预测超临界 CO_2 中甘油三酯溶解度的新模型研究

于 海[1,2]，陈可可[1,2]，滕桂平[1,2]，余德顺[1,2]
(1. 贵州大学 化学与化工学院，贵州 贵阳 550025；
2. 中国科学院地球化学研究所 环境地球化学国家重点实验室 超临界流体研究中心，贵州 贵阳 550081)

摘要： 用一个新模型预测甘油三酯(TAGs)在超临界 CO_2(SC-CO_2)中溶解度，模型采用基于遗传算法(GA)优化的支持向量机(SVM)建立，同时采用应用域检查(AD)对模型的鲁棒性进行改善。通过GA优化，从大量的分子描述符中选出了与溶解度性质相关性最高的5个描述符；训练集和测试集拟合的相关系数(R)分别为0.986和0.982，相应的均方根差(*RMSE*)是14.10%和19.30%；在删掉异常点后，训练集和测试集的 R 值分别提高为0.992和0.984，对应的 *RMSE* 降低，分别为10.70%和11.70%。由结果可知：研究中建立的GA-SVM新模型提供了一个有效的方法预测甘油三酯在SC-CO_2中的溶解度，可为设计超临界 CO_2 萃取过程参数提供理论指导。

关键词： 超临界 CO_2；溶解度模型；遗传算法；支持向量机；甘油三酯
中图分类号： TQ 641 **文献标识码：** A **文章编号：** 1005-9954(2018)12-0037-05
DOI： 10.3969/j.issn.1005-9954.2018.12.008

Study on prediction for solubility of triglycerides in supercritical CO_2 using a new model

YU Hai[1,2], CHEN Ke-ke[1,2], TENG Gui-ping[1,2], YU De-shun[1,2]
(1. College of Chemistry and Chemical Engineering, Guizhou University, Guiyang 550025, Guizhou Province, China; 2. Supercritical Fluids Research Center, State Key Laboratory of Environmental Geochemistry, Institute of Geochemistry, Chinese Academy of Sciences, Guiyang 550081, Guizhou Province, China)

Abstract: A new support vector machine (SVM) model based on the genetic algorithm (GA) was developed to predict solubility of triglycerides (TAGs) in supercritical CO_2 (SC-CO_2). Besides, the application domain check (AD) was used to improve the robustness of the model. Five descriptors highly related to solubility are selected from a large number of molecular descriptors through the optimization of the GA. The correlation coefficients (R) for the training set and the test set are 0.986 and 0.982, and the corresponding root mean square errors (*RMSE*) are 14.10% and 19.30%, respectively. After removing the outliers, values of R for the modified training set and the modified test set increase from 0.986 to 0.992 and from 0.982 to 0.984, values of the corresponding *RMSE* decrease from 14.10% to 10.70% and from 19.30% to 11.70%. The results show that the developed model provides an effective method for predicting solubility of TAGs in SC-CO_2 and theoretical guides for the process design in the SC-CO_2 extraction.

Key words: supercritical CO_2; solubility model; genetic algorithm; support vector machine; triglycerides

SC-CO_2 技术已经是一种被广泛研究和应用的绿色技术。在植物油的SC-CO_2萃取过程中，充分利用该技术的前提就是要了解SC-CO_2和植物油的相平衡关系[1]。植物油是多种纯TAGs的混合物，因此获取高质量的纯TAGs在SC-CO_2中溶解度数据对理解其工业上应用显得尤为重要[2]。但在SC-CO_2中进行溶解度实验测定的时间长，成本高，所以人们开始用模型来预测TAGs在SC-CO_2中的溶解度。

收稿日期：2018-04-06
作者简介：于海(1992—)，男，硕士研究生，研究方向为精细化工分离技术，电话：17678917967，E-mail：2212752768@qq.com；余德顺(1963—)，男，硕士，研究员，通信联系人，电话：13608582488，E-mail：yudeshun@vip.skleg.cn。

