Classification Model of Seed Cotton Grade Based on Least Square Support Vector Machine Regression Method

Si Chen
Department of Electrical Enginnering
University at Buffalo, The State University of New York
Buffalo, New York, US
schen23@buffalo.edu

Ling Li-na[1] ,Yuan Rong-chang[2] ,Sun Long-qing[3]
[1,3]*China Agricultural University*
[2]*Power Automation Department*
[1]*645966069@qq.com*
[2]*rongchangyuan@qq.com*
[3]*sunlq@cau.edu.cn*

***Abstract*— Grade classification of seed cotton is a major problem that has an significant impact on the agricultural economy. According to characteristics like impurities, yellowness and brightness that extract from images of seed cotton, constructing classification model of seed cotton base on the least square method. Using support vector machine regression to come up with a well improved algorithm. After full learning, seed cotton classification accuracy satisfy the actual application needs.**

***Index Terms*— seed cotton, fuzzy math, pattern recognition, least square method, support vector machine**

I. Introduction

The seed cotton acquisition plays an important role in the chain of the cotton industry, and the grade of seed cotton is the main basis for decision seed cotton price. "Feel & eye-measurement" is current the main method that we rely on when we do seed cotton grade classification test. This manual inspection is limited by the experience of the examiner and also vulnerable to human feelings, bias, benefits and many other factors [1]. Thus, the acquisition price is highly related to the examiner, it may not keep fairness all the time and it frequently damaged the benefits of the farmers [2].

According to the China national seed cotton grade classification standard, using image processing techniques, feature extraction and analysis in the image of seed cotton impurities, yellowness, brightness and other characteristics, this paper proposed a seed cotton grade classification model based on the least squares method and use of support vector machines to achieve automatic classification of the seed cotton grades.

II. Image Acquisition and Feature Extraction

A. Image Acquisition

This paper is based on stratified sampling method to collect different grades of seed cotton sample image. For samples in grade 3-5, amount of 15,20 and 15 samples are selected. The seed cotton sample are positioned in black background and images are generated by digital camera.

B. Feature Extraction

In order to better extract the characteristic parameters of the seed cotton image, several pro-processing methods are being applied, such as image noise reduction, brightness transform and seed cotton image segmentation, the results shows in Fig. 1. Thus, through color space conversion, we get the yellow degrees, magenta degrees, brightness, cyan degrees, hue and saturation information[7] [8].

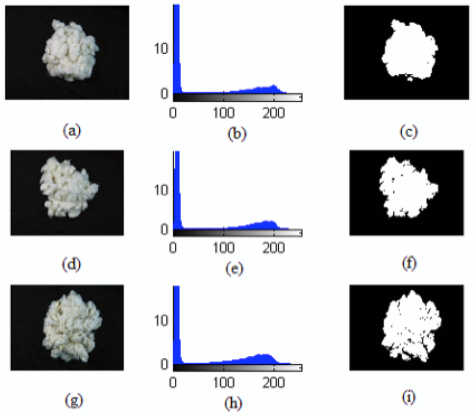

Fig. 1. Seed cotton image acquisition and extraction efficiency.

(a). Seed cotton image sample (degree of 3)
(b). Seed cotton image gray histogram (degree of 3)
(c). Seed cotton image extracted contours (degree of 3)
(d). Seed cotton image sample (degree of 4)
(e). Seed cotton image gray histogram (degree of 4)
(f). Seed cotton image extracted contours (degree of 4)
(g). Seed cotton image sample (degree of 5)
(h). Seed cotton image gray histogram (degree of 5)
(i). Seed cotton image extracted contours (degree of 5)

The image size of seed cotton can be measured by counting the number of its pixels after contour extraction, before that, a split threshold value Th must be set first.

The algorithm for counting seed cotton pixels can be described as the following steps:

978-1-4673-1975-1/12/$31.00 ©2012 IEEE

a) Set the threshold Th, initialize the seed cotton total number of pixels num
b) Search each pixel component points to determine whether the pixel gray value is larger than the threshold, if it is, then $num + 1$, if not, go to step (c)
c) Judge whether the pixel is the last one, if it is, then the search is completed, if not, go to step (b), if the program ends, return the final value.

The seed cotton yellow degrees, magenta degrees, brightness, cyan degrees, hue, saturation and the number of seed cotton pixel ratio feature extraction method is build on the basis of the statics of the number of seed cotton pixel extraction algorithm. Therefore, it's similar to the above one. Given the yellow degree as an example.

After doing contour extraction of the seed cotton image, the number of pixels of seed cotton num is already calculated by the algorithm we described above. Therefore, we only need to calculated the sum of the yellow degree within the contour, we write it as $yellow$. After that we need to seek the proportions of num and $yellow$.

