Computers and Electronics in Agriculture 145 (2018) 27–34

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Visible-near infrared spectrum-based classification of apple chilling injury on cloud computing platform

Xia Ji'An[a], Yang YuWang[a,*], Cao HongXin[b], Han Chen[a], Ge DaoKuo[b], Zhang WenYu[b]

[a] College of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[b] Institute of Agricultural Information, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China

ARTICLE INFO

Keywords:
Apple
Chilling injury
Visible–near-infrared spectrum
Cloud computing
Classification

ABSTRACT

This paper evaluates the feasibility of applying cloud computing technology for spectrum-based classification of apple chilling injury. The reflectance spectra of Fuji apples with four different levels of chilling injury (none, slight, medium, and severe) were collected. During data processing, the spectra at 400–1000 nm were selected, and first- and second-order-derivative spectral data sets were obtained through integral transformations. Five optimal wavebands were chosen as inputs for the classification models. A cloud computing framework based on Spark and the MLlib machine learning library was used to realize multivariate classification models based on an artificial neural network (ANN) and support vector machine (SVM). The ANN and SVM classification models were used for multivariate classification and analysis of the spectral data sets (raw, first derivative, second derivative) and corresponding optimal wavebands. Of the total data samples, 70% were used for training, while the remaining 30% were used for prediction. The experimental results showed that, by using the cloud computing platform, we could establish an efficient spectrum classification model of apple chilling injury; the ANN model had slightly higher accuracy than the SVM model (not including the second-derivative spectra), but the SVM model was more efficient. Moreover, the classification accuracy using full-waveband spectral data sets was higher than that of data sets using five optimal wavebands. Furthermore, the Spark framework and MLlib were used to implement binary classification models (decision tree and random forest), and these were compared with the multivariate classification model; the binary classification method had better performance in near-infrared spectrum-based classification of apple chilling injury. Finally, we extended the existing spectrum data set to verify the efficiency of the cloud computing platform and desktop PC for handling larger data sets. The results showed that the efficiency of the cloud computing platform was significantly improved by increasing the spectral data set capacity or number of working nodes. Owing to processor and memory limitations, the classification algorithm and model of abundant spectral data sets cannot complete all of the tasks on a desktop PC.

1. Introduction

The apple is one of the five 'healthy' fruits recommended by the Food and Agriculture Organization, and it is ranked first for recommendations. In the 2016–2017 production season, the total global apple yield is expected to be around 77.6 million tons, and the worldwide demand for fresh apples continues to grow (USDA, 2016). Planting and harvesting of apples are affected by many factors, including the growth season, harvest time, and storage conditions.

Around 40 main varieties of apples are planted around the world. Apple trees have stable growth cycles under normal climatic conditions, and apples of different varieties have different harvest times. Some late-maturing varieties are picked in late autumn or winter, and when the weather becomes extremely cold or the temperature declines sharply, apples can incur chilling injuries. During the storage and transportation of applies, the fruit may be stored in a low-temperature and oxygen-deficient environment, and the peel and pulp may become damaged due to the low temperature, low oxygen, and low concentration of carbon dioxide (Lumpkin et al., 2014).

Chilling injury can result in poor sales of apples (Watkins and Nock, 2004). It is difficult to detect and diagnose chilling injuries at an early stage, because as long as the apples remain in a low-temperature environment, they look normal on the outside; chilling injury symptoms only become apparent as the temperature rises (ElMasrya et al., 2009).

To date, many researchers have studied chilling injuries in apples. The research of Watkins and Liu (2010) showed a risk of fruit browning when stored at a temperature below 0 °C, and when apples were stored at a temperature above 3 °C, there is a risk of senescent breakdown. Val

* Corresponding author at: College of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China.
E-mail address: yuwangyang@njust.edu.cn (Y. Yang).

https://doi.org/10.1016/j.compag.2017.12.012
Received 9 June 2017; Received in revised form 4 November 2017; Accepted 11 December 2017
Available online 04 January 2018
0168-1699/

(1) Level 0 (none) (2) Level 1 (slight) (3) Level 2 (medium) (4) Level 3 (severe)

Fig. 1. Chilling injury categories.

et al. (2010) stored apple samples in a cold storage environment (0–4 °C) for 4 months, and used the low O_2 pre-treatment method to reduce the rate of stippen (14%). ElMasrya et al. (2009) used hyperspectral images and an artificial neural network (ANN) to build an analysis model of "Red Delicious" apple chilling injuries. Leisso et al. (2015) studied the influence of fruit peel damage to "Honeycrisp" apples, stored under low-temperature storage conditions, on oxidation resistance and lipid and phenolic metabolism.

