变量^[14,16], 筛选后的变量称为保留变量。SVR 的另一个缺点是可解释性差。进一步用多轮末尾淘汰法对保留变量进行强制筛选, 可给出各保留变量对预测精

度影响的相对重要性次序,因而具部分解释能力^[14-17].
最优核函数选择和非线性变量筛选应同步进行.

对表2数据,最优核函数为径向基核($t=2$),保留变

量为 x_9, x_4, x_7, x_8 , 且其对酶活 y 预测精度影响的相对重要性为 $x_8 > x_7 > x_4 > x_9$ (表3). 在表1给定的上下限范围内, 变量 x_1, x_2, x_3, x_5, x_6 对酶活影响甚微.

表3 非线性变量筛选与保留变量强制筛选及其MSE值(径向基核)

Table 3 Screening variables and compulsorily screening retained variables and their MSE ($\times 100$ values)

Stage	Round	Before screening	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	Variables screened
Screening variables	1	7.5	6.0	6.8	6.9	8.3	6.4	5.9	8.6	8.4	8.9	x_6
	2	5.9	5.2	5.5	6.4	7.1	5.7	-	8.3	7.7	8.9	x_1
	3	5.2	-	5.0	5.5	6.2	4.4	-	7.7	6.3	8.3	x_5
	4	4.4	-	4.4	5.0	5.4	-	-	7.2	5.5	6.4	x_2
	5	4.4	-	-	1.9	3.7	-	-	6.8	5.2	4.6	x_3
Compulsorily screening retained variables	6	1.9	-	-	-	4.6	-	-	6.1	4.4	3.5	x_9
	7	3.5	-	-	-	2.9	-	-	8.3	5.3	-	x_4
	8	2.9	-	-	-	-	-	-	6.1	7.3	-	x_7
	9	6.1	-	-	-	-	-	-	-	-	-	x_8

3.4 模型评估

至此获得了一个含4个输入变量的SVR模型,其拟合误差MSE为0.019.现以双重留一法评估该模型的泛化推广能力^[14]:从径向基核函数($t=2$)及其4个保留变量出发,从20个样本中每次拿出1个作为独立测试样本,其余19个样本再次以留一法Gridregression.py搜索寻找最优SVR参数,以最优参数 c, g, p 对该19个样本Svmtrain后预测(Svmpredict)独立测试样本.双重留一法预测结果MSE为0.022,比拟合误差MSE仅略有增加,表明所建模型有较好的泛化推广能力;其对酶活大于1.3的5个样本有4个预测较好,仅 N_8 预测稍差(数据未显示).由于关注的是预测的酶活较高的那些处理,如果一个回归模型整体预测精度较高但高酶活样本预测偏差较大,显然不是所期望的.通过双重留一法预测进行评估,核函数为径向基核,保留变量为 x_9, x_4, x_7, x_8 的SVR模型可以信赖.

3.5 预测与频次统计寻优

以径向基核为最优核函数,以 x_9, x_4, x_7, x_8 为保留变量,以表2的20个样本训练建立SVR模型,对 x_4 (10水平), x_7 (5水平), x_8 (10水平), x_9 (10水平)共5000个全组合待测样本进行预测,取预测酶活 $\hat{y} > 1.3$ 的519个处理统计各因子频次*i*如表4.在5000个预测样本中,最优组合为 $x_4=1.5, x_7=0.6, x_8=6.9, x_9=20$,对应最大预测值 $\hat{y}=1.574$.结合最优组合和表4, x_4 (谷氨酸)应取1.5或1.5以上(谷氨酸诱导谷氨酸脱羧酶产生,是必需的), x_7 (硫酸镁)可固定为0.6, x_8 (pH值)波动较大,可在6.51~7.22间取值再研究,峰值是6.88; x_9 (时间)应取 ≥ 20 ,因过长的培养时间对其多轮转化不利,可固定为20,不再外推.

3.6 第二轮均匀设计

经第一轮均匀设计及分析,9个因子中7个已基本固定(x_1, x_2, x_3, x_5, x_6 等非保留变量取基本发酵培养基中

表4 预测酶活大于1.3的519个样本的频次统计寻优

Table 4 Frequency (*i*) statistics of each factor in the 519 samples with predicted enzyme activity over 1.3

x_4	<i>i</i>	x_7	<i>i</i>	x_8	<i>i</i>	x_9	<i>i</i>
0.50	34	0.3	4	5.80	0	10.0	21
0.61	23	0.4	30	5.98	1	11.0	21
0.72	18	0.5	127	6.16	27	12.0	21
0.83	16	0.6	193	6.33	49	13.0	23
0.94	17	0.7	165	6.51	77	14.5	26
1.06	21			6.69	98	15.5	32
1.17	32			6.88	103	16.5	40
1.28	59			7.04	92	18.0	55
1.39	124			7.22	56	19.0	105
1.50	175			7.40	16	20.0	175

的基准值,若考虑成本因素可取最低水平值; x_7, x_9 经频次统计寻优分别固定为0.6和20.现变动因子只有2个,各4水平: x_4 上限由原来的1.5外推至2.8,下限调整为1.3; x_8 下限为方便起见调整为6.6,上限为7.2.全组合16处理稍多,均匀设计仅4处理偏少.故选择2因子8水平进行均匀设计,再归并为4水平,共8处理.第二轮均匀设计及各处理酶活预测值、实测值见表5.

