第 26 卷第 6 期　　四川理工学院学报(自然科学版)　　Vol. 26　No. 6
2013 年 12 月　　Journal of Sichuan University of Science & Engineering(Natural Science Edition)　　Dec. 2013
文章编号:1673-1549(2013)06-0051-05　　DOI:10.3969/j. issn. 1673-1549. 2013. 06. 011

PSO－SVR 在果酒生物活性物质预测中的应用

陈国超，成新文

(四川理工学院计算机学院，四川　自贡　643000)

摘　要:针对 BP 神经网络和遗传算法对果酒生物活性物质预测存在速度慢和精度低的缺点,建立了基于支持向量回归机(SVR)的果酒生物活性物质预测模型。鉴于支持向量机模型的精度和泛化能力很大程度取决于不敏感损失系数 ε、惩罚系数 C 和 RBF 核函数的宽度系数 γ 三个参数,模型采用粒子群算法对三个参数同时进行优化,实现了果酒生物活性物质的非线性预测。仿真结果表明:基于 PSO－SVR 算法的果酒生物活性物质预测模型性能优于所比较的 BP 神经网络模型和支持向量回归机模型,能有效提高果酒生物活性物质的预测精度和稳定性。

关键词:支持向量机;生物活性物质;预测模型

中图分类号:TP301　　**文献标志码**:A

引　言

对人类生命活动具有显著调节功能的生理活性成分称为生物活性物质。在柑橘、枸杞等植物中富含的生物活性物质有黄酮、多糖等[1-2]。其中,枸杞中的黄酮具有抗突变、抗菌、抗氧化、抗病毒、抗肿瘤和防治血管硬化、调节免疫、降血糖等功能[3-4]。果酒中生物活性物质的最终含量受酵母添加量、糖添加量、发酵温度及 pH 值等因素影响,建立果酒生物活性物质预测模型,正确预测活性物质的变化趋势,对于提高果酒的营养成份和保健价值具有重要的指导意义。

影响果酒生物活性物质的因素较多,具有非线性特征,采用线性回归法、响应面法等传统方法预测精度不高[5]。文献[6]采用人工神经网络对沙棘果酒发酵工艺进行优化,预测酒精度,但是速度慢,精度较低。文献[7]用遗传算法优化 BP 和 RBF 两种神经网络建立了苹果气味识别的分类预测模型,提高了预测准确率。近年来,支持向量机技术较好地解决了小样本、非线性等实际问题,表现出良好的预测性能,成为继神经网络研究之后新的研究热点[8-9]。

针对果酒生物活性物质预测的特点,本文以岗稔果酒发酵过程正交试验的测量数据为样本[10],将支持向量机方法应用于生物活性物质黄酮苷含量的预测。鉴于支持向量机的精度和泛化能力取决于不敏感损失系数 ε、惩罚系数 C 和 RBF 核函数的宽度系数 γ 三个参数,采用粒子群优化算法对支持向量机的三个重要参数进行优化选择,实现了果酒生物活性物质含量的非线性预测。

1　基于支持向量回归机的果酒生物活性物质预测模型

1.1　果酒生物活性物质模型

果酒发酵是一个复杂的生化反应过程,影响果酒发酵的因素有菌种、温度、酵母接种量、SO_2、自由基、pH 值、金属离子等[10]。采用常用的中温发酵法时,影响黄酮苷含量的主要因素有菌种、温度、酵母接种量和固化金属铝离子浓度。

优良的菌种是果酒发酵的基础条件,影响发酵速度的因素是酵母接种量,而影响菌种的生长以及酒精的产

收稿日期:2013-09-16

基金项目:酿酒生物技术及应用四川省重点实验室开放基金项目(NJ2011-09)

作者简介:陈国超(1969-),男,四川达县人,讲师,硕士,主要从事人工智能方面的研究,(E-mail)cgchwb@ suse. edu. cn

生速度的因素是温度,合理的 pH 值有助于抑制其他杂菌的污染和保证发酵过程的正常进行。例如,温度较低时,发酵慢、酵母不易衰老、发酵更彻底、生成的酒精浓度高,黄酮苷含量也高;温度较高时,发酵快、酵母衰老更快,从而使黄酮苷的含量反而低。

由于正交试验法具有均衡分散性和整齐可比性的优点,在发酵过程研究中应用较多。根据影响岗稔果酒的生物活性物质黄酮苷的四大因素,在正交试验因素水平表中,采用四因素三水平,见表 1。

