

蚁群算法在近红外光谱定量分析中的应用研究

郭亮, 吉海彦*

中国农业大学信息与电气工程学院, 北京 100083

摘要 蚁群算法是新近发展的基于群体智能的仿生优化算法, 它模拟蚂蚁的觅食行为来解决复杂的组合优化问题。蚁群算法的优点是智能搜索、全局优化、鲁棒性、分布式计算和容易与其他算法相结合等。近红外光谱定量分析技术在很多领域得到广泛的应用, 而其关键技术环节之一是建立近红外光谱测量数据的多元校正模型。文章将蚁群算法应用于近红外光谱定量分析中, 建立了谷物样品的傅里叶变换近红外漫反射光谱和谷物中蛋白质含量的定量分析模型, 得到了较好的结果。校准集的相关系数与相对标准偏差分别为 0.943 和 3.41%, 预测集的相关系数与相对标准偏差分别为 0.913 和 4.67%。

关键词 蚁群算法; 近红外光谱; 定量分析模型

中图分类号: O657.3 文献标识码: A 文章编号: 1000-0593(2007)09-1703-03

引言

现代近红外光谱分析技术以其分析速度快、效率高、成本低和易于实现在线分析等特点, 在农业、食品、医药、石化和烟草等行业得到了广泛应用^[1-3]。近红外光谱定量分析的关键技术环节之一是建立近红外光谱测量数据的多元校正模型。常用的建立模型的方法有: 逐步回归分析法(SRA)、主成分回归法(PCR)、偏最小二乘法(PLS)^[4]、人工神经网络法(ANN)^[5,6]等, 其目的是建立用于预测未知样品性质或组成的分析模型。目前使用较多的多元校正算法是 PLS 线性回归方法, 在非线性明显时可用人工神经网络方法等。这些传统的化学计量学算法的一个共同特点是它们都以经典的统计学的渐近理论为依据。

蚁群算法(Ant colony algorithm)是意大利学者 Dorigo 等在 20 世纪 90 年代初, 模拟蚂蚁的觅食行为而提出的一种全新的仿生算法^[7]。该算法具有正反馈原理、分布式计算、鲁棒性强和易与其它算法相结合等突出优点。经过近些年的不断发展, 该算法日趋成熟, 目前已被成功地应用于通讯、交通和人工智能等领域^[8-10]。近年来蚁群算法已成为国内外的一个研究热点, 蚁群算法在光谱分析方面的应用已有报道, 例如丁亚平等将蚁群算法应用于多组分显色体系的解析及多组分导数荧光光谱解析, 获得了满意的结果^[11,12]。

本文用蚁群算法建立了谷物样品的傅里叶变换近红外漫反射光谱和谷物中蛋白质含量的定量分析模型, 得到了较好

的结果。校准集的相关系数与相对标准偏差分别为 0.943% 和 3.41%, 预测集的相关系数与相对标准偏差分别为 0.913% 和 4.67%。

1 蚁群算法应用于光谱定量分析的原理

蚁群算法是一种仿生学算法, 其产生的灵感来自于蚂蚁在觅食过程中的群体协作。蚁群可以迅速准确地找到食物与蚁巢之间的最短路径, 并且能随环境的变化而变化, 适应性地搜索新路径, 产生新的选择。图 1 形象地表现了蚁群觅食的过程。

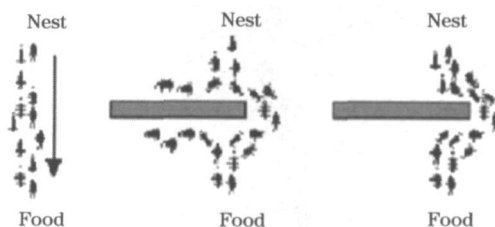


Fig 1 The process of foraging behavior of ants

现实中的蚁群通过一种称为信息素(Pheromone)的物质来实现信息的正反馈, 蚂蚁会在其所通过的路径上释放一种被称为信息素的物质, 使得一定范围内的其他蚂蚁能够察觉并由此影响它们以后的行为。当一些路径上通过的蚂蚁越来越多时, 其留下的信息素轨迹(Trail)也越来越多, 以致信息