在第一轮均匀设计中,20个处理实测最高酶活为1.447;经SVR指导的第二轮均匀设计,8个处理中4个超过1.447,其中3个超过2.0,预测最高与实测最高组合吻合,整体预测成功率较高.对第一轮均匀设计实测数据经非线性变量筛选,从9个因子中筛选出4个保留变量(x_8, x_7, x_4, x_9),经频次统计寻优进一步固定了 x_7 和 x_9 .从表5看,筛选的保留变量 x_4 和 x_8 的重要性得到了验证.第二轮均匀设计最优组合为 N_{25} ,其中 x_4, x_8 分别为2.3和6.8;基于第一轮均匀设计数据,经SVR和频次统计寻优分析 x_4 应外推至1.5以上、 x_8 峰值为6.88得到了验证.UD-SVR优化配方与发酵条件为:牛肉膏5 g/L,蛋白胨10 g/L,NaCl 3 g/L,谷氨酸2.3 g/L,葡萄糖2 g/L, KH_2PO_4 3 g/L, MgSO_4 0.6 g/L, pH 6.8,发酵时间20 h,优化方案的实测酶活性为2.303.

表5 第二轮均匀设计及结果

Table 5 The second uniform design and the activity of GAD

Scheme	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	GAD activity (OD ₆₃₀)	
										Predicted value	Observed value
N_{21}	5	10	3.0	1.3	2	3	0.60	6.8	20	1.493	2.010
N_{22}	5	10	3.0	1.3	2	3	0.60	7.2	20	1.409	1.196
N_{23}	5	10	3.0	1.8	2	3	0.60	6.6	20	1.702	2.264
N_{24}	5	10	3.0	1.8	2	3	0.60	7.0	20	1.624	1.413
N_{25}	5	10	3.0	2.3	2	3	0.60	6.8	20	1.776	2.303
N_{26}	5	10	3.0	2.3	2	3	0.60	7.2	20	1.600	1.780
N_{27}	5	10	3.0	2.8	2	3	0.60	6.6	20	1.725	1.184
N_{28}	5	10	3.0	2.8	2	3	0.60	7.0	20	1.605	1.279
N_A	5	10	3.0	1.5	2	3	0.70	7.4	20	1.611	1.263
N_B	8	15	2.0	1.5	3	4	0.69	7.4	20	2.090	0.927
N_C	8	15	2.0	1.5	3	4	0.69	7.4	20	2.241	0.927
N_D	8	15	2.8	0.5	3	4	0.56	6.9	18	1.920	1.933

Note: N_{21} ~ N_{28} were schemes based on uniform design; N_A was SLR model predicted optimal scheme in all 5000 schemes. N_B , N_C and N_D were the optimal schemes predicted by the model of PLS-QPR with reciprocation and square, PLS-QPR with reciprocation only and PLS-QPR with square only, respectively^[2].

参比经验风险最小模型中, GAD 酶活的 SLR 模型为 $\hat{y} = -1.4028 + 0.1915x_4 + 0.7571x_7 + 0.1707x_8 + 0.0467x_9$ (调整相关系数 $R_a = 0.6222$, \hat{y} 为 OD₆₃₀ 值), PLS-QPR 考虑交互项和平方项模型、PLS-QPR 仅考虑交互项模型、PLS-QPR 仅考虑平方项模型均相当冗长, 且可依表 2 数据由 DPS 软件快捷获得^[2], 故未列出. 前 3 个参比模型实测值均明显低于推测值, 这与其 x_4 均为 1.5(偏低)、 x_8 均为 7.4(偏高)有关; PLS-QPR 仅考虑平方项模型推测值与实测值接近但酶活仅为 1.933, 这与其 x_4 仅为 0.5(明显偏低)有关. 总之, 在小样本前提下, 结构风险最小的 SVR 模型其泛化推广能力明显优于经验风险最小模型.

4 结论

对一个 9 因子的复杂配方寻优, UD-SVR 仅经过 2 轮均匀设计, 实施 28 个处理, 实测酶活性就从初始方案的 1.528 上升到 2.303. 优化后的产 GAD 大肠杆菌诱变菌株配方与发酵条件为: 牛肉膏 5 g/L, 蛋白胨 10 g/L, NaCl 3 g/L, 谷氨酸 2.3 g/L, 葡萄糖 2 g/L, KH₂PO₄ 3 g/L, MgSO₄ 0.6 g/L, pH 6.8, 发酵时间 20 h. UD-SVR 为解决多因素多水平的配方优化试验设计与分析提供了新的有效途径, 但其有效性仍需更多独立研究的支持.

