Research on Apple Origins Classification Optimization Based on Least Angle Regression in Instance Selection

Bin Li
East China Jiaotong University

Yuqi Wang
East China Jiaotong University

Lisha Li
East China Jiaotong University

Yande Liu (✉️ jxliuyd@163.com)
East China Jiaotong University

Research Article

Keywords: least angle regression, instance selection, classification, SVM

Posted Date: September 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2050870/v1

License: ©️ This work is licensed under a Creative Commons Attribution 4.0 International License.  Read Full License
Abstract

The combined near-infrared spectral analysis technology and machine learning is widely used in fruit quality detection. The train set split from the heavy sample set contains redundant samples, and modeling directly may cause larger errors. Instance selection methods can be used to improve the performance of the raw train set by optimally selecting the effective samples. So, the least angle regression-based instance selection (LARIS) method was proposed to improve the model accuracy of fruits classification in this paper; an experiment was conducted on a total of 952 apples from four origins. Before splitting, the spectral reconstruction methods were used for preprocessing and the 19 outliers were eliminated by statistics. The sample set partitioning based on joint x-y distance (SPXY) was used to split the sample set into the raw train set of 699 samples and the test set of 234 samples. 511 samples with a compression ratio of 26.90% and the random train set with the same compression ratio were built based on training samples. Compared these classifiers trained by three train sets, the model's accuracy established by the optimal train set is 96.6%, which are 4.7% and 6.4% higher than the raw and random train sets. And the average precision and recall of four origins are higher 6% than the raw and random train set. Therefore, the prediction accuracy of apple origins classification model is improved by LARIS. The LARIS method enriches the application and it provides an experimental support for the least angle regression algorithm in instance selection.

1 Introduction

Apple is one of the most exported fruits in China. The biochemical quality depends on growing conditions such as light, soil, climate, and orchard production management ways. Different origins have different conditions, and it is difficult to distinguish just from appearances, so trade in shoddy goods is regular (Li et al., 2022). The near-infrared spectroscopy (NIRS) analysis technology combined with machine learning is used to establish the prediction model for apple origins classification to solve this problem (Eisenstecken et al., 2019; Li et al., 2021).

In the modeling process, the train set is a key to the quality of the model. The more effective the train set is, the more accurate the prediction is, especially the heavy train set (Nawar and Mouazen, 2018). The heavy train set contains some redundant samples, which may cover valuable information, increase the model error, reduce the model performance (Liu et al., 2010; Brodinová et al., 2019), and increase the cost of storage space, calculation, economics even labor. Zhou et al. (2002) had shown that the representative samples in multivariate calibration can be used to improve in some situations. So, it was necessary to extract effective samples to relieve generalization errors and the impact of redundant samples (Sáiz-Abajo et al., 2005; Lucà et al., 2017; Song et al., 2017). This problem was known as instance selection (IS) (Liu et al., 2017). IS can be used to solve the above problems and reduce the storage of the model library, which is also significant in the transmission (Li et al., 2015).

In recent years, with the increasing popularity of heavy datasets, IS algorithms have had new developments (Zhou et al., 2021). In 2003 years, Sang-Woon Kim and B. John Oommen (2003) summarized the IS algorithm into three categories according to the sample distribution, i.e., Condensation (Liu et al., 2017), Edition (Song et al., 2017), and Hybrid. In 2022 years, Soumitra Saha added clustering (Shen et al., 2016; Kordos et al., 2022), boosting (García-Pedrajas and de Haro-García, 2014; Feng et al., 2016), ranking (Cavalcanti and Soares, 2020; de Santana Pereira and Cavalcanti, 2011), adaptive (Djouzi and Amamra, 2022), and other methods supplementing the IS algorithm according to the technical characteristics. They innovated a cluster-oriented instance selection (CIS) algorithm that could control the selection rate. However, all these various algorithms have advantages and disadvantages. No single algorithm dominates the others generally (Saha et al., 2022). The IS algorithms that refer to distance, like Euclidean distance, can be used to improve the consistency between datasets and establish good models. However, they omit the complexity of the boundary, which leads to overlapping border samples being removed and an increased generalization error of the
model for such samples. For example, the border samples have a more significant effect than the internal samples on the hyperplane. Excluding boundary samples directly can change the hyperplane and lead to inaccurate predictions. Seriously, the model established by the train set without the missed border samples at overlapping is meaningless (Liu et al.2017). The overlapping and border sample means that classes of samples are very similar at the spectra.