The algorithm for extracting the yellow degree and the number of seed cotton pixel ratio characteristics can be described as the following steps:

a) Set the threshold Th, initialize the total number of yellow degree $yellow$
b) Search each pixel component points to determine whether the pixel gray value gv is larger than the threshold Th, if it is, then $yellow + gv$, if not, go to step (c)
c) Judge whether the pixel is the last one, if it is, then the search is completed, if not, go to step (b), if the program ends, return the value $yellow/num$

For the extraction of the average seed cotton impurities size and number, we taking the following methods: After doing m times seed cotton image dilation, we can calculate the corresponding Euler number $E(m)$, and we are able to calculate the disappear impurities number $f(m)$. See (1)

$$f(m) = E(m) - E(m-1) \tag{1}$$

Gradually increases m, $f(m)$ will tend to be stable. Set when m achieve a relatively stable value $m = M$, then the corresponding impurities number M can be calculate in (2)

$$N = E(0) - E(M) \tag{2}$$

Using (3) to estimate the total number of seed cotton impurities pixels.

$$A = \sum_{k=1}^{M} k^2 * f(k) \tag{3}$$

The above characteristics were divided by the number of pixels of the seed cotton num, and we get yellow degrees, magenta degrees, cyan degrees, brightness, hue, saturation, impurities pixels, the impurity quantity of seed cotton pixel number ratio, all of which compose a feature vector space :

$$(yellow, magenta, cyan, intensity, hue, saturation, impurity\_area, impurity\_number)$$

III. Modeling and Solution Algorithm

A. Grade Classification Least Square Method Model

The general level of the classification model is equivalent to construct a reasonable classification rules f and making the output y_i represents the grade that the sample cotton belongs to. See (4)

$$f(yellow, magenta, cyan, intensity, hue, saturation, impurity\_area, impurity\_number) = y_i \tag{4}$$

Which the input of the classification rules f corresponding to the degree of yellowness, the magenta degrees, cyan degrees, brightness, hue, saturation, impurities pixels, pixel ratio of impurities in the number of seed cotton.

Through training and learning, we get the best classification rule f, the objective function for optimization is shown in (5)

$$min \sum_{j=1}^{m} (f_j(yellow, magenta, cyan, intensity, hue, saturation, impurity\_area, impurity\_number) - y_k)^2 \tag{5}$$

m represents the number of samples, we define sample j belongs to grade k.

y_i represents the output of cotton with grade i, in this paper, we assume that cotton only has grade 3, 4, 5, which means for cotton grade i, the y_i will take 3, 4, 5 orderly.

B. Based on Particle Swarm Optimization for Least Squares Support Vector Machines (PSO-LSSVM) Solution Algorithm

1) LSSVM: Principle of least squares support vector machine (LSSVM) is as follows:

First consider a liner regression program with training samples n, set the training data set $(x_i, y_i), i = 1, 2, \ldots, n$, $x_i \in R^d$ is the input mode of sample i . $y_i \in R$ is corresponding to the expected output of sample i. Then, the linear regression function can be written in (6)

$$y(x) = w^T x + b \tag{6}$$

w for weights, b for bias. By minimizing the target function, LSSVM determines the regression coefficients, see (7)

$$\min J(w, \xi) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{n} \xi_i^2 \tag{7}$$

Subject to:

$$y_i = w^T x_i + b + \xi_i, i = 1, 2, \ldots, n \tag{8}$$

For nonlinear problems, we can use non-linear transformation to transform the question into a high-dimensional space and seeking optimal classification in the in the transformed space. With non-linear mapping $\phi : R^d \rightarrow H$, the input sample is mapping to high-dimensional space H. In feature space H it uses only space dot product to write an optimal hyperplane algorithm, intended to find function K which allows:

$$K(x_i, x) = \phi(x_i)\phi(x) \tag{9}$$

According to the relevant function theory, as long as the conditions of a kernel function $K(x_i, x)$ satisfies Mercer condition, it will be able map to a inner product space in a certain transformation space.

