2014

ISSN : 0974 - 7435 Volume 10 Issue 22

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(22), 2014 [13839-13845]

Prediction of naringin content based on machine learning methods

Yan Zeng[1], Xinwen Cheng[1], Qi Li[1], Xiao Wang[1], Yuyun Chen[2]

[1]School of Computer Science, Sichuan University of Science & Engineering, Zigong, 643000, (CHINA)

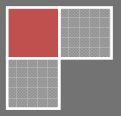
[2]School of Chemistry and Pharmaceutical Engineering, Sichuan University of Science & Engineering, Zigong, 643000, (CHINA)

ABSTRACT

To increase the accuracy and speed of measurement of Naringin extraction rate, the prediction of Naringin extraction rate is raised based on Weighted Least Square Support Vector Machine (WLSSVM) and improved Artificial Bee Colony (ABC) of the machine learning methods. Taking the ratio of material to solvent, the extracting time, ethanol concentration and extracting temperature which influence Naringin extraction rate as the input of WLSSVM and Naringin extraction rate as output, learn extracting Naringin test data from shaddock peels. The results of simulation indicate that the prediction of improved ABC algorithm and advanced WLSSVM acquires better prediction speed, accuracy and stability and is appropriate for the prediction of Naringin extraction, compared with the methods of LSSVM and ABC-LSSVM.

KEYWORDS

Naringin extraction rate; Weighted least square support vector machine (WLSSVM); Artificial bee colony algorithm (ABC algorithm); Prediction.



INTRODUCTION

Naringin, as a flavanone compound, is abundant in shaddock peels. Modern medical research discovers that Naringin has pharmacological effects including antioxidant, cancer prevention, anti-microbial, reducing blood pressure and cholesterol, and etc. Meanwhile, Naringin is used as natural pigment, flavor improving agent and bitterant in food industry. Therefore, extracting Naringin from shaddock peels not only reasonably and effectively adopts the waste shaddock peels, but also further realize the deep development and utilization of Naringin.

Generally, the measurement of Naringin extraction rate adopts the methods of chromatography and spectrophotometry. The traditional measurement methods are influenced by some unpredictable elements such as artificial operation and the variation of the environment, resulting decreasing the accuracy of the data. Thus, machine learning methods in artificial intelligence are effective means to solve biomass parameter prediction, for instance, neural network, support vector machine (SVM), and etc. Reference[5] applies neural network for the prediction of crop evapotranspiration. But the generalization ability of neural network is limited, so the application of small sample data is also limited. Reference[6] applies SVM for the prediction of lycopene content. SVM can deal with the problem of nonlinear small sample data, but the generalization ability is limited. Least Square Support Vector Machine (LSSVM) not only solves the problem of the application limitation of small sample data by neural network, but also copes with the problems of the limited generalization ability of SVM and the sensitivity of SVM to the abnormal samples. As a result, LSSVM is a research hotspot of current prediction methods. The important parameter indexes to affect the prediction accuracy of LSSVM are regularization parameter and kernel bandwidth coefficient. The traditional cross-validation method can ensure high accuracy of prediction but with slow prediction speed and bad robustness, constraining its practical application. Artificial Bee Colony algorithm (ABC algorithm) is a swarm intelligence optimization algorithm with convenience and global optimization ability. ABC algorithm is often applied for parameter optimization of neural network and LSSVM. But its rate of convergence is slow and easy to fall into local optimum, so its application is also limited.

First of all, this paper improves ABC algorithm, introducing initialization strategy of chaos sequence and selection policy following by bee based on "pheromone-sensitivity". Consequently, the rate of convergence and the ability of local optimization of algorithm are improved. Meanwhile, introducing weight vector into LSSVM can solve the problem of bad robustness. Secondly, apply improved ABC algorithm for optimizing the prediction of WLSSVM, conducting simulation contrast test of the measurement of Naringin extraction rate.

ALGORITHM PRINCIPLE

Improved ABC algorithm

ABC algorithm simulates the intelligence searching behavior of bee colonies, with the advantage of convenience and easy to realization, attracting wide attention of worldwide scientists in recent years. However, normal ABC algorithm is easy to fall into local optimum, occurring prematurity and stagnation. Thus, it shall be adopted the following strategies to improve.