Based on that, the least angle regression-based instance selection (LARIS) method is proposed. The least angle regression (LAR) is a regular ranking and selecting algorithm, but it has never been used for IS. This method can be used to increase the accuracy and optimize the train set by retaining part of overlapping border samples. The paper structure is as follows: The sample preparation and spectrum acquisition are introduced in section 2; the proposed LARIS algorithm is described in section 3; the experimental results and discusses are shown in section 4, and the paper is concluded in section 5.

2 Samples Preparation And Spectrum Acquisition

A total of uniform shape 952 apples with intact epidermis were prepared and stored at a temperature of 5°C storage cabinet. Among them, No.1 ~ 280, 280 apples from Aksu were Class 1, No.281 ~ 524, 244 Panzhihua apples were Class 2, 228 Luochuan apples were Class 3, No.525 to No.752, and No.753 ~ 952, 200 Yantai apples were the Class 4, all single samples were with a weight of 200 ± 10 g. Before spectra were collected, all samples were cleared, marked four points at the equatorial part evenly, and placed at 25°C under laboratory ambient for 24 hours.

Four spectra of each apple were collected and then the average spectrum was calculated to represent this sample. The spectra were collected by the United States Felix F750 portable near-infrared spectrometer ranging from 729 to 975 nm with the 3 nm resolution. There were 83 variables in total, each marker was automatically scanned seven times by the spectrometer detection port, and the average spectra were outputted from it.

3 Theory And Algorithm

The experiment's steps are as follows. Firstly, the best preprocessing method and outliers are determined by the cross-validation and entire sample classification model (Cavalcanti and Soares 2020). Secondly, three train sets are obtained to analyze the effectiveness and applicability of LARIS: the raw train set is split from the sample set by the SPXY method, and the optimal train set is optimized from the raw train set by the LARIS algorithm, the random train set is selected from the raw train set and its compression ratio is same as the optimal train set. All processes are shown in Fig. 1. The details of the LARIS algorithm, preprocessing, outliers, classifiers, and evaluation are introduced.

3.1 LARIS

The Least Angle Regression-based instance selection (LARIS) is used to calculate the correlation between samples and the standard spectrum of each class by least angle regression apart. The samples with a zero correlation are eliminated. K sets of optimal samples are obtained and used to take a union. This union is used to select the optimal train set whose accuracy is highest. The k classes samples in the raw train set are denoted as the \( X = (X_1, X_2, \ldots, X_k) \), and the \( X_i = (x_{i1}, x_{i2}, \ldots, x_{im}) \) are used to represent the spectra of each class of samples. Where, \( m \) is the sample size of Class \( i \), and \( n \) denotes the total number of samples. Details are as follows.

a) The first step is to calculate the response of each class. Each response is approximated by the mean spectrum, \( \bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k \).
\( L \) denotes the linear space consisting of the selected samples in \( X \). \( S \) denotes the unselected samples in \( X \). \( \hat{X}_i \) is the estimated value of \( X_i \). Assuming \( j \) samples have been added to \( L \), and \( n-j \) samples remain in \( S \).

b) The second step is to select samples from \( X \) by LAR respond with \( X_i \). The total of \( k \) sets of optimal samples \( X'_1, X'_2, \ldots, X'_k \) are obtained when all \( k \) classes are selected in turn.

To obtain the optimal samples set of Class \( i \) by LAR, all raw training samples' correlation \( \beta_1 \) are started with 0. Assuming that, the \( x_j \) is the most correlated sample with the response \( \hat{X}_i \) in \( S \). Let \( x_1 = x_j \). The estimate \( \hat{X}_i \) takes the largest step possible to approach \( \hat{X}_i \) in the direction of sample \( x_1 \) until some other sample, i.e., \( x_{j+1} \), has as much correlation with the current residual \( \hat{X}_i - x_1 \). The correlation of \( x_j \) is updated to \( \beta_1 \). At this point, \( L = \{x_1, x_2\} \), \( \hat{X}_i = \beta_1 x_1 \). Let \( x_2 = x_{j+1} \).

