Academic Journal of Engineering and Technology Science
ISSN 2616-5767 Vol.3, Issue 7: 201-209, DOI: 10.25236/AJETS.2020.030720

Design of storage time identification system of lettuce based on Adaboost

Mao Xinyuan

Jiangsu University, 212000, Zhenjiang, China

ABSTRACT. *The storage time of lettuce is an important factor affecting the quality and nutritional value of lettuce. In order to identify the storage time of lettuce quickly, effectively and non-destructively, Adaboost algorithm was used to identify and classify lettuce by near infrared spectroscopy. Taking 100 fresh lettuce samples as the research object, the near infrared diffuse reflectance spectrum of lettuce was detected every 12 hours by Antaris II near infrared spectrum analyzer for three times, and the spectrum scanning wave number ranged from 10000-4000cm^{-1}.Firstly, the standard orthogonal transform (SNV) is used to preprocess the collected data to eliminate the influence of noise, and then the principal component analysis (PCA) is used to reduce the dimension of the 1557-dimensional lettuce near infrared spectrum data. After PCA, the spectral data information after dimension reduction is extracted by LDA algorithm, which improves the accuracy of clustering and achieves the maximum distance between classes and the minimum distance within classes. Then, the weak classifier in Adaboost algorithm is constructed by using K-nearest neighbor rule.After 10 iterations, the strong classifier composed of 10 K-nearest neighbor weak classifiers can achieve classification accuracy higher than 98%, and the classification time is about 3.6s.Experiments show that Adaboost technology combined with dimensionality reduction clustering algorithms such as PCA and LDA provides a new idea for quickly and efficiently identifying the storage time of lettuce.*

KEYWORDS: *AdaBoost; Lettuce; Near infrared light; Storage time*

1. Introduction

The people's living standards are constantly improving in the process of economic development. As an important vegetable on people's table, the freshness of lettuce is an important reference index for people's purchase and selection, and its nutrients will be lost with the extension of storage time [1].Therefore, people urgently need a rapid, accurate and non-destructive technique to detect the storage time of lettuce.

Near-infrared spectral analysis technology is mainly used to study near-infrared spectral analysis technology. Most of the information in organic compounds can be reflected in the near infrared spectrum. Because groups in organic matter have different reflection or absorption for different wavelengths of near-infrared light in

ISSN 2616-5767 Vol.3, Issue 7: 201-209, DOI: 10.25236/AJETS.2020.030720

different environments, near-infrared spectroscopy can be used as a means to detect sample information [2].Yin Shikui et al.[3] , taking typical conifer species in northeast forest areas as the research object, combined with near infrared spectroscopy technology, constructed a near infrared estimation model of wood basic density of Korean pine, larch and spruce fir, and realized the optimization of the wood density estimation model.

Adaboost algorithm, as a boost algorithm, its basic principle is to obtain the weak classifier with the smallest discourse power from the trained weak classifiers by allocating sample weights and discourse power of weak classifiers, and then to obtain a strong classifier by weighting, and even to add weak classifiers continuously, so that the strong classifier can achieve high accuracy. Therefore, it is widely used in various classification scenarios. Cheng Jiabing et al. [4] collected 150 green litchi images, used Adaboost algorithm to construct weak classifier, and finally formed a 16-layer cascade classifier, with the final recognition accuracy of 92.7% and recognition time of 1.27s s.In his research, the integral graph technique was used to analyze the eigenvalues quickly, which reduced a lot of time consumption.

Because the near infrared spectra of lettuce are different after different storage time, the near infrared spectra of lettuce can be detected by machine learning Adaboost to achieve classification. And the dimension of near infrared spectrum data of lettuce collected by near infrared spectrum analyzer is very high, principal component analysis (PCA) and linear discriminant analysis (LDA) are used to reduce the dimension of sample data, and then Adaboost weak classification is constructed by K nearest neighbor rule to identify the storage time of lettuce.

2. Related Classification Algorithm

2.1 Adaboost algorithm

Adaboost algorithm, as a boost algorithm, can effectively improve the classification accuracy by increasing the number of iterations. Different from other algorithms, in this algorithm, the weight of samples will change with iteration, and the weight of wrong samples will increase, while the weight of correctly classified samples will decrease. After updating the sample weight, the weight distribution of the sample will be applied to the training of weak classification $h_t(x)$ and the next sample weight w_i and the discourse power $h_t(x)$ of the weak classifier in the strong classifier a_t will be obtained.Iterate all the way to the end condition.