收稿日期: 2006-06-30, 修订日期: 2006-09-28

基金项目: 国家高技术研究发展计划项目(“863”计划)(2002AA24805F2)资助

作者简介: 郭亮, 1981 年生, 中国农业大学信息与电气工程学院硕士研究生 * 通讯联系人 e-mail: yuntian@cau.edu.cn

素强度增大, 后来蚂蚁选择该条路径的概率也越高, 从而更增加了该条路径的信息素强度。

按照蚁群算法, 有:

$$P_i = \begin{cases} \text{Pheromone}(i) / \sum \text{Pheromone}(u) & i \in J(u) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$\text{Pheromone}(i) = (1 - \rho) \cdot \text{Pheromone}(i) + Q/g \quad (2)$$

$$\text{Pheromone}(i) = (1 - \rho) \cdot \text{Pheromone}(u) \quad (3)$$

这里 $\text{Pheromone}(i)$ 是被选中路径上的信息素, $J(u)$ 表示未被选中的路径的集合。蚂蚁完成一次循环后, 被选中的路径的信息素将按照(2)式进行调整, 未被选中的按照(3)式进行调整。 Q 是一个常数, 一般称作显著性因子, g 为目标函数。

在光谱建模的过程中, 要得到样品的含量, 就需要将特定样品与特定的光谱之间建立数学模型。如下式:

$$F = C_1X_1 + C_2X_2 + C_3X_3 + \dots + C_nX_n + E \quad (4)$$

在得到了近红外光谱仪器扫描的光谱数据之后, 首先采用主成分分析的方法(PCA)对数据进行降维处理。其中, $X_1, X_2, X_3, \dots, X_n$ 代表过主成分分析(PCA)处理之后的主成分值, n 为主成分数, 本实验取 5 个主成分进行分析。

按照实验要求选择目标函数 e 。目标函数有多种形式, 可以是方差, 偏差, 标准偏差等等, 本文选择

$$e = \sqrt{\sum_{i=1}^n (F_{\text{real}} - F_{\text{count}})^2 / (n - 1)} \quad (5)$$

于是光谱建模的问题就转化为了一个以 e 为优化目标的优化问题, 按照上面的数学模型, 将 $C_1, C_2, C_3, \dots, C_n$ 一组随机数组成的二进制数码串作为求解目标。通过蚁群算法搜索, 并通过目标函数检验选出 $C_1, C_2, C_3, \dots, C_n$ 最优个体作为问题的解。

2 样品光谱及化学值的获得

谷子样品经脱壳后, 用 TECATOR 旋风磨将其磨成粉末并过 60 目筛, 按规范的方法制样。以硫酸钡为背景, 用傅里叶变换近红外光谱仪, 对样品扫描 130 次取平均, 得到样品的近红外漫反射光谱。样品光谱如图 2 所示。在待测成份信息量最为丰富的 4 000~ 5 600 cm^{-1} 间每隔 16 cm^{-1} 取一数据点, 将图谱送入微机。

样品中蛋白质含量的测定是采用国标(GB2905—82)凯氏定氮法, 重复测量 4 次以上取其平均值。

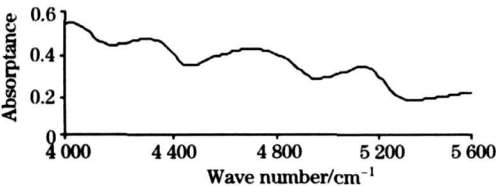


Fig 2 Fourier transform near infrared diffuse reflectance spectroscopy of cereal

3 结果与讨论

3.1 目标函数的选择

目标函数是判断种群中个体优劣和群体优化程度的标准, 选择一个合适的目标函数可以加速算法收敛, 提高计算精度。目标函数包括方差、标准差等。试验结果表明, 选择均方差作为目标函数能够较快地收敛。因此本文选择了均方差作为目标函数。

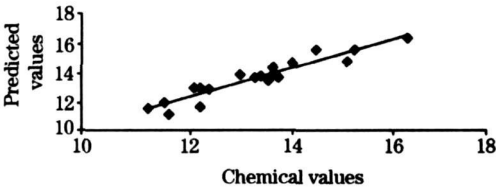


Fig 3 The correlation plot of chemical values and ant colony algorithm predicted values for protein in calibration set

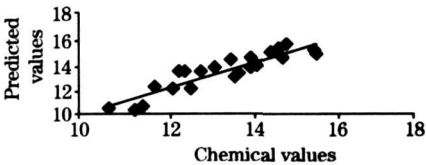


Fig 4 The correlation plot of chemical values and ant colony algorithm predicted values for protein in prediction set

Table 1 Chemical values, predicted values and relative deviation in prediction set

序号	化学值	预测值	相对偏差 / %
1	12.66	12.04	4.9
2	14.26	13.82	3.1
3	11.33	10.36	8.6
4	11.54	10.56	8.5
5	12.90	13.44	-4.1
6	11.80	12.10	-2.6
7	15.68	14.84	5.4
8	12.53	13.34	-6.5
9	10.73	10.52	2.0
10	12.23	12.00	1.9
11	14.85	14.43	2.8
12	14.55	15.00	-3.1
13	13.82	13.17	4.7
14	15.63	15.03	3.8
15	13.70	12.93	5.6
16	14.09	13.71	2.7
17	13.23	13.67	-3.3
18	12.38	13.33	-7.7
19	14.78	16.25	-3.2
20	14.09	14.40	-2.2
21	14.96	15.55	-4.0
22	13.60	14.23	-4.7