Chaos initialization population

The quality of population directly influences the rate of convergence and solving accuracy of ABC algorithm. Bring in chaos sequence at the stage of initializing population and take advantage of ergodicity of chaos to ensure the diversity of initialization population, which can increase the solving rate and improve the quality of solution.

Produce randomly a D dimension vector $X_1=(x_{1,1}, x_{1,2}, \ldots, x_{1,D})$, and $x_{1,j} \in (0,1)$. D is the dimension of solution space. Adopt Logistic equation:

$$X_{i+1} = \mu X_i (1 - X_i) \quad \textbf{(1)}$$

Among the equation, μ is controlled parameter. When μ=4, Logistic lies in chaos and produces chaos sequences X_1, X_2, …, X_k. Map chaos sequences to the search space to begin chaos searching, producing a number of candidate nectar sources. Then choose the initialization nectar source with good fitness in order to increase the rate of convergence and accuracy.

Pheromone-sensitivity selection model

In the normal ABC algorithm, the following bees choose an employed bee to follow based on the fitness value of the employed bee population by the means of roulette, which is easy to result that the algorithm converges early. Thus, the improving algorithm combines pheromone and sensitivity to substitute the chosen strategy of roulette, which is shown as following:

$$p(i) = \begin{cases} \dfrac{f(i) - f_{\min}}{f_{\max} - f_{\min}} & f_{\max} \neq f_{\min} \\ 0 & p(\mathrm{k}) \leq s(i) \end{cases} \quad \textbf{(2)}$$

Among the equation, $f_{(i)}$ is the individual fitness value; f_{max} is the maximal fitness value and f_{min} is the minimal fitness value; p(k) is number k pheromone of food source which is not equal to *i*; s(i) is the sensitivity of number i following bee: s(i)~U (0,1).

The value of pheromone is in direct proportion to the value of objective function and its value reflects the quality of solution. Pheromone updates dynamically after finishing every search process. The sensitivity confirms the searching region direction, while pheromone must match the sensitivity in its searching region. Therefore, Pheromone-Sensitivity selection model prevents algorithm falling into local optimum to some extent and ensures the direction of rapid evolution of algorithm.

WLSSVM

LSSVM improves SVM, solving the problems of small samples, nonlinear and high dimensions and increasing the solving speed and generalization ability. Through constructing the following regression function, LSSVM converts nonlinear problem into linear estimation in high dimensions feature space:

$$f(x) = \omega^T \varphi(x) + b \tag{3}$$

Among the function, ω and b are undetermined parameters. The normal solution of LSSVM has bad robustness, so it can be solved by adding a weight v_i to the error variance ξ_i on the basis of normal LSSVM. Improved objective function of LSSVM is presented as:

$$\begin{aligned} &\min J(\omega, \xi) = \frac{1}{2}\omega^T \omega + \frac{1}{2}\gamma \sum_{i=1}^{n} v_i \xi_i^2 \\ &s.t. \quad y_i = \omega^T \varphi(x_i) + b + \xi_i, i = 1, 2, \ldots, n \end{aligned} \tag{4}$$

Among this function, ξ_i is the error between the true value and the prediction value of number i sample; γ is the regularization parameter which can adjust dynamically.

To the convenience of solving the function, Lagrange function can be utilized to solve the optimization:

$$\begin{aligned} L(\omega, \beta, \xi, \alpha) = &\frac{1}{2}\omega^T \omega + \frac{1}{2}\gamma \sum_{i=1}^{n} v_i \xi_i^2 - \\ &\sum_{i=1}^{n} \alpha_i \left[\omega \varphi(x_i) + b + \xi_i - y_i \right] \end{aligned} \tag{5}$$

Among this function, $a_1 \in R$ *(i=1,2,...,N)* is the multiplier of Lagrange. Weight v_i is confirmed by $\xi_i = a_i / \gamma$.