2.1.1 Adaboost algorithm flow

Step 1: Initialize sample weights: if there are n samples, the initial weight of each sample is *w=1/N*.

$$D_1 = \{w_{11}, w_{12}, \dots, w_{1i}, \dots, w_{1N}\}, w_{1i} = 1/N, i = 1,2, \dots, N$$

ISSN 2616-5767 Vol.3, Issue 7: 201-209, DOI: 10.25236/AJETS.2020.030720

Step 2: Iterate T times to train the weak classifier. $t = 1, 2, ..., T$, And *t* is the number of iterations.

a. Start training from the pair of samples obtained in the first step $D_t(t = 1,2,\dots,T)$, according to the formula

$$e_t = P(h_t(x_n) \neq y_n) = \sum_{n=1}^{N} w_{tn} I(h_t(x_n) \neq y_n) \quad (1)$$

Calculate the error of the weak classifier e_t, and select the one with the smallest error function as the weak classifier $h_t(x), x \to \{-1, +1\}$.

b. Calculate the weight a_t of the weak classifier $h_t(x)$

$$a_t = \frac{1}{2} log \frac{1-e_t}{e_t} \quad (2)$$

c. Update the weight distribution of samples in the sample set for the next iteration, in which the weight of the wrong samples will increase in the last round, while the weight of the correctly classified samples will decrease.

$$D_{t+1} = (, w_{t+1,2}, \dots, w_{t+1,i}, \dots, w_{t+1,N}) \quad (3)$$

$$w_{t+1,i} = \frac{w_{ti}}{Z_t} exp(-a_t y_i h_t(x_i)) \quad Z_t = \sum_{i=1}^{N} w_{ti} \, exp(-a_t y_i h_t(x_i)) \quad (4)$$

In the next iteration, the weight distribution D_{t+1} is used as the sample, and w_{t+1i} is also used as the weight of the updated I-th sample.The product of $y_i \, h_i(x_i)$ is defined in {+1,-1}.Parameter Z_t is introduced as normalization factors, so that the sum of updated sample weights is 1.

Step 3: Combine strong classifiers $H(x)$

$$H(x) = sign(\sum_{t=1}^{T} a_t h_t(x)) \quad (5)$$

Strong classifier $H(x)$ is the weighted sum of weak classifier $h_t(x)$ and corresponding discourse power a_t , and then the result of symbolic function $sign(f)$ operation.

2.1.2 Advantages and disadvantages of AdaBoost

In the derivation of Adaboost algorithm, we can see that Adaboost has the following advantages: (1) High precision;(2) Flexible construction of weak classifiers;(3) Consider the discourse power of each weak classifier. However, Adaboost also has some problems, such as slow training speed and sensitivity to abnormal samples.

ISSN 2616-5767 Vol.3, Issue 7: 201-209, DOI: 10.25236/AJETS.2020.030720

2.2 LDA algorithm and PCA algorithm

The LDA algorithm makes use of the features with the greatest differences among different categories in the high-dimensional features of samples, which make samples of different categories dispersed in the low-dimensional space, while samples of the same category are as close as possible in the low-dimensional space, so as to achieve the maximum ratio between the inter-class dispersion matrix and the intra-class dispersion matrix of samples. Therefore, LDA algorithm can not only reduce the dimension of data, but also facilitate direct classification[5].

The main idea of PCA is to get a new k-dimensional sample set from n-dimensional features, and the features in the new sample set are orthogonal to each other.Generally speaking, the selected k dimension can contain 90% or more of the information in the original data set, so as to achieve the purpose of data dimension reduction.

2.2.1 The difference between LDA and PCA

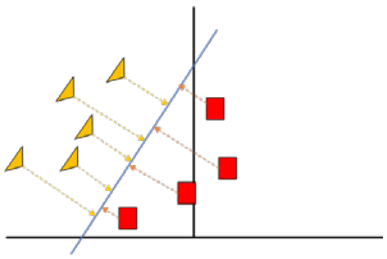

Fig. 1 Two-dimensional PCA projection schematic

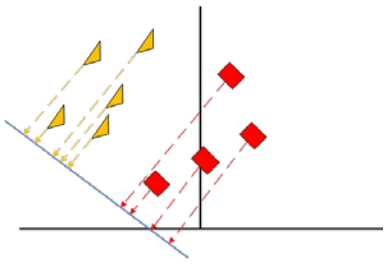

Fig. 2 Two-dimensional LDA projection schematic

Fig.1 is a schematic projection diagram of two-dimensional PCA, which shows that PCA algorithm is only a simple projection of data.Fig.2 is a projection diagram of two-dimensional LDA. Compared with PCA, LDA not only reduces the dimension, but also considers the differences among different categories. It can be seen that LDA can not only reduce the dimension, but also classify it simply [6].However, LDA, as a linear discriminant method, can only be reduced to k-1 dimension at most, while PCA has no limit.

