The Application of Machine learning in Analysis Traditional herbal Seeds

Junjian Bai
702008401

what is Licorice?

Licorice is a traditional Chinese herbal medicine. Its seed is characterized by hardness. Usually, the hard rate of licorice-seed is determined by soaking the seeds, but this method is time-consuming and sometimes destroys the seeds. Therefore, developing a fast and nondestructive analysis technique for determining hard rate of licorice seeds is important and could promote the application of hard seeds in cultivation.

what is Near-infrared (NIR)spectroscopy

Near-infrared (NIR) spectroscopy [Wang, Xue and Sun: (2012); Yang, Gao and Sun: (2015)] is based on the absorption of electromagnetic radiation ranging from 4000 to 12000cm-1. NIR spectra can provide rich information on molecular structure. Recently, it has demonstrated great potential in the analysis of complex samples owing to its simplicity, rapidity and nondestructivity, and has been successfully applied to analyze the chemical ingredients and quality parameters of compounds.

magic machine learning

People used machine learning to solve this problem before. They tried to apply Support Vector Machines which owns better generalization regression ability compared with other machine learning methods like artificial neural network (ANN).

But

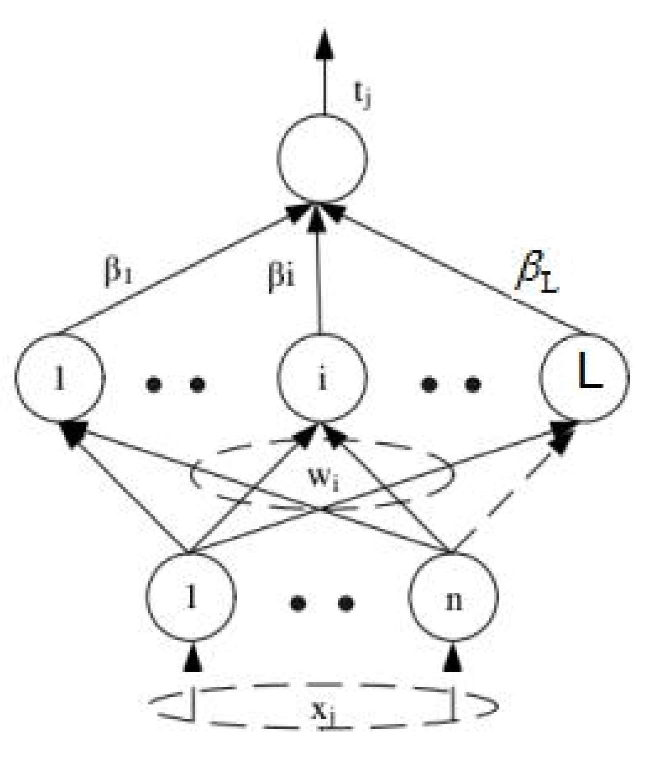
The learning speed of classical Support Vector Regression (SVR) is low, since it is constructed based on the minimization of a convex quadratic function for all training samples.
the learning speed of Back Propagation networks(BP) is low, since the slow gradient-based learning algorithms are extensively used to train neural networks, and all the parameters of the networks are tuned iteratively by using such learning algorithms.

what should we do?

In order to solve this problem, I want to use a algorithm called extreme learning machine (ELM) which randomly chooses hidden layer biases and input weights and doesn't need to adjust them during the training. In theory, this algorithm tends to provide good generalization performance at extremely fast learning speed.

Extreme learning machine (ELM) is a type of single-hidden layer feedforward neural networks (SLFNs) and has been successfully applied to both classifi cation and regression problems. Here, I give a brief definition of ELM on regression; a more detailed description of ELM is available in the literature [Huang, Siew and Zhu:(2006); Jose,Martinez and Pablo:(2011)].

ELM



The essence of ELM is that its hidden-layer parameters are not necessarily tuned, and training error is minimized. Specifically, given a set of N patterns:

$D=\{(x_i,t_i)\,|\,x_i\in\mathbf{R}^n,\quad \boldsymbol{t}_i=[t_{i1},t_{i2},\quad\cdots,t_{im}]^T\in\mathbf{R}^m,i=1,\cdots,N\}$ which the number of hidden-layer is L, where xi is the input vector, and ti is the target value. We expect to find a standard SLFN with L hidden nodes to approximate these N patterns with zero error, which means that the desired output for the j-th pattern is

$$\sum_{i=1}^{L}\boldsymbol{\beta}_i g(\boldsymbol{w}_i\cdot x_j+b_i)=\boldsymbol{t}_j,\ j=1,\cdots,N$$

where wi is the weight vector connecting the i-th hidden node with the input node, and bi denotes the bias term of the i-th hidden node. The β i is the output weight from the i-th hidden node to the output node. The g(x) is an activation function and g(wi・xj +bi) is the output of the i-th hidden node. The linear system is equivalent to the following matrix equation

$$\mathbf{H}\beta=T$$

with

$$\mathbf{H}=\begin{bmatrix} g(\boldsymbol{w}_1\cdot x_1+b_1) & \cdots & g(\boldsymbol{w}_L\cdot x_1+b_L) \\ \vdots & \cdots & \vdots \\ g(\boldsymbol{w}_1\cdot x_N+b_1) & \cdots & g(\boldsymbol{w}_L\cdot x_N+b_L) \end{bmatrix}_{N\cdot L}\quad \beta=\begin{bmatrix}\boldsymbol{\beta}_1^T\\ \vdots \\ \boldsymbol{\beta}_L^T\end{bmatrix}_{L\cdot N}\quad and\quad \mathbf{T}=\begin{bmatrix}t_1^T\\ \vdots \\ t_L^T\end{bmatrix}_{N\cdot m}$$

where β = (β 1, β 2,…., β L). H is defined as the hidden layer output matrix, the i-th column of which is the i-th hidden node output with respect to the input xi. The T is the desired output. Huang et al pointed out that the input weights wi and hidden layer biases bi for the SLFN are not necessarily tuned during training and may be assigned values randomly. Based on this scheme, Huang et al proposed a simple SLFN algorithm, called ELM, the goal of which is to find a least-squares solution of the linear system . This can be posed as the following optimization

$$\| \mathbf{H}\hat{\beta} - \mathbf{T} \|_2^2 = \min_{\beta} \| H\beta - \mathbf{T} \|_2^2$$

which is a normal quadratical programming with no constraints. With $\mathrm{H}^T\mathrm{H}$ being positive definite, its optimal solution $\hat{\beta}$ can be obtained by

$\hat{\beta} = \mathbf{H}^{\dagger}\mathbf{T}$ where $\mathbf{H}^{\dagger} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$

where H† is the Moore-Penrose generalized inverse of matrix H.

Experiments design

In this section, the performance of the proposed ELM learning algorithm is compared with the popular algorithms of feedforward neural networks like the conventional back propagation(BP)algorithm and support vector machine(SVM). All the simulations for the BP, SVM and ELM algorithms are carried out in MATLAB 13 environment.

The activation function used in ELM algorithm is a simple sigmoidal function g(x)=1/(1+exp(-x))
The parameter C is tuned and set as C = 100 in SVR algorithm.

I use these 5 agorithm to anlyz the performance of the proposed models

standard error of calibration (SEC)

$$SEC = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-1}}$$

standard error of prediction (SEP)

$$SEP = \sqrt{\frac{\sum_{i=1}^{n^*}(y_i - \hat{y}_i)^2}{n^*-1}}$$

sum-squared error of test

$$SSE = \sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

sum-squared deviation of test samples

$$SST = \sum_{i=1}^{m}(y_i - \hat{y})^2$$

Sum-Squared regression of test

$$SSR = \sum_{i=1}^{m}(\hat{y}_i - \overline{y})^2$$

In general, a small SSE/SST means the estimates are consistent with the real values. Typically the SSR/SST increases as the SSE/SST decreases. In fact, an extremely small value for the SSE/SST is not desirable because it probably means that the regressor is overfitting. Therefore, a good estimator should strike the balance between the SSE/SST and SSR/SST.

Data set

I use the data of licorice seeds of Near-infrared (NIR) spectroscopy to analysis the hard rate of licorice seeds.（The data is from the China Agriculture University）

The licorice seeds used in this experiment were harvested between 2002 and 2007, from various locations within China A total of 112 licorice seeds were used in the experiment.

Consequently, the spectral data set contains 112 samples measured at 2100 wavelengths in the range of 4000 to 12000cm−1. The NIR spectra of the licorice seeds are shown in Figure 1.