$$v_i = \begin{cases} 1, & \left|\xi_i / \hat{s}\right| < c_1 \\ \dfrac{c_2 - \left|\xi_i / \hat{s}\right|}{c_2 - c_1}, & c_2 \le \left|\xi_i / \hat{s}\right| \le c_1 \\ 10^{-4}, & others \end{cases} \tag{6}$$

Among this function, $\hat{s}$ is the robustness estimation of standard deviation of error, and $\hat{s} = \dfrac{IQR}{2 \times 0.6745}$. *IQR* is the inter-quartile range of error ξ_i, which is used to measure the deviation from the estimation error of Gaussian distribution. Through equation (4), add weight vector. Compared with normal LSSVM, even if the error deviation is not subject to Gaussian distribution, the better analysis can be achieved. As a result, the robustness of algorithm has been increased. According to the density estimation of error deviation, the constant c_1 and c_2 usually are c_1=2.5 and c_2=3.

The solution of *a* and *b* is related to the selection of kernel function. Radical Basis Function (RBF) whose speed of learning is fast usually is chosen:

$$K(x_i, x_j) = \exp\left(-\frac{\left| x_i - x_j \right|^2}{2\sigma^2} \right) \tag{7}$$

Among this function, σ is kernel bandwidth and the functional estimation of WLSSVM is achieved as following:

$$f(x)=\sum_{i=1}^{n} a_i k(x_i, x)+b, \tag{8}$$

In a word, kernel bandwidth σ and regularization parameter γ are important parameters of LSSVM. The selection of these parameters directly affects the learning ability and generalization ability of LSSVM.

NARINGIN EXTRACTION METHOD AND PREDICTIVE MODELING

The extraction and measurement of naringin

After dying the mature fresh shaddock peels, smash them into powder and weigh quantitative shaddock peel powder. Then, add ethanol-water solution at certain concentration. Ultrasonic extraction is done for a while at setting temperature. Next, vacuum filtration can get the extracting solution. At last, measure absorbance by the means of UV spectrophotometry and calculate Naringin extraction rate based on the absorbance. The equation is shown as:

$$\text{Naringin extraction rate } \% = \frac{(A+0.0003)V}{26.026\times 1000M}\times 100 \tag{9}$$

Among this equation, A is absorbance; V is the quantity of extracting solution (ml); M is the mass of shaddock peel powder (g).

It can be concluded that the process of measuring Naringin in the shaddock peels is complicated and the measuring time is long.

The selection of data samples

In the experiment of extracting Naringin from shaddock peels, the first step is to carry on the single-factor experiment on the condition of five ratios of material to solvent (1:10, 1:20, 1:30, 1:40, 1:50), five ethanol concentrations (30%, 50%, 70%, 90%, 100%), four extracting time (20min, 40min, 60mim, 70min) and four extracting temperature (30℃, 40℃, 50℃, 60℃) to measure the influence on Naringin extracting rate of four factors including ethanol concentration, the ratio of material to solvent, extracting temperature and time. The result is that four factor can affect Naringin extraction rate at a different degree.

On the basis of single-factor experiment, choose ethanol concentration, ratio of material to solvent, extracting time and temperature as research factors and choose Naringin extraction rate as evaluation index. Then, four factors and three levels orthogonal test is made. Orthogonal factor level table is shown as following TABLE 1.

TABLE 1 : Orthogonal factor level table

	Factors			
Level	A Extracting temperature (℃)	B Extracting time (min)	C Ratio of material to solvent	D Ethanol concentration (%)
1	30	35	1:25	65
2	40	40	1:30	70
3	50	45	1:35	75

According to the analysis of the results of orthogonal test, it can be concluded that there is a significant influence on Naringin extraction rate by four nonlinear factors, including ratio of material to solvent, ethanol concentration, extracting time and temperature and there is small interaction between four factors, which are easy to measure. Consequently, choose ratio of material to solvent, ethanol concentration, extracting time and temperature as auxiliary variables of prediction model to input and Naringin extraction rate as primary variable to output.

Take 30 groups of data got from Naringin extraction test as sample data. In order to increase the learning ability and prediction accuracy of the model, the priority is to normalize the data and choose the front 18 groups of data as the training data. Construct the improved WLSSVM prediction model based on improved ABC.