表 1 果酒黄酮苷正交试验设计因素水平表

水平	菌种	酵母接种量	发酵温度	固化 Al^{3+} 浓度
1	安琪	0.10%	15℃	0
2	BM45	0.20%	20℃	1.00%
3	D254	0.30%	25℃	2.00%

靳桂敏等人经过正交试验发现,使用中温酵母发酵时,发酵岗稔果酒的优良菌种为 BM45,在接种量为 0.2%,经 3.0% 海藻酸铝固定化后,在 25℃ 发酵温度下,可获得较高黄酮苷含量的果酒[10]。

1.2 支持向量回归预测模型构建

生物活性物质含量预测的核心问题是预测模型的构建。1995 年,Vapnik 等人提出了支持向量机(Support Vector Machine,SVM)技术[8]。SVM 以统计学习理论(SLT)为基础,基于结构风险最小化,具有结构简单,推广能力高的特点。SVM 在解决小样本、非线性及高维模式识别问题中具有许多特有的优势,并可以应用到函数拟合等问题中,产生了支持向量回归算法(Support Vector Regression,SVR)。

支持向量回归模型常用于解决非线性预测问题,应用价值高。给定样本集:$\{x_i,y_i\}$,$(i=1,2,\cdots,n;x_i \in R^n, y_i \in R)$,$x_i$ 为输入矢量,y_i 为相应的输出值,支持向量回归模型可用下面的回归函数定义:

$$f(x) = \omega \cdot \varphi(x) + b \tag{1}$$

其中,ω 为权重系数因子,b 为偏差系数。引入非负松弛变量上限 ξ 和下限 ξ^*,最佳回归函数通过求以下函数的最小极值,得出参数 ω 和 b:

$$\varphi(\omega,\xi,\xi^*) = \frac{1}{2}|\omega|^2 + C\left(\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\xi_i^*\right)$$

$$\text{s.t:}\quad \omega \cdot \varpi(x) + b - y_i \leqslant \varepsilon + \xi_i^*$$

$$y_i - \omega \cdot \varpi(x) - b \leqslant \varepsilon + \xi_i$$

$$\xi_i,\xi_i^* \geqslant 0 \tag{2}$$

其中,ε 为损失系数;C 为惩罚参数。引入拉格朗日乘子 a_i 和 a_i^*,建立拉格朗日函数,并根据 KKT 条件求解,得到原约束问题的对偶问题为:

$$\min\left(\frac{1}{2}\sum_{i,j=1}^{n}(a_i - a_i^*)(a_j - a_j^*)K(x_i,y_i) + \varepsilon\sum_{i=1}^{n}(a_i + a_i^*) - \sum_{i=1}^{n}y_i(a_i^* - a_i)\right)$$

$$\text{s.t:}\quad \sum_{i=1}^{n}(a_i - a_i^*) = 0, C \geqslant a_i, a_i^* \geqslant 0 \tag{3}$$

其中,$K(x_i,y_i)$ 为核函数,当采用径向基核函数构造支持向量回归模型时,可得到回归函数为:

$$f(x) = \sum_{i=1}^{n}(a_i^* - a_i)K(x_i,x) + b \tag{4}$$

其中,在没有足够的先验知识时,常采用径向基核函数(Radial Basis Function,RBF)如下:

$$K(x_i,x_j) = \exp\left(-\frac{|x_i - x_j|^2}{2\gamma^2}\right) \tag{5}$$

其中,SVR 的不敏感损失系数 ε、惩罚系数 C 和 RBF 核函数的宽度系数 γ 三个系数的优化选择决定回归模型的学习精度和泛化能力。

1.3 支持向量回归机参数分析

采用 RBF 核时,需要事先确定三个参数:RBF 核函数的宽度系数 γ、惩罚系数 C 和 SVR 的不敏感损失系数 ε。对于三个参数的选择,大多以经验和试凑为基础,交叉验证法和留一法是较常用的方法。

不敏感损失系数 ε 控制回归函数对样本数据的不敏感区域宽度,影响支持向量的数目。如果不敏感损失系数 ε 的选择太小,参与回归的支持向量将增多,虽然可以提高回归精度,但可能导致过拟合,降低泛化能力;如果不敏感损失系数 ε 太大,支持向量数减少,可能造成欠拟合,使训练和预测误差变大[11]。

惩罚系数 C 反映了支持向量回归机对超出 ε 管道样本数据的惩罚程度。如果惩罚小,那么训练误差变大,所以不能选择太小的惩罚系数。相反选择过大的惩罚系数,虽然可以提高模型的稳定性,但将导致模型泛化能力的降低。因此,选择合理的惩罚系数有助于保持模型和泛化能力的稳定性。