Instead of \( \hat{X}_i \) continuing along \( x_1 \), \( \hat{X}_i \) proceeds in a direction equiangular between \( x_1 \) and \( x_2 \) until a third sample, i.e., \( x_{j+2} \) earns its way into the “most correlated” set. The correlation of \( x_2 \) is updated to \( \beta_2 \). At this point, \( L = \{x_1, x_2, x_3\} \), \( \hat{X}_i = \beta_1 x_1 + \beta_2 x_2 \). The \( \hat{X}_i \) proceeds equiangularly between \( x_1, x_2 \), and \( x_{j+2} \). \( \hat{X}_i \) is approached along the “least angle direction” until a fourth sample enters, etc.

Until there are no samples in \( S \) or the residual value \( \hat{X}_i - \hat{X}_i \) less than \( \varepsilon \), the optimization of Class \( i \) in \( X \) is ended, and the optimal sample set \( X'_i \) is obtained by the samples in \( L \). The correlations in \( S \) samples are still equal to 0.

The \( \hat{X}_i \) and \( \beta_1 \) are defined as:

\[
\left\{ \begin{array}{l}
\hat{X}_i = 0 \\
\hat{X}_i = \sum_{j=1}^{n} \beta_j x_j
\end{array} \right.
\]

1

\[
\beta_i = L (L^T L)^{-1} L^T x_i, x_i \in S
\]

2

\( L \) is added with other class samples when Class \( k \) is optimized, for which these samples can be considered as the overlapping and border samples.

c) The third step is to obtain the union of optimal samples, \( X' = X'_1 \cup X'_2 \cup \ldots \cup X'_k \).

d) The final step is to calculate the response \( X' \) of \( X' \) and run step b) to calculate the correlation \( \beta_1 \) of each sample in \( X' \) by LAR again. The non-zero correlation samples are sorted from largest to smallest according to the absolute value of the \( \beta_1 \). The sub-models are established and determined by ten-fold cross-validation. The training samples corresponding to the most accurate model determine the optimal train set. The first sub-model is established by the first 10k samples initially, while the subsequent samples are entered one after another.

The LAR algorithm is used twice: each class and the union. This operation reduces the sample size, considers the boundaries’ complexity, and eliminates samples with strong covariance while retaining similar samples with overlapping distributions and different types. The algorithm description is shown in Fig. 2.
3.2 Preprocessing and Outliers

In this case, the partial least squares (PLS) spectrum reconstruction was determined to be the best preprocessing method by establishing cross-validation classification models (Chen et al. 2021; Liu and Wang 2021; Huang and Bais 2022), which could be used to extract independent principal components from all variables according to labels. Hotelling’s T-square and Q-residuals statistics are often used for principal component analysis. They are combined with an F-test of a 5% confidence interval to reject large statistical samples. Hotelling’s T-square statistic describes the sample and tests whether the multivariate samples are from the same population. The smaller the statistical value is, the better the model is. The Q-residuals is derived from the error matrix to describe the variables that the classifiers cannot be explained. The Q-residuals of outliers is larger.

3.3 Model

SVM is a popular classifier whose classification boundary is the hyperplane controlled by the training samples. The validity of the samples selected by LARIS can be judged according to the prediction results (Liu et al. 2017). To solve the imbalance problem happening in split and optimization, the imbalance rate IR (Ramentol et al. 2014) is introduced to describe the class distribution of the set.

\[
IR = \frac{N_{\text{max}}}{N_{\text{min}}}
\]

\(N_{\text{max}}\) \(N_{\text{min}}\) is the number of the majority and minority class samples, respectively. The set with IR higher than 1.5 is defined as an unbalanced dataset. The dataset with IR less than 1.5 is defined as the balanced dataset. The prediction may be affected by imbalanced datasets. The classifier combined with the imbalance strategies is necessary to improve the prediction. The decomposition and improvisation are the common imbalance strategies (Li et al. 2022).