迄今为止，溶解度模型主要分为以下 4 类: 理论模型、半经验模型，智能模型和 SVM。预测效果好的理论模型是状态方程[3]，但该模型参数多。另一类模型是 Chrastil 根据缔合理论提出半经验模型，能够仅用溶剂密度和温度 2 个变量关联溶解度，计算简便准确[4-5]，但此类方程受到缔合理论的影响，对压力和浓度有限制。神经网络模型可以很好地解决这些问题，而且预测准确[6-7]，但神经网络模型普遍存在局部最优、泛化能力差、网络结构参数选择难等问题。因此，理论性强、泛化能力更好的 SVM 模型得到了广泛的使用，在此过程中由于计算得到的大量描述符中，会存在数据冗余现象和与溶解度相关性小的描述符，所以在 SVM 建模过程中，通常会采用一些描述符选择方法，删除重复的和相关性小的描述符。常用的描述符选择方法有多元线性回归法(MLR) 和遗传算法(GA) [8-9]。本文采用基于 GA 优化的 SVM 模型预测纯 TAGs 在 SC-CO_2 中的溶解度。

1 支持向量机(SVM) 理论[8-9]

SVM 已经广泛被用于数据挖掘和变量回归问题; 基于统计学习理论和结构风险最小化，同时通过一个非线性的映射和一个核函数，将低维的非线性回归转化为高维的线性回归问题。

支持向量机中的输入数据集的一般表达式如下:

$$\{(x_i,y_i),i=1,2,\cdots,N\}$$

其基本回归函数

$$f(x)=\langle\omega,x\rangle+b \tag{1}$$

SVM 回归中所用的结构风险函数 $R(f)$ 为

$$R(f)=\frac{1}{N}\sum_{i=1}^{N}L(f(x_i)-y_i)+\frac{1}{2}\|\omega\|^2 \tag{2}$$

式(2) 中的损失函数为

$$L(f(x)-y)=\begin{cases}|L(f(x)-y)-\varepsilon|,|f(x)-y|\geqslant\varepsilon\\0,|f(x)-y|<\varepsilon\end{cases} \tag{3}$$

将方程(3) 代入方程(2) ，并引入松弛变量 ξ 和 ξ^* ，得到目标函数:

$$\text{minimize}\quad\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{N}(\xi+\xi^*)$$

$$S.t.\begin{cases}y_i-\langle w,x_i\rangle-b\leqslant\varepsilon+\xi_i\\\langle w,x_i\rangle+b-y_i\leqslant\varepsilon+\xi_i^*\\\xi_i,\xi_i^*\geqslant0\\i=1,2,\cdots,N\end{cases} \tag{4}$$

将目标函数代入到拉格朗日(Lagrange) 函数，并且执行一个对偶操作，使得下列目标函数最大化:

$$\text{maximize}\left[-\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)(x_i,x_j)-\varepsilon\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)+\sum_{i=1}^{N}y_i(\alpha_i-\alpha_i^*)\right]$$

$$S.t.\begin{cases}\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)=0\\0\leqslant\alpha_i\leqslant C\\0\leqslant\alpha_i^*\leqslant C\\i,j=1,2,\cdots,L,N\end{cases} \tag{5}$$

联立方程(6) 和(7) 得到 SVM 回归模型

$$f(x)=\sum_{i}^{N}(\alpha_i-\alpha_i^*)\langle x_i,x\rangle+b \tag{6}$$

对于高维特征空间 2 个向量的内积问题，通常用非线性条件下的高斯径向基核函数(RBF) 进行求解，RBF 用方程(9) 表示如下:

$$K(x_i,x_j)=\exp(-\gamma\|x_i,x_j\|),(\gamma>0) \tag{7}$$

最后，SVM 回归模型，用方程(10) 表示:

$$f(x)=(\alpha_i-\alpha_i^*)K(x_i,x)+b \tag{8}$$

2 模型建立

2.1 数据的收集

数据集是由文献中压力 p 和温度 T 范围分别在 8—50 MPa 和 308—353 K，9 个 TAGs 的 171 个溶解度数据组成[10-16]。其中，9 个甘油三酯分别是三丁酸甘油酯、三己酸甘油酯、三辛酸甘油酯、三癸酸甘油酯、三月桂酸甘油酯、三豆蔻酸甘油酯、三棕榈酸甘油酯、三油酸甘油酯和三硬脂酸甘油酯(纯度为摩尔分数≥99%) 。为了方便计算和观察，将溶解度值 S(g/g) 转化为其对数形式(log S) 用于模型的建立。由于 TAGs 在 SC-CO_2 中的溶解度很小，因此可以认为混合物密度是相同条件下纯 CO_2 的密度，纯 CO_2 密度从美国国家标准与技术研究院(NIST) 数据库中得到。然后随机选取数据集中 75% 的实验溶解度数据建立模型，剩下 25% 的实验溶解度数据用来验证模型。

2.2 分子描述符计算和预处理

ChemDraw Ultra11. 0 和 Hyperchem8. 0. 7 软件分别绘出 TAGs 分子的 2D 结构和 3D 结构，通过 AM1 半经验法优化，得到 TAGs 稳定的空间结构; 然后利用 Hyperchem 8. 0. 7 和 PaDEL2. 20 软件计算分子描述符。

将常数和几乎是常数的描述符从数据矩阵中删除，计算剩下的 1 247 个描述符的协方差。任意两两描述符之间皮尔逊系数大于 0. 9 的描述符对，其

中的一个被删除，以保证输入变量是线性无关的。另外，SC-CO_2 的溶解度受到温度和压力变化很大，因此温度和压力是两个和溶解度相关的实验描述符。最后留下 63 个描述符用于 GA 优化。

2.3　模型参数优化和最优描述符选择

本工作采用 RBF 核函数，训练时采用 ε-SVM 函数模型。利用 Matlab2017b 软件和 LIBSVM 工具箱进行扩展编程，选用 GA 作为模型的优化算法，建立 GA-SVM 模型。GA-SVM 的一般步骤如图 1 所示。

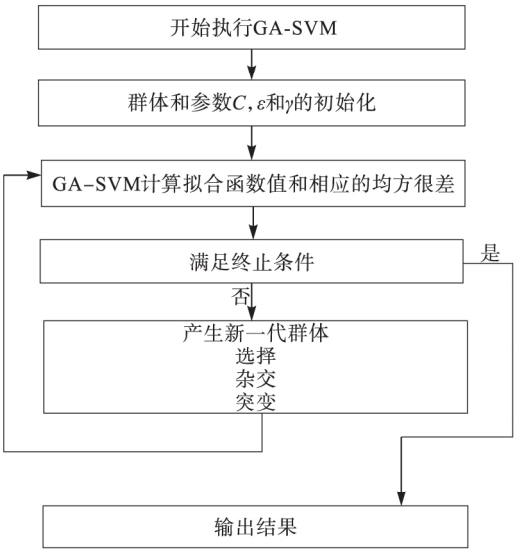

图 1　遗传算法优化的支持向量机模型的一般流程

Fig. 1　General flow chart of GA-SVM model

当遗传代数、拟合函数值、*RMSE* 和 *R* 分别是 1611，0.133，0.189 和 0.977 时，达到了 GA 的终止条件。经过 GA 优化的 3 个模型参数 C，ε，γ 值分别是 721.551，0.368 和 0.1，并且选择的 5 个分子描述符分别是温度、压力、VED3，TDB8m 和 PPSA-3。

VED3 描述符代表特征向量的系数被用作局部垂直方向上的不变量(LOVIs)，而 LOVIs 能够对不同的图谱顶点的差异性进行区别，较高的值对应较低度数的顶点，它们离中心或者高度数的垂直方向十分远，基于这些 LOVIs 的总和，提出 VED 指数作为分子描述符。TDB8m 是一个基于距离的 3D 拓扑描述符。它包含了在分子中，原子空间(S_k)、电子(X_k)和原子种类(I_k)的信息。PPSA-3 表示原子电荷权重 PPSA，它是一个 3D 带电部分表面积描述符，它是通过将分子表面积和部分原子电荷信息结合起来得到一种新型描述符，它能够通过更好地描述分子间的相互作用，来改善一些物理化学性质的实验值，而这些分子间的相互作用最初是由分子自身的极性引起的[17-18]。