The predictive modeling of improved WLSSVM based on improved ABC

Through the analysis of 3.2, it has been confirmed that the procedures of constructing the prediction model of Naringin extraction rate based on IABC-WLSSVM are presented as following, which takes four measurable variations including ratio of material to solvent, ethanol concentration, extracting time and temperature as the input variations of the model and Naringin extraction rate as output variation, combined with improved ABC algorithm and WLSSVM theory and the trial data of Naringin extraction from shaddock peels.

1) Normalize the related data of the experiment of extracting Naringin from shaddock peels.

2) Initialize parameters. The maximum search limit is 30; the maximum iteration is 200; the total number of bee colonies is 40. Produce chaos sequence based on equation (1) and map the chaos sequence into the solution space to produce initial solution and calculate the fitness function value.
3) Record the optimal value for the employed bee at step n and develop neighborhood search to produce a new location.
4) When the solution of neighborhood search by employed bees is better than the recorded optimal solution, substitute the recorded solution; otherwise, keep the recorded solution.
5) After finishing neighborhood search, all employed bees dance the waggle dance to share food source information with the following bees. The following bees choose the employed bees according to the Pheromone-Sensitivity selection model.
6) The same as step 3) and step 4), the following bees record the optimal fitness value and according parameters after the colony finishes its last update.
7) When the neighborhood search number by the following bees achieves its threshold *Limit*, if the following bees still do not find the better location, the scout bees initialize the location of food source again.
8) If the maximum iteration is achieved, stop calculating and output the optimal value; otherwise, repeat step 3).
9) Improve the regularization parameter γ and kernel bandwidth σ of WLSSVM according to the optimal solution. Construct IABC-WLSSVM prediction model and take simulation and test by using Naringin extraction data.

THE RESULTS OF SIMULATION AND THE ANALYSIS

To verify the effectiveness of IABC-WLSSVM prediction model, take 30 groups of data achieved from Naringin extraction experiment as sample data and the front 18 groups of data as training data to construct model. The latter 12 groups of data are taken as test data to verify this model.

During the simulation test, the computer configuration is Lenovo Yonah E5800@3, 2GHz with 2G memory and Windows XP operation system. The testing environment is Matlab 7.0.

The optimal parameter of WLSSVM is improved by IABC algorithm: the kernel bandwidth σ=62.1147 and regularization parameter γ=98.937. Compare the simulation data got from the prediction model based on IABC-WLSSVM with the data of LSSVM model and normal ABC-WLSSVM model, the result is shown as TABLE 2.

TABLE 2 : The result comparison of naringin extraction rate

Number	The actual extraction rate	LSSVM		ABC-LSSVM		IABC-WLSSVM	
		Simulation results	Relative error (%)	Simulation results	Relative error (%)	Simulation results	Relative Error (%)
1	3.455	3.178	-8.017	3.227	-6.599	3.294	-4.660
2	3.214	3.050	-5.103	3.108	-3.298	3.128	-2.676
3	2.981	2.953	-0.939	2.942	-1.308	2.965	-0.537
4	3.127	2.734	-12.568	2.875	-8.059	2.993	-4.285
5	3.595	3.499	-2.670	3.493	-2.837	3.569	-0.723
6	3.475	3.762	8.259	3.876	11.540	3.642	4.806
7	3.443	3.381	-1.801	3.352	-2.643	3.376	-1.946
8	3.705	3.516	-5.101	3.532	-4.669	3.607	-2.645
9	2.928	3.147	7.480	3.124	6.694	3.042	3.893
10	3.366	2.983	-11.378	3.190	-5.229	3.141	-6.684
11	3.115	2.975	-4.494	2.991	-3.981	3.062	-1.701
12	3.432	3.338	-2.739	3.366	-1.923	3.410	-0.641

It can be concluded from TABLE 2 that the maximal relative error between the simulation data achieved by adopting IABC-WLSSVM prediction model and the actual data is 6.684, and the minimal relative error is 0.537%, and the average relative error is 2.933%. The average relative error of LSSVM model is 5.879%. The average relative error of normal ABC-LSSVM is 4.898%. Compared with the latter two relative errors, IABC-WLSSVM prediction model has better learning ability and higher accuracy.