3.4 Evaluation

In this case, the specific evaluation metrics of precision, recall, and accuracy are considered to evaluate the models’ performance in each class and overall samples. The precision is interested in measuring the correctness among the total positive labels. The recall is significant for the true positive rate and expresses how the positive class is predicted (Keskes et al. 2022). The ACC is used to measure the percent for the true labels. The ACC_CV is denoted as the accuracy rate of the train set cross-validation. The ACC_P is denoted as the rate of the test set of the model. The equations are expressed as follows.

\[
\begin{align*}
\text{precision} &= \frac{TP}{TP + FP} \\
\text{recall} &= \frac{TP}{TP + FN} \\
\text{ACC} &= \frac{TP + TN}{TP + FP + TN + FN}
\end{align*}
\]


4 Results And Discussions

4.1 Spectral analysis and preprocessing
Figure 3 shows the raw and preprocessing average spectra of each origin. In Fig. 3(a), the spectral shapes of four classes of samples are very similar, i.e., the significant differences in average spectra happen at the 735–755 nm, 915–945 nm, and 955–970 nm wave peaks and troughs, so these positions have some consistent spectral absorption peak. Many overlapping bands and slight differences in absorbance cause difficulties in the sample origins. Thus, they need to be preprocessed. In this case, preprocessing methods are performed the combination of Savitzky-Golay filter(S-G), derivatives, PLS spectral reconstruction combined, and multiple scattering correction (MSC). The most effective preprocessing method is PLS spectral reconstruction and reconstruction combined with MSC. Figure 3(b) shows the average spectra of the four types of samples after reconstruction. After reconstruction, the primary information of the spectrum is in the first 20 latent variables in terms of trend.

Table 1 shows the modeling prediction results of various preprocessing methods after ten-fold cross-validation of the sample set. The classification accuracy of the raw set without preprocessing is the lowest. The results of numerical derivatives applying the S-G filter are as bad as raw. After preprocessing by combining the S-G filter, derivatives, and MSC, the accuracy is improved by about 15%. The 100% accuracy is obtained by the preprocessing method of PLS spectral reconstruction and MSC. The accuracy of PLS reconstruction is exceeded 95%.

<table>
<thead>
<tr>
<th>Pre-processing</th>
<th>ACC</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>precision</td>
<td>recall</td>
<td>precision</td>
<td>recall</td>
</tr>
<tr>
<td>None</td>
<td>55.6%</td>
<td>43.5%</td>
<td>87.5%</td>
<td>69.3%</td>
<td>69.3%</td>
</tr>
<tr>
<td>MSC</td>
<td>65.7%</td>
<td>50.4%</td>
<td>93.9%</td>
<td>93.0%</td>
<td>70.9%</td>
</tr>
<tr>
<td>S-G(0)</td>
<td>56.1%</td>
<td>44.0%</td>
<td>87.5%</td>
<td>69.0%</td>
<td>69.3%</td>
</tr>
<tr>
<td>S-G(0)+MSC</td>
<td>66.5%</td>
<td>51.2%</td>
<td>93.9%</td>
<td>70.9%</td>
<td>93.0%</td>
</tr>
<tr>
<td>S-G(1)</td>
<td>57.5%</td>
<td>46.2%</td>
<td>86.4%</td>
<td>73.4%</td>
<td>62.4%</td>
</tr>
<tr>
<td>S-G(1)+MSC</td>
<td>71.9%</td>
<td>54.8%</td>
<td>95.7%</td>
<td>87.3%</td>
<td>92.2%</td>
</tr>
<tr>
<td>S-G(2)</td>
<td>46.4%</td>
<td>41.5%</td>
<td>75.4%</td>
<td>50.2%</td>
<td>60.7%</td>
</tr>
<tr>
<td>S-G(2)+MSC</td>
<td>64.9%</td>
<td>49.3%</td>
<td>91.4%</td>
<td>93.4%</td>
<td>99.2%</td>
</tr>
<tr>
<td>Reconstruct</td>
<td>96.4%</td>
<td>95.8%</td>
<td>96.8%</td>
<td>94.7%</td>
<td>94.7%</td>
</tr>
<tr>
<td>Reconstruct+MSC</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Each model accuracy was calculated with spectra reconstructed by a particular latent variable. Fig.4 shows the variation in the accuracy of two preprocessing methods with PLS spectral reconstruction. In Fig.4, the accuracy is increased with the first 20 latent variables, raised to maximum, and remained constant when latent variables are greater than 20. After preprocessing with PLS reconstruction with two latent variables, the model’s accuracy is 100%, and its classification effect is the best. However, since this model does not show the superiority of IS, and the impact of subsequent latent variables is not good, the PLS reconstruction with the 83 latent variables factor is determined for preprocessing.