3　模型训练和预测分析结果

随机地选取 75% 的溶解度数据作为训练集建立模型，剩下的 25% 用作测试集验证模型。训练集和测试集的拟合结果如图 2 所示。

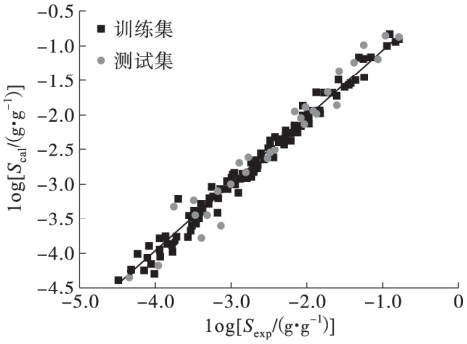

图 2　训练集和测试集各自溶解度实验值和溶解度计算值的散点分布

Fig. 2　Scatter plot of experimental log *S* values and calculated log *S* values for training set and test set

图 2 表明：训练集和测试集中的 log *S* 实验值和相应的 log *S* 计算值几乎都落在了 45°对角线附近，仅有个别的几个数据点偏离程度较大，2 个模型的 *R* 分别是 0.986 和 0.982，表明模型预测效果较好。采用 *RMSE* 来定量评价模型：

$$RMSE = \sqrt{\frac{\sum_{i-1}^{n} (y_i^{exp} - y_i^{pre})^2}{n}} \tag{9}$$

训练集和测试集的 *RMSE* 分别是 0.141 和 0.193。

4　排除异常值

为了进一步保证模型的鲁棒性，采用维里图和帽子矩阵排除实验数据中异常的溶解度数据[19]。帽子矩阵：

$$H_i = \boldsymbol{x}_i^{T}(\boldsymbol{X}^{T}\boldsymbol{X})^{-1}x_i \tag{10}$$

式中：X 是训练集的描述符矩阵，x_i 是第 i 个溶解度数据的描述符向量。用方程(11)定义杠杆值的阈值(H^*)如下：

$$H^* = \frac{3(p+1)}{n} \tag{11}$$

式中：n 是训练集中溶解度数据的数量，p 是 GA 选择的描述符的数量，经计算该研究中的 H^* 值是 0.131，H^* 的有效范围是 $0 \leqslant H_i \leqslant H^*$。标准残差(*SR*)定义为

$$SR = \frac{y_i^{\exp} - y_i^{\mathrm{pre}}}{\sqrt{\frac{1}{n-m-1}\sum_{i=1}^{n}\left(y_i^{\exp} - y_i^{\mathrm{pre}}\right)^2\left(1 - H_{ii}\right)}} \quad (12)$$

一般选择 SR 的有效范围为 $-3 \leqslant SR \leqslant 3$。运用上述方法得到图 3。

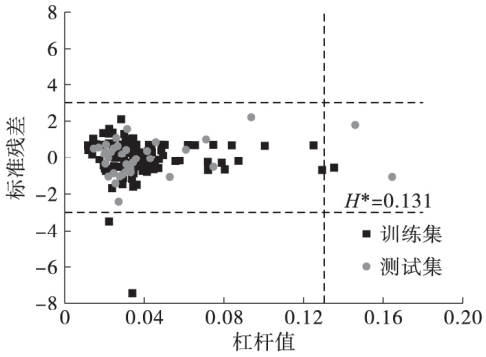

图 3 训练集和测试集中各自含有的异常值分布情况

Fig. 3 Distribution plot of outliers existing in training set and test set

图 3 表明: 在训练集中有 3 个异常值,而在测试集中有 2 个异常值; 正是由于它们的存在,降低了模型的预测性能和稳定性,因此,这些异常值必须从训练集和测试集中移除。

当把这些异常值删除之后,我们用修正后的训练集和测试集分别建立模型和验证模型。修正后的训练集和测试集的 log S 实验值和 log S 计算值的散点分布图被描绘在图 4 中。