RBF 核函数的宽度系数 γ 影响支持向量回归机的各支持向量间的相关程度。宽度系数不能选择太小,否则支持向量间的联系相应就比较小,学习较复杂;宽度系数的选择太大,支持向量间的影响过强,容易产生欠拟合现象,影响模型的精度,训练误差将变大。

因此,SVR 模型的精确度和泛化能力取决于 ε、C、γ 三个参数,彼此密切相关,需要综合考虑,作出合理的参数选择。可以将粒子群算法引入到 SVR 中,对 SVR 的不敏感损失系数 ε、惩罚系数 C 和 RBF 核函数的宽度系数 γ 三个参数进行优化选择。

1.4　粒子群算法优化支持向量回归参数

1995 年,美国社会心理学家 Kennedy 博士和电气工程师 Eberhart 博士提出了粒子群优化算法(PSO)。粒子群优化算法的基本思想是通过群体中个体之间的协作和信息共享来寻找最优解。

如果 $X_i = (x_{i1}, x_{i2}, \cdots, x_{iD})$ 为 D 维目标搜索空间中粒子 i 的当前位置; $V_i = (v_{i1}, v_{i2}, \cdots, v_{iD})$ 为第 i 个粒子的当前"飞行"速度; $p_{best} = (p_{i1}, p_{i2}, \cdots, p_{iD})$, $i = 1, 2, \cdots, N$ 为粒子 i 所经历过的具有最好适应值的位置(个体极值),整个粒子群迄今为止搜索到的最优位置(全局极值) $g_{best} = (p_{g1}, p_{g2}, \cdots, p_{gD})$。粒子更新自己的速度和位置的方法如下:

$$v_i(t+1) = wv_i(t) + c_1 r_1 [p_{best} - x_i(t)] + c_2 r_2 [g_{best} - x_i(t)] \quad (6)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (7)$$

其中, c_1、c_2 为加速常数, w 为惯性因子, r_1 和 r_2 为两两相互独立的[0,1]范围内变化的随机数。

采用粒子群算法优化 SVR 参数对(ε、C、γ)时,每个粒子的位置和速度由参数(ε、C、γ)决定。选择反映 SVR 回归性能的均方差作为粒子群算法的适应度函数:

$$fitness = \left(\frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i)]^2\right)^{\frac{1}{2}} \quad (8)$$

1.5　粒子群算法优化支持向量回归参数的基本流程

每个粒子由参数 ε、C、γ 组成,参数的取值范围设定为: $C = [1\ 1000]$, $\varepsilon = [0.001\ 0.1]$, $\gamma = [0.01\ 5.0]$,惯性因子 $w = 0.6$,加速常数 c_1、c_2 为 1.7 和 1.9,粒子群数量为 20,最大迭代次数为 200,采用粒子群算法优化 SVR 参数的具体步骤如下:

(1) 算法参数设置和粒子群初始化。设定粒子群规模 m、最大迭代次数 Iter_{max}、学习因子和惯性权重等参数值,初始化粒子群的位置和速度。

(2) 计算粒子的适应度值,产生粒子的个体最优值和全局最优值。设置粒子的当前位置为个体极值 p_{best},按照公式(8)计算粒子适应度从中选择适应度值最佳的粒子个体位置作为初始全局极值 g_{best}。

(3) 按照式(6)和式(7)更新粒子位置和粒子速度。

(4) 按照公式(8)计算每个粒子的适应度值。

(5) 将各个粒子的当前适应度值与个体极值 p_{best} 对应的适应值比较,如果优于个体极值那么更新 p_{best},否则原值保持不变。

(6) 比较更新后的每个粒子的 p_{best} 和 g_{best},若得出优值则更新 g_{best},否则原值保持不变。

(7) 检查终止条件,如果接近最大迭代次数 Iter_{max} 则迭代终止,否则返回步骤 3。

最后,根据以上步骤得出支持向量回归模型的参数,构建生物活性物质的预测模型。

2　仿真实验与分析

以岗稔果酒发酵过程的生物活性物质黄酮苷为例,根据中温发酵时,影响黄酮苷含量的主要因素为菌种、温度、酵母接种量和固化金属铝离子浓度等,采用正交试验的结果数据作为样本数据。

表 2 列出了文献[5]中,采用中温菌种发酵正交试验的实验方案及得到的结果数据共 16 例。作为 SVR 训练数据的样本共 11 例,随机选择 9 例样本数据作为 SVR 测试样本。由于黄酮苷各影响因子的量纲差异很大,在模型构建时需要对样本数据进行归一化处理,其计算方法如下:

$$y = (y_{max} - y_{min}) \times \frac{x - x_{min}}{x_{max} - x_{min}} + y_{min} \quad (9)$$