Visible–near infrared (VNIR) spectrum technology has been used widely for quality inspection of apples and for natural disaster-related inspections. VNIR can be used to determine the harvest time in accordance with changes in apple pigment and internal components (Bertone et al., 2012), and to predict storage quality (Ignat et al., 2014). VNIR can also be used to analyze common types of damage to apples, such as chilling (ElMasrya et al., 2009), sunburn (Torres et al., 2016), and scratch injuries (Luo et al., 2012).

Cloud computing has an important role in the development of modern agriculture, where it has been a driving factor of progress in the field of precision agriculture (Nguyen et al., 2017). With the wide range of information collected in the agricultural field, and increases in the overall volume of data, high-performance computation based on cloud computing can be used in agriculture and other data-intensive disciplines to collect, save, analyze, mine, and make predictions based on big data information. High-performance computation may also allow for faster and more accurate agricultural management, which could improve decision-making quality, reduce information asymmetry, and increase profits (Woodard, 2016; Lokers et al., 2016). At present, researchers are applying cloud computing technology to the field of precision agriculture, such as for greenhouse environment monitoring and decision-making, intelligent agricultural environment management, and crop performance analysis and recommendations (Vatari et al., 2016; Radu et al., 2016; Jayaraman et al., 2016).

This study was performed to evaluate the application of VNIR spectroscopy and cloud computing to classify apple chilling injury. This study had the following research objectives: (1) classification of the chilling injury of Fuji apple under low-temperature storage conditions; (2) determination of the optimal waveband for apple chilling injury detection; (3) use of cloud computing technology to develop and verify the spectrum classification model for apple chilling injury; and (4) performance analysis of a cloud computing platform. The results of this study should be useful for agricultural product detection and contribute to the agricultural application of cloud computing technology.

2. Materials and methods

2.1. Apple samples

In total, 240 Fuji apples were purchased from a local retailer; these apples were fresh, with no apparent mechanical damage, plant disease, or insect pests. Experiments were conducted at Jiangsu Academy of Agricultural Sciences. First, all apples were stored at 20 °C for 24 h. Then, the apples were divided into two groups: one group consisted of 140 apples that were stored at 0 °C with a relative humidity of 80–90% for 11 weeks, while the other 100 apples were stored at −1 °C with a relative humidity of 80–90% for 14 weeks. Before spectral measurements, all 240 apples were stored at 20 °C for 24 h to observe any symptoms of chilling injury. Then, we collected the external reflectance spectra of all of the apple samples. Classification was conducted in accordance with the degree of damage to the apple surface.

2.2. External visible–near infrared reflectance spectroscopy measurements

The ASD FieldSpec HandHeld^2 portable spectrometer was used to collect external reflectance spectra of the sample apples. The spectra were collected over the wavelength range of 325–1075 nm; the spectral resolution was less than 3 nm and the integral time was 8.5 ms. A circular, white reference plate with a size of 3.6 cm^2 was used for spectral equilibration and calibration. For each apple sample, the reflectance spectrum samples of chilling injured area was scanned twice, and the mean value was calculated. Fig. 1 shows the external condition of apples with different degrees of chilling injury.

2.3. Grading of apple chilling injury

All apples kept under two different storage conditions were used for the classification of apple chilling injury. The apples suffered more severe chilling injury at lower temperature with a longer storage time. During low-temperature storage of fruit, the classification of chilling injury can be performed according to the texture, browning, dryness, and smell of the fruit (Cai et al., 2010; Sun et al., 2017). For example, the harvest time, size, and maturity of an apple all affect the likelihood of chilling injury. Under similar storage conditions, not all apples will present with the same degree of chilling injury at low temperatures; that is, the development of injury varies among individual apples. Thus, in accordance with the texture, taste, and smell of an apple, chilling injury can be classified into four levels: 0 = none, 1 = slight, 2 = medium, and 3 = severe.

In accordance with the classification principles described in Table 1, manual observation and screening of the 240 sample apples were

Table 1
Sensory evaluation of degree of apple chilling injury.

Level	Texture	Taste	Smell	Classification
Level 0	Lustrous and smooth	Crisp and sweet	Obvious fragrance	None
Level 1	Slight wrinkling	Slight rottenness, still edible	Slight fragrance	Slight
Level 2	Partial softening	Partial rottenness, still edible	Slight rotten smell	Medium
Level 3	Soft and not crisp	Soft, not edible	Obvious rotten smell	Severe

conducted. Two groups of researchers were involved: one group was responsible for screening, and the other was responsible for the review and statistics. The samples were divided into four levels: 51 apples were of level 0, 72 were of level 1, 52 were of level 2, and 65 were of level 3.