To evaluate the performance of the proposed models, the data is carried out on four different spectral regions: 4000-6000cm−1, 6000-8000cm−1, 8000-10000cm−1, 10000-12000cm−1

Fig. 1 NIR of licorice seeds

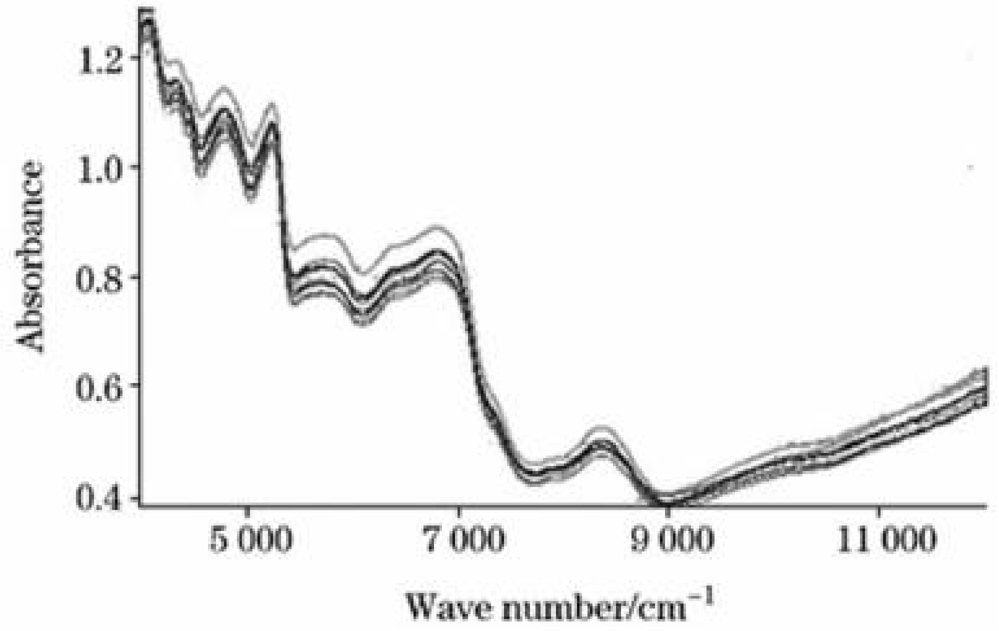


Table 1 The information on four spectral data sets

Dataset	Spectral range cm^{-1}	Number of samples	Number of wavelengths
Set A	10,000-12,000	112	525
Set B	8000-10,000	112	525
Set C	6000-8000	112	525
Set D	4000-6000	112	525

each spectrum in the 4000-12000cm−1 wavelength range was represented as a column vector; the length of the vector was defined by the number of wavelengths.

Table 2 Comparison of ELM, SVR and BP in terms of SEC, SEP, SSE/SST and SSR/SST

Data set	Methods	SEC	SSE/SST	SSR/SST	SEP
Set A	ELM	0.1463	0.3012	0.7156	0.2049
	SVR	0.1523	0.3257	0.6520	0.2958
	BP	0.1595	0.3100	0.7099	0.2029
Set B	ELM	0.0641	0.0589	0.9393	0.1065
	SVR	0.0752	0.0708	0.8200	0.2733
	BP	0.0705	0.0855	0.9244	0.1059
Set C	ELM	0.0491	0.0657	0.9012	0.0831
	SVR	0.0587	0.0794	0.9278	0.0939
	BP	0.0593	0.0791	0.8411	0.0821
Set D	ELM	0.0551	0.0839	0.9684	0.0783
	SVR	0.1323	0.0823	0.9097	0.0931
	BP	0.1452	0.0787	0.9502	0.0748

Table 3 Comparison of ELM, SVR and BP in terms of CPU-time

	Methods	Set A	Set B	Set C	Set D
Time (s)	ELM	0.0140	0.0149	0.0145	0.0153
	SVR	0.0449	0.0472	0.0443	0.0457
	BP	0.0611	0.0608	0.0620	0.0632

Experiment results

From the Table 2 and Table 3, Comparing with Support Vector Regression （SVR） and Back Propagation （BP） network, experimental results in different spectral regions show that the feasibility and effectiveness of the proposed method.

The ELM learning algorithm spent less CPU time obtaining the better testing result, however, it takes longer CPU time for BP,SVM algorithm to reach a much higher testing error.

So, the ELM has faster speed and more accurate learning result than BP and SVR.

Conclusions and Future Directions

This investigation will provide the theoretical support and practical method for the hardness of licorice seeds using ELM and NIR technology.

However, the output weight of ELM was estimated by least square estimation (LSE) method, and it makes ELM network lack robustness since least square estimation is relatively sensitive to outlier.

In the future, maybe we can solve this problem by using least absolute estimation instand of least square estimation in the extreme learning machine.