参考文献:

- [1] 方开泰. 均匀设计—数论方法在试验设计的应用 [J]. 应用数学学报, 1980, 3(4): 363-372.
- [2] 唐君义, 冯明光. 实用统计分析及其 DPS 数据处理系统 [M]. 北京: 科学出版社, 2002. 147-170.
- [3] Rosipal R, Be P P, Trejo L J, et al. Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space [J]. Journal of Machine Learning Research, 2001, (2): 97-123.
- [4] Gupta S, Manohar C S. An Improved Response Surface Method for

the Determination of Failure Probability and Importance Measures [J]. Structural Safety, 2004, 26: 123-139.

- [5] Chakraborty K, Mehrotra K, Mohan C K, et al. Forecasting the Behavior of Multivariate Time Series Using Neural Networks [J]. Neural Networks, 1992, 5: 961-970.
- [6] Vapnik V N. The Nature of Statistical Learning Theory [M]. New York: Springer-Verlag Press, 1995. 5-78.
- [7] Liang Y C, Sun Y F. An Improved Method of Support Vector Machine and Its Applications to Financial Time Series Forecasting [J]. Progress in Natural Science, 2003, 13(9): 696-700.
- [8] Huang C M, Lee Y J, Lin D K J, et al. Model Selection for Support Vector Machines via Uniform Design [J]. Computational Statistics and Data Analysis, 2007, 52(1): 335-346.
- [9] 林毅, 蔡福营, 张光亚. 苏云金杆菌杀虫晶体蛋白活性预测的支持向量机模型 [J]. 生物工程学报, 2007, 23(1): 127-132.
- [10] Boehm S L, Ponomarev I, Blednov Y A, et al. From Gene to Behavior and Back Again: New Perspectives on GABA_A Receptor Subunit Selectivity of Alcohol Actions [J]. Adv. Pharmacol., 2006, 54: 171-203.
- [11] Baekkeskov S, Aanstoot H J, Christgau S, et al. Identification of the 64 K Autoantigen in Insulin-dependent Diabetes as the GABA-synthesizing Enzyme Glutamic Acid Decarboxylase [J]. Nature, 1990, 347(6289): 151-156.
- [12] 许建军, 江波, 许时婴. 比色法快速测定乳酸菌谷氨酸脱羧酶活力及其应用 [J]. 微生物学通报, 2004, 31(2): 66-71.
- [13] Chang C C, Lin C J. LIBSVM: A Library for Support Vector Machines [CP]. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, 2001-03-31.
- [14] 袁哲明, 熊洁仪, 张永生. 基于 SVR 和 K-近邻群的组合预测在 QSAR 中的应用 [J]. 分子科学学报, 2007, 23(3): 163-169.
- [15] 张永生, 袁哲明, 熊洁仪, 等. 基于 SVR 和 CAR 的多维时间序列分析及其在生态学中的应用 [J]. 生态学报, 2007, 27(6): 2419-2424.
- [16] 谭显胜, 袁哲明, 周铁军, 等. Multi-KNN-SVR 组合预测在含氟化合物 QSAR 研究中的应用 [J]. 高等学校化学学报, 2008, 29(1): 95-99.
- [17] 谭泗桥, 袁哲明, 柏连阳, 等. 基于支持向量机回归与 K-最近邻法的组合预测用于除草剂 QSAR 建模 [J]. 农药学报, 2007, 9(4): 324-329.

Experimental Design and Analysis of the Optimal Fermentation Medium Based on Uniform Design and Support Vector Regression

YUAN Zhe-ming^{1,2}, ZUO Bin³, TAN Si-qiao², TAN Xian-sheng², XIONG Xing-yao¹

(1. Pre-State Key Lab. Germplasm Innovation and Utilization of Crop, Hunan Agricultural University, Changsha, Hunan 410128, China;

2. College of Bio-safety Science and Technology, Hunan Agricultural University, Changsha, Hunan 410128, China;

3. College of Bioscience and Technology, Hunan Agricultural University, Changsha, Hunan 410128, China)

Abstract: Based on uniform design and support vector regression, UD-SVR, a novel experimental design and analysis method for the prescription optimization was proposed. It was applied to optimize the complicated fermentation medium including nine factors for a variant of *Escherichia coli*. The variant could produce glutamate decarboxylase which transformed glutamic acid into gamma aminobutyric acid *in vitro*. The optimization results of the medium by UD-SVR showed that OD₆₃₀, an activity index of glutamate decarboxylase, increased from 1.528 in the initial median to 2.303 in the optimal medium after testing of 28 schemes. UD-SVR is more efficient than the reference models and has the potential to be widely used for experimental design and analysis of the prescription optimization.

Key words: uniform design; support vector regression; prescription optimization