Introducing Lagrange function, calculate each variable's partial derivatives and ordered partial derivatives to zero. By using this method, we turn the solution of the optimization problem into solving linear equations. Solving this problem and we get the least squares support vector machine linear regression model. See (10)

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b \tag{10}$$

2) PSO: Particle Swarm Optimization (PSO) is proposed by United States social psychologist Kennedy and electrical engineers Eberhart in 1995. They came up with a Bionic optimization algorithms. At the beginning of this algorithm, PSO is initialized to a random particles. In each iteration, particles update itself by track two "extremes": the first is the best solution that the particle find by itself, it called a local extremum points (using $pbest$ to represent its location). Another extreme point is a best solution which found by the whole particle group (or it can be called as "particle population"), it called the global extreme points (using $gbest$ to represent its location). After you found the two best solution, particles will update it's own position and speed by using the following formula (11) and (12). The information in particle i can be expressed in using d dimensional vector, the position is $Y_i = (y_{i1}, y_{i2}, \ldots, y_{id})^T$. The speed is $V = (\boldsymbol{v}_{i1}, \boldsymbol{v}_{i2}, \ldots, \boldsymbol{v}_{id})^T$, for the other vectors:

$$\begin{aligned} \boldsymbol{v}_{id}^{k+1} = w \times v_{id}^k + c_1 \times rand_1^k \times (pbest_{id}^k - x_{id}^k) \\ + c_2 \times rand_2^k \times (gbest_{id}^k - x_{id}^k) \end{aligned} \tag{11}$$

$$y_{id}^{k+1} = y_{id}^k + v_{id}^{k+1} \tag{12}$$

Where $\boldsymbol{v}_{id}^k$ is particle i's speed at k times iteration in dimension d; c_1, c_2 is accelerated coefficient; $rand_{1,2}$ is a random number in $[0, 1]$; x_{id}^k is particle i's position at k times iteration in dimension d; $pbest_d$ is particle i's local extremum postion that in dimension d, $gbest_d$ is the whole particle group's global extremum postion that in dimension d, W for initial inertia right value.

3) Algorithm Steps: Based on cross-validation, we use adaptive evolutionary principle in PSO to optimizing LSSVM. The algorithm steps shows below:

a) Do data normalization in the data pre-processing step and divide training set S into n independent subsets;
b) Set the PSO and LSSVM initial parameters.
c) $S_j (j = 1, 2, \ldots, n)$ serve as training samples and the remaining $n - 1$ subsets are training by LSSVM with hyper-parameter $(c_i(t), r_i(t))$. After training, algorithm get testing error of LSSVM to S_i.It will use the average of final testing error as the error of cross examination.
d) Based on the cross examination error, algorithm determines the current individual optimal value and the current group optimal value. Using formula (11) (12) to update new particle's position and speed.
e) Determine whether to achieve the maximum evolution generation. If achieved, then the optimization process stop. Otherwise, go back to step (c).
f) Using the training sample set to construct a LSSVM using (c_{best}, r_{best}) as the hyper-parameter and do predetermination.

C. Based on the distance between the closeness of fuzzy decision method

We can also deal with the output real-number y', by using the output value and determine which grade's value it is close to. And we classified it to this certain grade. Another option is to use fuzzy function to define the degree of belonging to each grade.

In distinguish step, fuzzy nearness degree $\delta(i)$ for grade i is based on the equation of closeness of distance between classes. See (13)

$$\delta(i) = 1 - \frac{|y_i - y'|}{\lambda} \tag{13}$$

Where $\lambda = \max y - \min y$

Compare with $\delta(1), \delta(2), \ldots, \delta(i), \ldots, \delta(7)$, the largest $\delta_{max} = \delta(j)$ is considered to belong to the grade j.

IV. Model Test and its Result Analysis

For the harvested seed cotton sample, we pick 4 samples in each grade as the test sample, and another 48 samples as training samples. In the first step, we are using competitive learning neural network [6]-[13] to train the sample. After training is completed, we test this training model by using the test sample. The result shown in Table I.

Using the least squares support vector machine and the same training and examination of samples, after training is complete, test results as shown in the Table II.

TABLE I

ORIGINAL NEURAL NETWORK TEST RESULT

Sample	Original Grade	Computer Output	Grade Result
1	3	2.8551	3
2	3	2.7883	3
3	3	2.8418	3
4	3	2.7823	3
5	4	3.5780	4
6	4	3.1033	3
7	4	4.4973	4
8	4	4.2271	4
9	5	4.0152	4
10	5	4.8547	5
11	5	4.8786	5
12	5	5.2851	5

TABLE II

LEAST SQUARES SUPPORT VECTOR MACHINES FOR TEST RESULTS

Sample	Original Grade	Computer Output	Grade Result
1	3	3.2608	3
2	3	3.2775	3
3	3	2.9998	3
4	3	3.2173	3
5	4	3.8254	4
6	4	3.5701	4
7	4	3.7906	4
8	4	3.7107	4
9	5	4.2803	4
10	5	4.8903	5
11	5	5.3844	5
12	5	5.3366	5

After the training, the result as shown in Fig. 2, Mean squared error of train data = 0.0548386 (regression); Squared correlation coefficient of train data= 0.709294 (regression); Mean Squared error of test data = 0.0276279 (regression); Squared correlation coefficient of test data = 0.839666 (regression).