3.2 参数的选择

参数 ρ 和 Q 通过实验确定。 $1-\rho$ 是信息素消失因子; ρ 太大或者太小都对结果不利,经试验,本文选择 ρ 为0.3, Q 为1.000。

3.3 谷物中蛋白质含量的预测结果

将42个样品随机分为两组,第一组20个样品作为校准集,第二组22个作为检验集。用VC6.0为编程工具,对校准集样品建立校正模型,并用此模型对检验集的22个样品进行预测。结果为:校准集的相关系数为0.943,相对标准偏差为3.41%;预测集的相关系数为0.913,相对标准偏差为4.67%。图3和图4分别为校准集和预测集中谷子样品蛋白质含量化学值与蚁群算法预测值的相关图。表1为预测集中

样品的化学值、蚁群算法预测值及相对偏差。

4 结 论

蚁群算法是一种新型的仿生学算法,本文将这种算法应用于近红外光谱定量分析中,建立了谷物样品的傅里叶变换近红外漫反射光谱和谷物中蛋白质含量的定量分析模型。建模结果为:校准集的相关系数为0.943,相对标准偏差为3.41%;预测集的相关系数为0.913,相对标准偏差为4.67%。结果表明,蚁群算法能较好地建立近红外光谱定量分析的校准模型。蚁群算法在近红外光谱分析中的进一步应用,尚需进行深入的研究。

参 考 文 献

- [1] Burns Donald A, Ciurczak Emil W. Handbook of Near-Infrared Analysis. New York: Marcel Dekker, Inc., 1992.
- [2] LU Wai zhen, YUAN Hong-fu, XU Guang-tong, et al(陆婉珍,袁洪福,徐广通,等). Modern NIR Spectroscopic Analysis Techniques(现代近红外光谱分析技术). Beijing: China Petrochemistry Press(北京:中国石化出版社), 2000.
- [3] YAN Yan-lu, ZHAO Long-lian, HAN Dong-hai, et al(严衍禄,赵龙莲,韩东海,等). The Foundations and Applications of Near Infrared Spectroscopy Analysis(近红外光谱分析基础及应用). Beijing: China Light Industry Press(北京:中国轻工业出版社), 2005.
- [4] Geladi P, Kowalski B R. Analytica Chimica Acta, 1986, 185: 1.
- [5] Long J R, Groggion V G, Gemperline P J. Anal. Chem., 1990, 62(17): 1791.
- [6] JI Hai-yan, WEN Ming, HAO Bin(吉海彦,闻明,郝斌). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26(1): 57.
- [7] Colomi A, Dorigo M, Maniezzo V. Proc. 1st European Conf. Artificial Life. Paris, France: Elsevier, 1991, 134.
- [8] Bonabeau E, Dorigo M, Theraulaz G. Nature, 2000, 406: 39.
- [9] Dorigo M, Gambardella Luca Maria. Biosystems Engineering, 1997, 43: 73.
- [10] Marcoulides George A, Drezner Zvi. Structural Equation Modeling, 2003, 10(1): 154.
- [11] DING Ya-ping, LIU Ping-yang, SU Qing-de, et al(丁亚平,刘平阳,苏庆德,等). Computers and Applied Chemistry(计算机与应用化学), 2002, 19(3): 326.
- [12] DING Ya-ping, SU Qing-de, WU Qing-sheng(丁亚平,苏庆德,吴庆生). Chemical Journal of Chinese Universities(高等学校化学学报), 2002, 23(9): 1695.

Application Study of Ant Colony Algorithm in Near Infrared Spectroscopy Quantitative Analysis

GUO Liang, JI Hai-yan*

College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China

Abstract Ant colony algorithm is a novel bio-inspired optimization algorithm, which simulates the foraging behavior of ants for solving various complex combinatorial optimization problems. The advantages of ant colony algorithm are intelligent search, global optimization, robustness, distributed computation and easy combination with other heuristic method. Near infrared spectroscopy quantitative analysis has been applied in many fields, whereas the key step is building the calibration model of measured data. In the present paper, ant colony algorithm was used to build the quantitative analysis model of Fourier transform near infrared diffuse spectroscopy for protein in cereal. Satisfied results were obtained. For calibration set, the correlation coefficient and relative standard deviation were 0.943 and 3.41%, respectively, while for prediction set, the correlation coefficient and relative standard deviation were 0.913 and 4.67%, respectively.

Keywords Ant colony algorithm; Near infrared spectroscopy; Quantitative analysis model

* Corresponding author

(Received Jun. 30, 2006; accepted Sep. 28, 2006)