4.2 Outliers

The outliers were eliminated by cross-validation of the sample set through two statistics after preprocessing. Python computing codes were referenced in (Daniel 2018) for these two statistics. The distribution of the samples’ Hotelling’s T-square and Q-residuals estimates is shown in Fig. 5. According to the F-distribution, the critical values of Hotelling’s T-square and Q-residuals are 115.29 and 349.94, respectively. The samples outside the boundaries can be regarded as
abnormal samples for removal, which are marked with orange in Fig. 5. In Fig. 5, a total of 19 samples are eliminated, among which are 2 samples from Class 1, 10 from Class 2, 6 from Class 3, and 1 from Class 4. The sample set contains 933 samples, among which are 278 samples from Class 1, 234 from Class 2, 222 from Class 3, and 199 from Class 4.

4.3 Sample split and train set optimization

After outliers eliminating, the sample set was split into the raw train and test sets by the SPXY method in the ratio of 3:1. The raw train set was optimized by LARIS to establish the optimal train set. To illustrate the quality of the optimal train set, a random train set from the raw with the same compression ratio was established. The details of the raw, optimal, and random train sets are shown in Table 2. The samples size in three train sets are similar. All included the highest of Class 1 samples, the same as Class 2 between Class 3, and the lowest number of fourth-class samples. 699 samples are contained in the raw train set, and LARIS selects 511, the compression ratio is 26.90%. The random train set is selected randomly from the raw. The imbalance rates of the raw and optimal train sets are almost the same and both slightly greater than 1.5, so they are both imbalanced datasets. The random train set is balanced, but the unbalanced strategy did not affect its accuracy (Li et al. 2022). The rest (OVR) decomposition strategy was used to combine with the SVM classifier to establish the model for the unbalanced set (Ramentol et al. 2014; Li et al. 2022).

<table>
<thead>
<tr>
<th>Train set</th>
<th>Class</th>
<th>Train samples</th>
<th>Test samples</th>
<th>IR</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw</td>
<td>210/175/177/137</td>
<td>699</td>
<td>234</td>
<td>1.53</td>
</tr>
<tr>
<td>optimal</td>
<td>155/134/122/100</td>
<td>511</td>
<td>234</td>
<td>1.55</td>
</tr>
<tr>
<td>random</td>
<td>147/127/133/104</td>
<td>511</td>
<td>234</td>
<td>1.41</td>
</tr>
</tbody>
</table>

Figure 6(a) to Fig. 6(h) shows the process and results of the optimization. Figure 6(a), Fig. 6(c), Fig. 6(e) and Fig. 6(g) show the correlation values of all samples when each class is selected apart. The average correlation value and the number of optimal training samples are counted in Fig. 6(b), Fig. 6(d), Fig. 6(f), and Fig. 6(h). Because the testing samples are not involved in the selection, the correlation coefficient of unselected samples and testing samples are all assigned to 0. Statistically, 392, 391, 390, and 392 samples are optimally selected for each class, and 688 samples are obtained after taking a union.

The most significant effect is served on the Class $i$ samples in terms of the size and correlation when Class $i$ is optimized by LARIS. In Fig. 6(a) and Fig. 6(b), 392 samples are selected from the raw train set when optimizing the Class 1 samples, of which 140 samples from Class 1 and other classes less than 100. Class 1 has the largest number both in size and correlation means. The results are consistent with the characteristics optimized by using the Class 1 standard spectrum. In Fig. 6(c) and Fig. 6(d), 391 samples are selected from the raw train set when optimizing Class 2 samples, of which 114 samples from Class 1 are the largest, and the correlation means of Class 4 is the largest. Although Class 2 samples are ranked second on the size and correlation mean, the sum of correlations and standard spectrum of Class 2 is the largest. In Fig. 6(e) and Fig. 6(f), the same of the first three classes’ sizes are counted responding to the Class 3 standard spectrum, but the correlation sum of Class 3 is the largest. It is consistent with the optimization characteristics. In Fig. 6(g) and Fig. 6(h), the number of Class 4 samples is the lowest, but the correlation means of Class 4 is much larger than the other class. In addition, it is found that the Class 4 samples are always least
regardless of any class standard spectrum, which indicates that the fourth-class samples are more different from the rest of the raw train set.