2.4. Data processing and analysis

The original data collected with the spectrometer contained spectral noise and baseline drift, so the Savitzky-Golay convolution smoothing method was used for denoising and smoothing (Turton, 1992). A five-point convolution smoothing method was used to process the original spectral data. As significant noise existed at the two ends of the spectral interval (325–400 nm and 1000–1075 nm), the 400–1000 nm region was used in this study.

Through integral operations based on the raw spectrum data set of the 400–1000 nm band, we obtained first- and second-derivative spectrum data sets. The three spectrum data sets (raw, first and second derivatives) were used as inputs for the classification algorithm. By combining the first- and second-derivative spectra, we selected the most sensitive wavebands from the apple chilling injury spectrum, and used these optimal wavebands for analysis and research into chilling injury. In this way, the prediction model could be simplified, improving the spectrum collection efficiency and satisfying the requirements for online applications.

2.5. Cloud computing platform and classification model

The cloud computing platform uses a high-performance computer and VMware virtual machine (VM). In the CentOS 6.5 operating system, Spark 2.1 was used as the cluster framework for cloud computing, and for handling the multivariate and binary classification algorithms. Using the cloud computing platform, classification modeling and analysis of the three spectrum data sets, and their corresponding optimal wavebands, were conducted.

We used the MLlib machine learning library provided by the Spark framework to develop two multivariate classification models, namely, an artificial neural network (ANN) and a support vector machine (SVM), and two binary classification models, namely, a decision tree (DT) and random forest (RF). The VNIR of apples with different levels of chilling injury was used as the data input, and different classification models were built on the cloud platform for training and prediction.

3. Results and discussion

3.1. External reflectance spectrum

After maturation, the carotenoid (orange and yellow) and anthocyanin (red and yellow) contents of the apple fruit are the most important factors affecting the color of the peel, which also plays an important role in light absorption. The reflectance spectrum of apple peel contains information on the external characteristics of the peel, such as the color and texture, as well as the water content. Fig. 2 shows the average external reflectance spectrum of apples under different cold storage environments.

For apple chilling injury, all spectrum curves had a similar pattern, and the reflectivity declined with increasing degree of chilling injury. For level 0, as the apple still has bright skin, it shows high reflectivity; with increasing degree of chilling injury, the reflectivity of the chilling area will gradually decline. The reflectivity of apples with severe chilling injury (level 4) shows a sharp decline, because the skin has become soft and developed brown staining.

Nagy et al. (2016) studied “Golden Reinders” apples in terms of the reflectance curve of the peel, and found a similar reflectance spectrum curve. At the blue light waveband (400–480 nm), carotenoids have the highest absorptivity, so the reflectivity is low. At the green light waveband (500–600 nm), because high anthocyanin content of Fuji apples

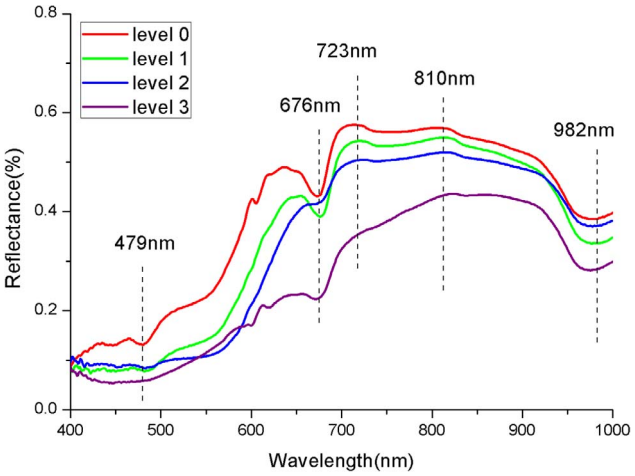

Fig. 2. Average reflectance spectra of apples with different degrees of chilling injury.

have low absorptivity, the reflectivity increases rapidly (Mark et al., 2003). At the 540–650 nm band, after the Fuji apple matures its chlorophyll content declines, whereas the anthocyanin content increases, so there is a wave crest between 600 and 650 nm. A wave crest and trough were apparent in the red light waveband (600–700 nm). Although Nagy et al. found a red edge at 678 ± 30 nm, the curve in this interval did not show a significant change in the present study. This may have been related to the specific apple variety studied (Torres et al., 2016). However, there was a trough between 900 and 970 nm, and the curve of this waveband was related to the water content of the apple.

3.2. Selection of optimum waveband

Our raw spectrum data set contains spectrum information for the visible light and NIR regions, as well as information on many spectrum-insensitive bands. To reduce the data volume and computation time, the raw spectrum curves and derivative spectra were analyzed to find the sensitive bands within the apple chilling injury spectra.