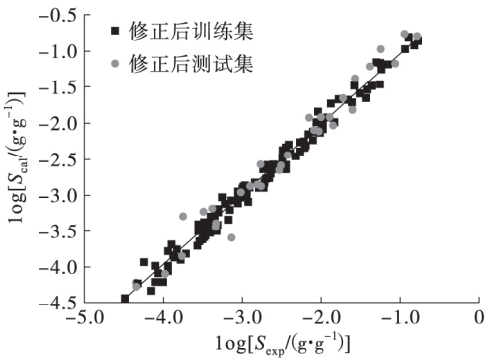

图 4 修正后的训练集和测试集中各自溶解度实验值和溶解度计算值的散点分布情况

Fig. 4 Scatter plot of experimental log S values and calculated log S values for modified training set and test set

图 4 表明: 修正后 GA-SVM 模型的 R 和 $RMSE$ 分别是 0.992 和 0.107,而用修正后的测试集验证模型的稳定性,拟合结果的统计学数据分别为 R 是 0.984 和 $RMSE$ 是 0.177。很明显,模型的预测能力被提高,预测误差减小,从而说明 AD 可以检验实验数据的可靠性,提高模型的性能。

5 结论

(1) 建立了一个 GA-SVM 新模型,为预测 TAGs 在 SC-CO_2 中的溶解度提供了一种新思路。

(2) 采用 171 个文献溶解度实验数据进行该 GA-SVM 模型的建立和验证,结果表明新模型预测 TAGs 在 SC-CO_2 中的溶解度精度较高,训练集和预测集的 R 值分别为 0.992 和 0.984,对应的 $RMSE$ 分别为 0.107 和 0.177。

(3) 基于 AD,排除了实验数据当中的异常值点,进一步提高了模型的预测能力,使模型鲁棒性更好。

符号说明:

b 偏移量

C 惩罚因子

H 帽子矩阵

H_{ii} 第 i 个数据点帽子矩阵的对角线元素

m 模型输入变量的个数

N 样本个数

n 每组甘油三酯相对应的数据点总数

p 压力,MPa

p 遗传算法选择的最优描述符数量

R 相关系数

$RMSE$ 均方根差

SR 模型的标准残差

T 温度,K

$\boldsymbol{X}$ 训练集描述符矩阵

$\boldsymbol{X}^{\mathrm{T}}$ 矩阵的转置

x $x_1, x_2, x_3 \cdots, x_{N-1}, x_N$ 的集合体

x_i 第 i 个学习样本输入值

y $y_1, y_2 \cdots, y_{N-1}, y_N$ 的集合体

y_i 第 i 个学习样本的输出值

y_i^{pre} 第 i 个学习样本的预测值

$y_i^{\exp}$ 第 i 个学习样本的实验值

$\alpha_i, \alpha_i^*, \alpha_j, \alpha_j^*$ 每个样本对应的拉格朗日乘子对

γ 核函数的宽度参数

ε 损失函数的损失因子

ξ, ξ^* 松弛变量

ω 权重

参考文献:

[1] MATUROVÁM, PREININGER V, SANTAVÝF. Supercritical carbon dioxide extraction and characterization of argentinean chia seed oil [J]. J Am Oil Chem Soc,