See Figure 1 in order to describe more intuitively the error between the estimated values of IABC-WLSSVM prediction model and the actual measured values.

It can be seen from Figure 1 that the data degree of fitting of the measured data and estimated data is good. Figure 2 is the relative error graph. The relative error keeps within 8%, presenting that the predicable data of the prediction model are reliable.

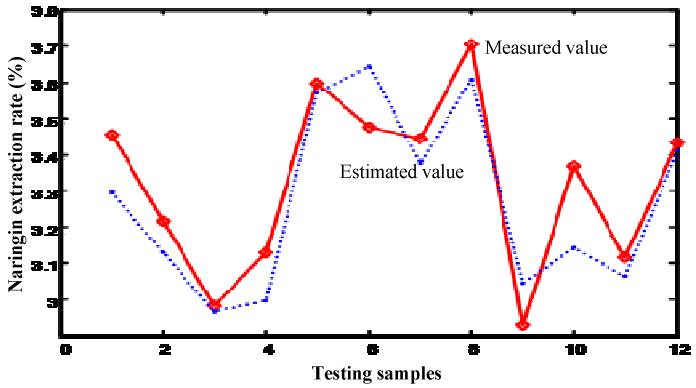


Figure 1 : The comparison of the measured data and simulation data of naringin extraction

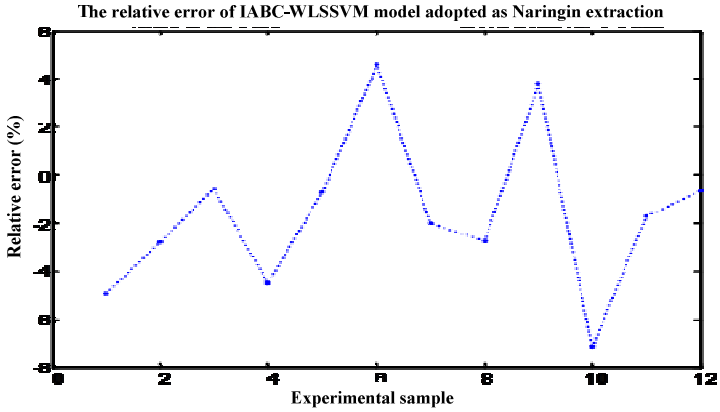


Figure 2 : The relative error of IABC-WLSSVM model

Figure 3 is the evolution graph of the prediction model of Naringin extraction rate based on IABC-WLSSVM. It can be seen from it that the rate of convergence is rapid and the property is stable of the prediction method of IABC-WLSSVM Naringin extraction rate.

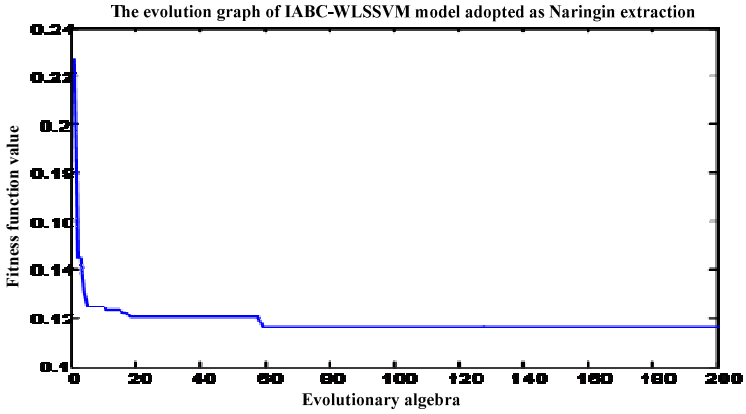


Figure 3 : The evolution graph of IABC-WLSSVM prediction model

CONCLUSION

This paper takes key index Naringin extraction rate of the experiment of extracting Naringin from shaddock peels as the object and applies machine learning prediction method combined with IABC algorithm and WLSSVM for the measurement of Naringin extraction rate. The results of simulation and comparison show that the average error of improved prediction method is 2.933%, which is better than 5.879% of LSSVM method and 4.898% of ABC-LSSVM method. It is shown that the prediction method of IABC-WLSSVM explained in this paper is easy to implement and acquires higher prediction speed and accuracy, providing a certain theory research foundation for industrial production of flavanone compounds such as Naringin.