Fig.7(a) shows the accuracy of 688 sorted samples according to the absolute correlation value from largest to smallest. The black line represents the cross-validation accuracy variation of the optimal sample union. The orange line represents the accuracy variation of the test set. The maximum points of the two lines are marked with circles. Since these four classes of samples are used to build ten-fold cross-validation sub-models, the first sub-model is established by the first 40 samples, the first 41 samples establish the second sub-model, and so on. Until the samples in the union are all added for establishing sub-models.

It can be seen from Fig.7(a) that value of $ACC_{CV}$ is maximum when the train set size is 511, and the value of $ACC_P$ is also the maximum simultaneously. Fig.7(b) shows these 511 samples. The value of $ACC_{CV}$ changes in a sawtooth pattern continually, and it decreases and then slowly raises back to the value close to maximum. But it is still less than the maximum. The value of $ACC_P$ slowly decreases in a stepwise pattern. The value of $ACC_P$ is already maximum at the first 268 training samples, but the value of $ACC_{CV}$ is not the maximum at this time. The value of $ACC_{CV}$ still rises, and it indicates that the model is underfitted and not reliable. The prediction results of the three models are shown in Fig.8(a), Fig.8(b), and Fig.8(c) and Table 3.

Fig.8(a) shows the confusion matrix of the model established by the raw train set. There are 19 misclassified samples, of which 18 are came from the first three classes, and only 1 sample of Class 4 is misclassified to Class 1. These 18 samples are misclassified from each other. The Class 4 samples are easily classified by the hyperplanes. Fig.8(b) shows only 8 samples misclassified between the first three classes. The sample of Class 1, which is classified as Class 4, is no long misclassified. The precision and recall rate of each class are increased significantly. Thus, the purpose of improving accuracy with effective training samples is achieved by LARIS. In Fig.8(c), the prediction of the random train set is worse than the raw. The SVM cannot be used to correctly classify the random selected training samples. This comparison has reinforced the conclusion that the quality of the optimal train set is higher than the random train set with the same compression ratio.

In Table 3, the model established by the optimal train set is best whether the whole or each class. The 73.1% of the samples of the raw training samples are selected by LARIS; the accuracy is improved from 91.9–96.6%; the prediction performance of raw and the random train set are followed by the optimal train set. The overlapped and misclassified samples in the first three classes are reduced. Their average precision and recall are improved by about 6% on the optimal. Among them, the misclassified size of Class 1 samples is reduced, and the precision of Class 1 is improved by 7.1%. The recall of Class 2 is improved by 10.2%, and its precision increases the most. Since the Class 4 samples are segmented by hyperplane well, the precision of Class 4 is 100%. The performance of the model established by the random train set is as quite as the raw train set.

Table 3 The results of the three models

<table>
<thead>
<tr>
<th>Train set</th>
<th>ACC_P</th>
<th>Class 1 precision</th>
<th>recall</th>
<th>Class 2 precision</th>
<th>recall</th>
<th>Class 3 precision</th>
<th>recall</th>
<th>Class 4 precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>91.9%</td>
<td>89.9%</td>
<td>91.2%</td>
<td>86.4%</td>
<td>86.4%</td>
<td>93.0%</td>
<td>88.9%</td>
<td>98.4%</td>
<td>100%</td>
</tr>
<tr>
<td>Optimal</td>
<td>96.6%</td>
<td>97.0%</td>
<td>95.6%</td>
<td>91.9%</td>
<td>96.6%</td>
<td>97.7%</td>
<td>93.3%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Random</td>
<td>90.2%</td>
<td>89.6%</td>
<td>88.2%</td>
<td>81.0%</td>
<td>86.4%</td>
<td>90.7%</td>
<td>86.7%</td>
<td>100%</td>
<td>98.4%</td>
</tr>
</tbody>
</table>
The classification boundary of SVM under high-dimension cannot be seen. However, the sample distribution and boundary of two variables can be observed, and the information on two-dimensional variables is incomplete. Figure 9(a)~(f) shows the model's boundary built by three two-dimensional train sets. The distribution of the whole sample set is shown in Fig. 9(a), Fig. 9(c), and Fig. 9(e), while Fig. 9(b), Fig. 9(d), and Fig. 9(f) contain only the test samples. Among them, the round and square points are used to represent the training and testing samples, respectively. The black-bordered points are used to represent the optimal samples. The dark green, purple, cyan, and yellow points are used to separately represent Class 1, Class 2, Class 3, and Class 4. Most of these colorful points are separately clustered in the light blue, dark blue, pink, and red regions. Table 4 shows the prediction results for the three two-dimensional train sets.