2011，88(2)：289-298.
[2] MOORTHY A S，LIST G R，ADLOF R O，et al. Using mettler dropping point data from dilute soybean oil-triglyceride mixtures to estimate thermodynamic properties for corresponding pure triglyceride [J]. J Am Oil Chem Soc，2017，94(4)：1-8.
[3] 文震，党志，宗敏华，等. 胆甾醇在超临界 CO_2 中的溶解度测定与关联 [J]. 化学工程，2006，34(11)：44-46.
[4] CHRASTIL J. Solubility of solids and liquids in supercritical gases [J]. J Phys Chem，1982，86(15)：3016-3021.
[5] 蒋春跃，吴建峰，孙志娟，等. 水在超临界二氧化碳中的溶解度 [J]. 化学工程，2014，42(7)：42-47.
[6] 胡德栋，王威强，杜爱玲. 超临界 CO_2 中固体溶解度的逆向传播人工神经网络模拟 [J]. 化学工程，2006，34(5)：45-48.
[7] 文震，李谦，党志，等. 紫苏油在超临界 CO_2 中溶解度的神经网络模型建立 [J]. 化学工程，2003，31(6)：67-70.
[8] 陈静，张倩，卞小强，等. 基于 GA-SVR 模型预测多环芳香烃在超临界 CO_2 中的溶解度 [J]. 石油化工，2017，46(3)：321-326.
[9] XU J，WANG L，WANG L，et al. QSPR study of Setschenow constants of organic compounds using MLR，ANN，and SVM analyses. [J]. J Comput Chem，2011，32(15)：3241-3252.
[10] GONÇALVES M，VASCONCELOS A M P，GOMESD A E J S，et al. On the application of supercritical fluid extraction to the deacidification of olive oils [J]. J Am Oil Chem Soc，1991，68(7)：474-480.
[11] NILSSON W B，GAUGLITZ E J，HUDSON J K. Solubilities of methyl oleate，oleic acid，oleyl glycerols，and oleyl glycerol mixtures in supercritical carbon dioxide [J]. J Am Oil Chem Soc，1991，68(2)：87-91.
[12] WEBER W，PETKOV S，BRUNNER G. Vapourliquide-quilibria and calculations using the RedlichKwon-gAspenequation of state for tristearin，tripalmitin，and triolein in CO_2 and propane [J]. Fluid Phase Equilib，1999，158/159/160(5)：695-706.
[13] BAMBERGER T，ERICKSON J C，COONEY C L，et al. Measurement and model prediction of solubilities of pure fatty acids，pure triglycerides，and mixtures of triglycerides in supercritical carbon dioxide [J]. J Chem & Eng Data，1988，33(3)：327-333.
[14] HAMMAM H. Solubilities of pure lipids in supercritical carbon dioxide [J]. J Supercrit Fluids，1992，5(2)：101-106.
[15] ASHOUR I，HAMMAM H. Equilibrium solubility of pure mono-，di-，and trilaurin in supercritical carbon dioxide experimental measurements and model prediction [J]. J Supercrit Fluids，1993，6(1)：3-8.
[16] PEARCE D L. Solubility of triglycerides in supercritical carbon dioxide [D]. Christchurch，New Zealand：Univ of Canterbury Chem & Process Eng，1990.
[17] TODESCHINI R，CONSONNI V，MANNHOLD R，et al. Molecular descriptors for chemoinformatics [M]. 2nd ed. Hoboken，New Jersey，USA：Wiley VCH Press，2009：27-37，714-726.
[18] STANTON D T，JURS P C. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies [J]. Anal Chem，1990，62(21)：2323-2329.

版 权 声 明

为适应我国信息化建设，扩大本刊及作者知识信息交流渠道，《化学工程》期刊已加入《中国知网 CNKI 系列期刊数据库》、《中国核心期刊(遴选)数据库》(万方数据——数字化期刊群)、《中文科技期刊数据库》、《中国科学引文数据库》、《中国学术期刊文摘(中文版)》、美国《化学文摘》(CA)、俄罗斯《文摘杂志》、《日本科学技术振兴机构中国文献数据库》、荷兰 Scopus、美国《乌利希期刊指南》等数据库。凡本刊发表的论文，将同时通过本刊加入的数据库进行网络出版或提供信息服务，稿件一经刊登，将在本刊稿酬中一次性支付著作权使用报酬(即包括印刷版、光盘版和网络版等各种使用方式的报酬)。如作者不同意论文被上述数据库收录，请向本刊提出书面说明，本刊将作适当处理。

为保护知识产权、杜绝学术不端行为，本刊对拟录用稿件均进行学术不端文献的检测，对于一旦发现一稿多投和抄袭稿件等学术不端行为，编辑部视其情节，作出禁止刊登或网上公告等处罚。

《化学工程》编辑部