在输出预测结果时,需要进行反归一化,可由公式(9)反推而得到。

表 2　黄酮苷含量 PSO – SVR 预测模型样本数据

样品	菌种	酵母接种量	发酵温度	Al^{3+} 浓度	总黄酮含量(mg/L)
1	1	1	1	1	503.8
2	2	2	2	2	617.3
3	3	3	3	3	513.8
4	4	4	4	4	508.3
5	1	2	3	1	615.5
6	2	1	4	2	636.4
7	3	4	1	3	661.9
8	4	3	2	4	666.4
9	1	3	4	1	583.7
10	2	4	3	2	543.7
11	3	1	2	3	490.1
12	4	2	1	4	420.2
13	1	4	2	1	310.3
14	2	3	1	2	566.5
15	3	2	4	3	518.3
16	4	1	3	4	509.2

采用粒子群算法优化 SVR,得到的较好结果为:不敏感损失系数 ε 为 0.024、惩罚系数 C 为 48.87 和 RBF 核函数的宽度系数 γ 为 0.015。

按照式(9),对所有样本数据进行规一化处理,结果见表 3。

对 9 组相同的测试样本数据,将采用 PSO – SVR 模型的生物活性物质黄酮苷含量的预测结果,与采用 BP 神经网络及支持向量机预测模型(交叉验证法)所得到

的实验结果进行比较,结果见表4。

表3 黄酮苷含量 PSO – SVR 模型样本数据归一化结果

No.	菌种	酵母接种量	发酵温度	Al^{3+}浓度	总黄酮含量
1	1.000	1.333	1.333	1.333	1.862
2	1.000	2.000	2.000	2.000	1.556
3	1.333	2.000	1.667	1.333	2.000
4	1.667	1.333	2.000	1.667	1.655
5	1.667	2.000	1.333	1.000	1.309
6	2.000	1.333	1.667	1.000	1.719
8	2.000	2.000	1.000	1.667	1.559
9	1.000	1.000	1.000	1.000	1.543
11	1.000	1.667	1.667	1.667	1.571
12	1.333	1.000	1.333	1.667	1.857
13	1.333	1.667	2.000	1.000	1.987
15	1.667	1.000	1.667	2.000	1.768
16	1.667	1.667	1.000	1.333	1.505

表4 黄酮苷含量三种预测模型的结果比较

NO	实际含量	BP 模型预测		SVR 预测		PSO – SVR 预测	
		预测含量	相对误差(%)	预测含量	相对误差(%)	预测含量	相对误差(%)
1	503.8	505.7	0.37	507.4	0.72	506.1	0.45
2	513.8	513.7	0.02	517.5	0.72	517.2	0.66
3	615.5	615.1	0.07	612.0	0.57	612.0	0.57
4	661.9	661.7	0.03	658.4	0.52	660.2	0.26
5	583.7	514.0	11.94	580.9	0.48	582.5	0.21
6	490.1	431.8	11.90	488.9	0.24	489.0	0.23
7	310.3	585.6	88.72	355.5	14.57	334.0	7.63
8	518.3	586.3	13.13	520.3	0.39	518.2	0.02
9	636.4	660.1	3.72	625.0	1.79	627.7	1.37