2.3 K nearest neighbor algorithm

K-nearest neighbor algorithm is a supervised algorithm. Its basic principle is to use some distance criterion to get the distance between training samples and test samples, and select the minimum distance to achieve classification. There are two

ISSN 2616-5767 Vol.3, Issue 7: 201-209, DOI: 10.25236/AJETS.2020.030720

key elements in K nearest neighbor algorithm: distance measurement and K value selection.

In K nearest neighbor algorithm, Euclidean distance, Manhattan distance, Mahalanobis distance and Chebyshev distance are generally used to measure the distance between two points in dimensional space. In this paper, Euclidean distance criterion $d_{XY} = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$ is used.

The algorithm K in K nearest neighbor plays a great role. If you choose a small value of K, it is equivalent to classifying near the test sample. In this case, the approximate error of classification will be relatively small, while the estimation error of classification will be very large, resulting in over-fitting. Similarly, when the value of K is larger, the estimation error of classification results will be smaller in this case. But at the same time, due to the influence of distant test samples, the approximate error of classification will be relatively large, resulting in the phenomenon of under-fitting [7].

3. Experiment and Result Analysis

3.1 Data acquisition and preprocessing

In this paper, the near infrared light data of 100 lettuce samples after three different storage times were collected, and each time was regarded as a category, with a total of 300 sample data, and each dimension data had 1557 dimensions.90 of them are taken as training samples, and the other 90 are taken as test samples.

Due to the influence of noise such as solid particle size and surface scattering on lettuce samples during collection, the original data collected contains large noise, so SNV algorithm is used to preprocess the near infrared spectral data of samples to reduce the influence of noise [8].

3.2 Sample dimension reduction

Because the original near infrared spectrum data has 1557 dimensions, which is large and contains a lot of secondary information, this paper first uses PCA algorithm to reduce the 1557-dimensional data to dimension. After testing, in this paper, the value is selected as 10, which can retain more than 99% of the original data information.

After PCA, LDA algorithm is used to further realize supervised dimension reduction, which reduces the test sample data from 10 dimensions to 2 dimensions, at the same time, makes the distance between classes larger and the distance within classes smaller, and improves the classification effect of subsequent K nearest neighbors.

ISSN 2616-5767 Vol.3, Issue 7: 201-209, DOI: 10.25236/AJETS.2020.030720

3.3 Build weak classifier and strong classifier

After dimensionality reduction, the weak classifier is constructed by K nearest neighbor rule. In K nearest neighbor classification, this paper adopts Euclidean distance criterion.

In this paper, there are 10 iterations, and a strong classifier of test samples will be generated by the weak classifier in each iteration. In the $t(t = 1,2,\cdots,10)$ iteration, t k-nearest neighbor weak classifiers are generated, the sample weights of t times are updated, and the discourse power of the weak classifiers of times is obtained. Finally, the strong classifier will be used to classify the test sample set.

Table 1 Classification accuracy and classification time

Times data	1	2	3	4	5	6	7	8	9	10
Test set (%)	86.67	92.21	93.33	95.56	98.89	100	98.89	100	100	100
Classification time (s)	0.404	0.634	0.928	1.272	1.535	1.858	2.184	2.416	2.718	3.104

It can be seen from Table 1 that the classification accuracy of test set can reach 100% when the iteration times exceed 5 times, and it can be seen that the classification time increases approximately linearly with the increase of iteration times. After the fifth iteration, the accuracy is close to 100%, so choosing an appropriate number of iterations can not only reduce a large amount of computation, but also ensure the accuracy.

3.4 Analysis and comparison

3.4.1 Weak classifier and strong classifier

In this study, PCA+LDA+K nearest neighbor is used to construct Adaboost's weak classifier, so the accuracy of classifying test samples with only weak classifier and Adaboost's strong classifier in 10 iterations can be compared.

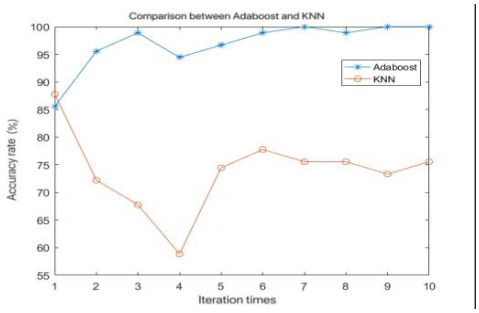

Fig. 3 classification accuracy of AdaBoost and KNN

ISSN 2616-5767 Vol.3, Issue 7: 201-209, DOI: 10.25236/AJETS.2020.030720

Table 2 Classification accuracy of Adaboost and KNN (%)

Number method	1	2	3	4	5	6	7	8	9	10	Average
Adaboost	85.55	95.55	98.88	94.44	96.66	98.88	100	98.88	100	100	96.87
KNN	87.77	72.22	67.77	58.88	74.44	77.77	75.55	75.55	73.33	75.55	73.89

It can be seen from fig. 3 and table 2 that the best classification accuracy is only 79.77%, and the worst case can be as low as 60% in 10 iterations when only weak classifiers with PCA+LDA+K neighbors are used.When Adaboost is used for classification, the best classification accuracy can reach 100%, and the worst case is 87.78%, which tends to 100%.