In classification, it is hard to correctly classify similar positive and negative samples because those near or far from the boundary are highly similar to the opponent. Such samples are not segmented well by the hyperplane of the SVM classification boundary does not in terms of the geometric distribution. According to Table 4 and Fig. 9(b) and Fig. 9(d), the classification boundary of the optimal train set is worse than the raw. The adjusted part of the boundary does not help improve the accuracy but draws the positive samples at the boundary into the negative region. The red boundary is concave inward, which leads to more yellow points exposed to the pink region and misclassifies as cyan points. Some dark green points are misclassified as purples, which are contained in the invaded region of the light blue. In Fig. 9(d), the biggest change of the classification boundaries is that the pink region is surrounded by the rest of the regions, and the dark blue regions appeared in the region in which the second principal component values are large. The probability of purple points classification is increased there. The distribution of dark blue and red regions in Fig. 9(e) are not changed much from Fig. 9(a), so the misclassified cyan and purple points are classified correctly by subtle adjustments. The model established by the two latent variables spectra is unreliable whether the train set is optimized or not. These spectral primary information amounts are not contained in these spectra.

Table 4. The prediction results for the three two-dimensional train sets

<table>
<thead>
<tr>
<th>Train set</th>
<th>ACC</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>precision</td>
<td>recall</td>
<td>precision</td>
<td>recall</td>
</tr>
<tr>
<td>Raw</td>
<td>60.3%</td>
<td>51.2%</td>
<td>44.9%</td>
<td>53.5%</td>
<td>39.0%</td>
</tr>
<tr>
<td>Optimal</td>
<td>59.0%</td>
<td>53.5%</td>
<td>55.9%</td>
<td>55.0%</td>
<td>55.9%</td>
</tr>
<tr>
<td>Random</td>
<td>62.0%</td>
<td>83.0%</td>
<td>64.7%</td>
<td>55.6%</td>
<td>42.4%</td>
</tr>
</tbody>
</table>

5 Conclusion

In this paper, a new instance selection method based on the least angle regression (LARIS) is proposed. LARIS can be used to select effective training samples from the raw train set, reduce the sample size, and improve the model prediction accuracy. However, this method needs to be used before splitting and after preprocessing and eliminate outliers to avoid affecting the optimal results. In the apple origins classification, 73.1% of the raw training samples were selected by LARIS; the accuracy, precision, and recall of the optimal train set were higher than others; the quality of the train set optimized by LARIS was higher than the raw and random training samples. Overall, under complete primary spectral information, the LARIS method is successfully used in the instance selection of fruits, which can effectively be used to select the representative samples of the train set and eliminate redundant samples to improve model prediction performance.
Declarations

Acknowledgements

Not applicable.

Author contributions

Bin Li, Yu-qi Wang, and Yan-de Liu designed the study, collected the data, performed the analysis, and drafted the manuscript. Li-sha Li supported the collection of data and interpretation of results. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China [31760344].

This work is not available for ethical approval.

The authors declare that they have no competing interests in this paper.

The datasets and materials generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

References


Figures

Figure 1

The flow
Figure 2

LARIS description
Figure 3

the average apple spectrum of four origins
Figure 4
The variation of accuracy in reconstructing

Figure 5
The distribution of Hotelling’s T-square and Q-residuals
Figure 6
The optimization statistics of each class standard spectrum
Figure 7
The results of models established by 688 sorted samples

(a) raw
(b) optimal
(c) random

Figure 8
The confusion matrix of the three models predicting the test set
Figure 9

The boundary of the SVM built by three two-dimensional train sets