采用 PSO – SVR 算法的果酒黄酮苷含量实测数据与预测数据对比,如图1所示。

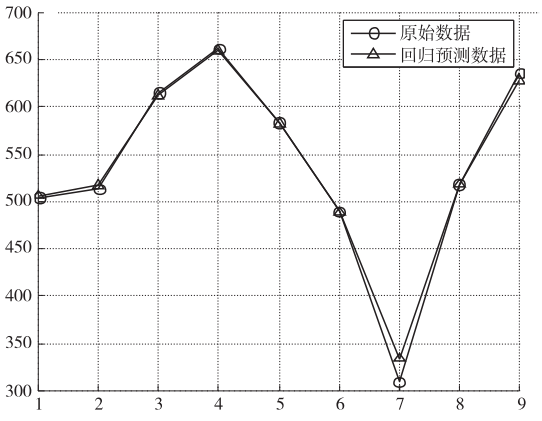

图1 黄酮苷实测数据与 PSO – SVR 预测数据对比

采用 PSO – SVR 算法的岗稔黄酮苷含量实测数据与预测数据的相对误差,如图2所示。

BP 神经网络预测模型的最大相对误差为 88.72%,最小相对误差为0.02%,平均相对误差为14.43%;采用

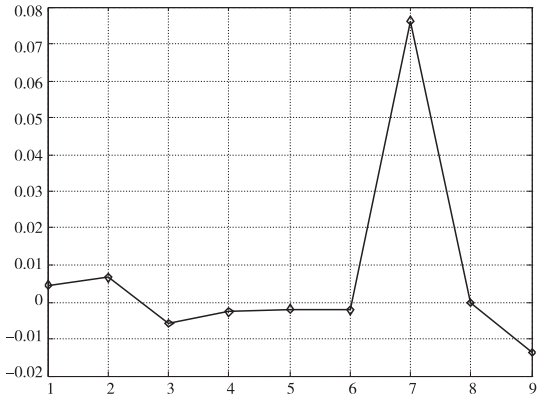

图2 黄酮苷含量 PSO – SVR 预测相对误差

交叉验证法的 SVR 预测模型,最大相对误差为14.57%,最小相对误差为0.24%,平均相对误差为2.22%;采用 PSO – SVR 预测模型,最大相对误差为3.85%,最小相对误差为0.22%,平均相对误差为1.27%。

分析三种预测模型相对误差的波动范围。其中,BP 预测模型波动范围最大,最大相对误差是最小相对误差的4436倍;其次是采用交叉验证法的 SVR 预测模型,基于 PSO – SVR 的预测模型波动范围最小。可见,基于 PSO – SVR 预测模型的预测精度更高、更稳定。

3 结束语

本文对影响果酒黄酮苷活性物质含量的四种因素进行正交试验,得出果酒总黄酮含量数据,构建了 PSO – SVR 预测模型,实现了对果酒生物活性物质的预测。通过对 BP 神经网络、SVR 和 PSO – SVR 三种模型比较可见,基于 PSO – SVR 的果酒总黄酮含量预测模型采用粒子群优化算法对支持向量机的重要参数进行优化,提高了果酒生物活性物质含量的预测精度和稳定性,具有较好的应用前景。

参 考 文 献:

[1] 张 玉,吴慧明,王 伟,等.不同品种柑橘果皮中类黄酮含量及其采后变化[J].食品科学,2010(6):202-204.

[2] 马虎飞,王思敏,杨章民.陕北野生枸杞多糖的体外抗氧化活性[J].食品科学,2011(3):60-63.

[3] 李淑珍,李 进.黑果枸杞总黄酮降血脂作用[J].时珍国医国药,2012(5):60-63.

[4] 叶兴乾,徐贵华,方忠祥,等.柑橘属类黄酮及其生理活性[J].中国食品学报,2008(5):1-7.

[5] 赵 爽,殷贝贝,刘 宇,等.响应面法优化毛木耳中总黄

酮提取工艺参数[J].食品工业,2012(2):1-4.

[6] 刘晓娜,韩建春,魏 婧,等.人工神经网络优化沙棘果酒主发酵工艺研究[J].中国酿造,2011(3):102-105.

[7] 赵杰文,刘木华,潘胤飞,等.遗传算法优化 BP 及 RBF 神经网络用于苹果气味分类[J].食品科学,2004,25(2): 39-42.

[8] Vapnik V.The nature of statistical learning theory[M]. Berlin:Springer Verlag,1995.

[9] 朱向荣,单 杨,李高阳.基于最小二乘支持向量机的国公酒中橙皮苷含量测定[J].光谱学与光谱分析,2009(9):2471-2474.

[10] 靳桂敏,林朝朋,钟瑞敏.岗稔黄酮苷在果酒发酵过程中稳定性研究[J].食品科技,2006(5):91-94.

[11] 成新文,陈国超,潘训海.果酒生物活性物质的 OED-LMBPNN 预测模型[J].电脑知识与技术,2012(32):7802-7806.

Prediction Model of Bioactive Substances From Wine Based on PSO-SVR

CHEN Guo-chao, *CHENG Xin-wen*

(School of Computer Science, Sichuan University of Science & Engineering, Zigong 643000, China)

Abstract: To solve the predict presence of slow speed and low precision of the bioactive substances from wine by using BP neural network and Genetic Algorithm, a prediction model based on Support Vector Regression (SVR) has been designed for metal corrosion rate. It is well known that the model complexity and generalization performance of this Support Vector Regression model depend on setting of the three parameters (ε, C, γ). Using the algorithm called Particle Swarm Optimization (PSO) to optimize the three parameters at the same time, nonlinear predictive of the bioactive substances from wine was achieved. Simulation results show that the proposed model is superior to the other two models for improving the forecast accuracy and stability.

Key words: Support Vector Machine; bioactive substances; prediction model