3.4.2 PCA and LDA

In this paper, PCA+LDA algorithm is used to reduce the dimension of training and test samples. Considering the influence of PCA and LDA on the weak classifier constructed by K nearest neighbor and the strong classifier finally generated.

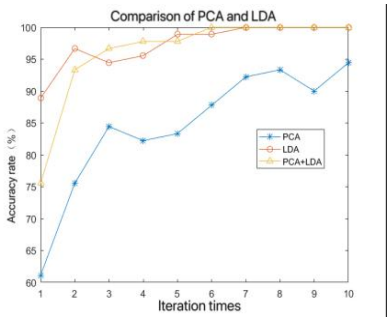

Fig. 4 Comparison of PCA and LDA

Table 3 PCA and LDA Consumption Schedule (S)

Number of times	1	2	3	4	5	6	7	8	9	10
PCA	0.390	0.777	1.249	1.482	1.860	2.202	2.586	2.928	3.355	3.624
LDA	3.644	6.994	10.38	13.85	16.99	19.45	22.59	26.39	29.10	33.21
PCA+LDA	0.418	0.809	1.369	1.765	2.023	2.517	2.941	3.321	4.062	4.303

ISSN 2616-5767 Vol.3, Issue 7: 201-209, DOI: 10.25236/AJETS.2020.030720

It can be seen from fig. 4 that with PCA dimension reduction, LDA dimension reduction and PCA+LDA dimension reduction, the classification accuracy of the final strong classifier will increase with the increase of iteration times. However, there are some gaps between PCA and the other two methods. However, using LDA and PCA+LDA has little difference in effect.

However, it can be seen from table 3 that although LDA is used, the classification effect can be improved. However, when calculating the inter-class scatter matrix and the intra-class scatter matrix, it takes a lot of time because the sample dimension is too high, which is about 10 times as long as the PCA algorithm is used to reduce the dimension, and there is also a small sample problem.

To sum up, PCA used in this paper has the characteristics of low computation but low accuracy, while LDA has the disadvantage of high accuracy but large computation. Therefore, using PCA+LDA algorithm can not only achieve high classification accuracy, but also save a lot of time.

4. Conclusion

The experimental results show that the Adaboost strong classifier composed of K-nearest neighbor classifier as weak classifier can quickly, accurately and non-destructively classify the storage time of lettuce. The experimental data show that the classification accuracy can reach about 98% after five iterations, and the classification time is only about 3.6s even after ten iterations. However, because the sample interval selected in this project is 12 hours, and the time interval is slightly larger, the classification system of samples with smaller time interval needs further study.

References

[1] Hu Yue, Cui Wen, Jin Minfeng, et al. Effects of different cultivation methods on the growth and nutritional quality of lettuce [J]. Journal of Shanghai Normal University (Natural Science Edition), 2019,48(05):566-573.

[2] Parrini S, Acciaioli A, Franci O, et al. Near Infrared Spectroscopy technology for prediction of chemical composition of natural fresh pastures[J]. Journal of applied animal research, 2019, 47(1): 514-520.

[3] Yin Shikui, Feng Guohong, Li Chunxu, et al. Optimization of coniferous wood basic density estimation model based on near infrared spectral band optimization [J]. Journal of Central South University of Forestry and Technology, 2020, (03):85-95.

[4] Cheng Jiabing, Zou Xiangjun, Lin Guichao, et al. Fast detection method of green litchi by cascade classifier based on AdaBoost algorithm [J]. Automation and Information Engineering, 2018,39(05):38-44.

[5] Anna K G E, Amir H E, Andres E. Classification models for heart disease prediction using feature selection and PCA[J]. Elsevier ltd, 2020, 19: 67-74.

[6] Tatiana V N, João C M. Barreira, et al. Phylogenetic insights on the isoflavone

ISSN 2616-5767 Vol.3, Issue 7: 201-209, DOI: 10.25236/AJETS.2020.030720

profile variations in Fabaceae spp: Assessment through PCA and LDA [J]. Food research international, 2015, 76:51-57.
[7] Li W, Chen Y M, Song Y P. Boosted K-nearest neighbor classifiers based on fuzzy granules[J]. Knowledge-Based systems, 2020, 195: 78-90.
[8] Xu Lihua, Xie Deti. Response of prediction accuracy of soil organic matter content to spectral pretreatment and characteristic band [J]. jiangsu journal of agricultural sciences, 2019(06):1340-1345.