For the existing sample, Table I shows that traditional neural network full training test result in the grade 5 sample has a classification precision 75%, grade 4 sample has a classification precision 75%, grade 3 sample has a classification precision 100%, the overall classification precision is 83%. Using the application support vector machine, Table II shows that after full training, the test result in the grade 5 sample has a classification precision 100%, grade 4 sample has a classification precision 75%, grade 3 sample has a classification precision 100%, the overall classification precision is 92%.

V. CONCLUSION

a) We extract the characteristics which reflect of seed cotton yellow degrees, magenta degrees, cyan degrees, brightness, hue, saturation and characteristics of impurity from the harvest seed cotton image.

b) Directly apply the traditional method of neural network, using samples of seed cotton for learning and training, general classification test accuracy is 83%. Tradition method of neural network classification accuracy it is difficult to improve.

c) Based on least-square principle, with the help of support vector machine, the overall classification accuracy enhanced from 83% to 92%, it improves the accuracy of classification.

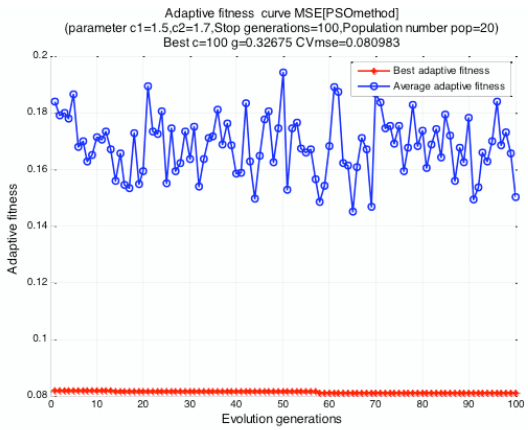

Fig. 2. PSO Support Vector Machine Training

REFERENCES

[1] Xiong Zongwei, Cai Pai, "Cotton quality and its international status in China," Chinese cotton [J],2002/10.

[2] Xiong Zongwei, "Cotton color characteristics," cotton [J],1995/04.

[3] Xiong Zongwei, Xiang Shikang, "China's cotton (cotton-flock) status and improving suggestion of standard," China cotton [J],1998/04.

[4] Xiang s-k, Xiong Zongwei, "On the reform of China's cotton standard," cotton [J],2000/02

[5] Li Ning, Liu Dongbo, Zang Yingming, "Cotton cotton grade standard in China and abroad study on difference of grading standards," Journal of Dalian Institute of light industry [j], the 20th volume 4th, December 2001.

[6] Ling Wang, Ji Changying, Chen Binglin, "Prior to harvest cotton in black background image based on morphological segmentation technology,"Journal of cotton [J],2006,18 (5): 299 303.

[7] Ling Wang, Ji Changying, Chen Binglin, Liu Shanjun, "Seed cotton before the harvest grade based on image feature analysis on clustering fusion,"Journal of crops [J],2007,33 (7): 1162-1167.

[8] Ling Wang, Ji Changying, "Classification of seed cotton grade in field sampling based on machine vision technology model," Chinese agricultural science [J],2007,40 (4): 704-711.

[9] Wang Ling, "Based on machine vision and prior to harvest cotton grade sample classification of multivariate regression model," Nanjing Agricultural University Master's thesis [M],2006.6

[10] Ling Wang, Ji Changying, "Seed cotton in fields of image segmentation based on competitive learning network," Journal of agricultural engineering [J], 24th 10th, October 2008

[11] Yong Wang , Xiaorong Zhu , Changying Ji, "MACHINE VISION BASED COTTON RECOGNITION FOR COTTON HARVESTING ROBOT," Machine Vision Based Cotton Recognition[J]1423

[12] Wang, L. and Ji, C., "BY USING MACHINE VISION RANKING FOR PREHARVEST COTTONS," IFIP International Federation for Information Processing, Volume 259;Computer and Computing Technologies in Agriculture[M], Vol. 2; Daoliang Li; (Boston: Springer),pp. 14651469, 2008.

[13] Wang, L. and Ji, C., "SUMMARY OF PIVOTAL TECHNIQUE OF COTTON-HARVEST ROBOT," in IFIP International Federation for Information Processing, Volume 259;pp. 14591463.Computer and Computing Technologies in Agriculture[M], Vol. 2; Daoliang Li; (Boston: Springer), 2008.