ACKNOWLEDGEMENT

(1) Liquor-making biotech and its application, open foundation item of key laboratories in Sichuan Province (No. NJ2011-09)

(2) Enterprise informatization and the technology of measurement and control of Internet of things, open foundation item of key universities' laboratories in Sichuan Province (No. 2014WYJ08)

(3) The scientific research project funding by Sichuan University of Science & Engineering(No. 2013KY04)

REFERENCES

[1] Qiong You, Keng Wu; The pharmacological effects of naringin on cardiovascular system [J], Guangdong Medical Journal, **31(22)**, 3006-3008 **(2010)**.

[2] Chunyan Jian, Runmin Guo, Danming Liu, etc; Effects of emodin on cell proliferation, FN expression and p38MAPK pathway in rat mesangial cells cultured under high glucose [J], Chinese Pharmacological Bulletin, **30(2)**, 238-243 **(2014)**.

[3] Yilin He, Lu Wang, Xiaohong Wu; Determination of naringin in nanofiber membranes by reversed-phase high performance liquid chromatographic [J], Journal of Chongqing Medical University, **39(9)**, 1296-1300 **(2014)**.

[4] Yanghe Luo, Xuefeng Wei, Qinglin Xie; Determination of total flavonoids from water chestnut peel by vis spectrophotometry [J], Food Research and Development, **30(6)**, 135-138 **(2009)**.

[5] Zhizheng Zhang, Yatao Jiao, Wei Li; Prediction of crop reference evapotranspiration based on GA-BP neural network [J], Journal of Agricultural Mechanization Research, **(1)**, 61-64 **(2011)**.

[6] Wei Liu, Jianping Wang, Changhong Liu, etc; Lycopene content prediction based on support vector machine with particle swarm optimization [J], Transactions of the Chinese Society for Agricultural Machinery, **43(4)**, 143-147, 155 **(2012)**.

[7] Liangzhi Xia, Hua Li, Keke Rao, etc; Research on soft sensor modeling of vinyl acetate polymerization rate based on hybrid QGA-LSSVM, Computer Measurement & Control, **20(4)**, 907-909, 913 **(2012)**.

[8] Xiaolin Huang, Lei Shi, Johan A.K.Suykens; Asymmetric least squares support vector machine classifiers [J], Computational Statistics & Data Analysis, **70(2)**, 395-405 **(2014)**.

[9] D.Karaboga; An idea based on honey bee swarm for numerical optimization[R], Erciyes University, Engineering Faculty, Computer Engineering Department, **(2005)**.

[10] Jiuchong Wang, Xiaoguang Fan, Sheng sheng, etc; Improved artificial bee coloby LS-SVM fault prediction [J], Journal of Air Force Engineering University (Natural Science Edition), **14(1)**, 16-19 **(2013)**.

[11] Xiaojun Bi, Yanjiao Wang; A modified artificial bee colony algorithm and its application [J], Journal of Harbin Engineering University, **33(1)**, 117-123 **(2012)**.

[12] Hongqiu Zhu, Chunhua Yang, Weihua Gui; Particle swarm optimization with chaotic mutation [J], Computer Science, **37(3)**, 215-217 **(2010)**.

[13] Huiying Wang, Jianjun Liu, Quanzhou Wang; Modified artificial bee colony algorithm for numerical function optimization [J], Computer Engineering and Applications, **48(19)**, 36-39 **(2012)**.

[14] V.Vapnik; The nature of statistical learning theory [M], [S.l.], Springer Verlag, **(1995)**.

[15] Shuxia Lu, Runa Tian; Structural weighted least squares support vector machine classifier [J], Computer Science, **40(12)**, 52-54, 80 **(2013)**.