The extreme value in the raw spectrum curve is point 0 in the first-derivative spectrum curve. In the second-derivative spectrum curve, the negative peak is consistent with the peak position in the raw spectrum curve. As shown in Fig. 3, by combining observations from the raw, first-order-derivative, and second-order-derivative spectra, we selected five wavebands, at 479, 676, 723, 810, and 982 nm, as the optimal bands for all three spectrum data sets. Among these bands, 479, 676, and 723 nm are in the blue and red visible light absorption regions. The 810 nm band is the infrared light band and the 982 nm band reflects the water content of the apple (Fan et al., 2009). These five wavebands represent the peak or valley values of the original spectrum wave, which can effectively express the spectrum curve characteristics of apples with different degrees of chilling injury.

3.3. Construction of cloud computing platform

Spark is a new cloud computing framework developed by the Apache Software Foundation. Spark uses a memory distribution data set to optimize the iteration workload for better data mining and machine learning algorithms. Furthermore, Spark provides a machine learning library, MLlib, which is optimized to support machine learning algorithms. In logistic regression computations, its computation speed is 100 times faster than that of Mapreduce in the Hadoop environment (Apache Software Foundation, 2017). Zaharia et al. (2016) showed that the unified programming framework provided by Spark and the large data application engine can effectively support modern workloads and provide users with substantial benefits.

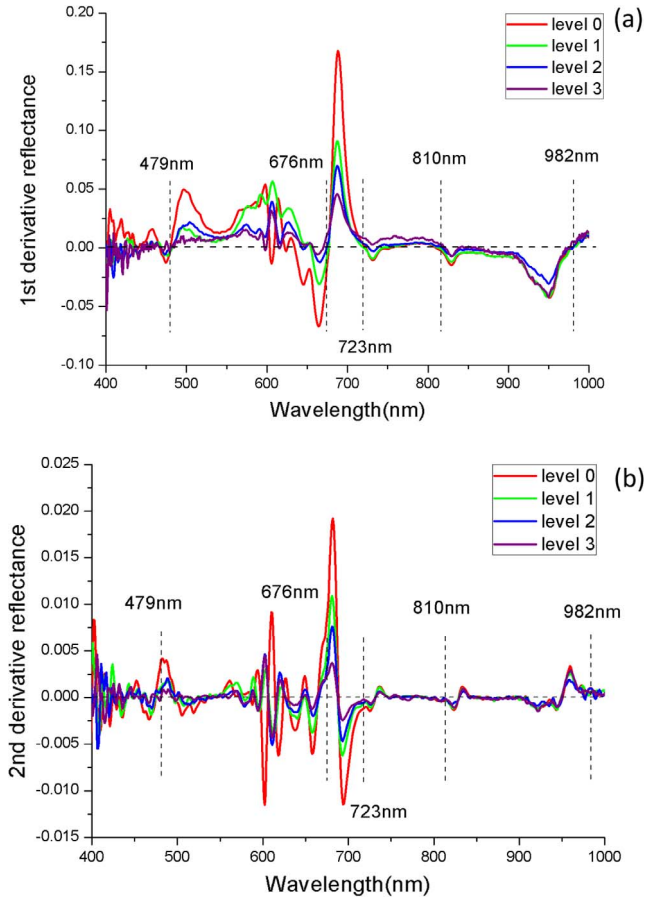

Fig. 3. Derivative reflectance spectra of apples with different degrees of chilling injury. (a) First-derivative spectra. (b) Second-derivative spectra.

Table 2
Spark cluster and node configuration.

Name	Model or edition
CPU	Intel Core i7 6700 K
Memory	Kingston DDR4 2133 MHz (8 GB × 2)
VMware Workstation	10.0.1
Java Development Kit (JDK)	1.8.0_60
Hadoop	2.6.0
Spark	2.1.0
Scala	2.11.8
IntelliJ IDEA	Community Edition 2016.3

Table 3
Experimental running environment.

Role	CPU number	Memory capacity	Storage capacity	Operating system
Master	1	4 GB	50 GB	CentOS 6.5
Node 1	1	3 GB	50 GB	CentOS 6.5
Node 2	1	3 GB	50 GB	CentOS 6.5

In this research, a high-performance computer and VM employing VMware 10.0 were used to support the cloud computing platform; the computer used an Intel i7 processor, 16 GB memory, and the CentOS 6.5 operating system. The Hadoop Distributed File System (HDFS), provided by Hadoop 2.6, was used for storage and management of distributed files. For the cloud computing framework, the integrated IntelliJ IDEA development environment was employed, and the Scala 2.11 program was used to realize the ANN, SVM, DT, and RF machine learning algorithms. The Spark cluster and node configuration were as shown in Table 2. The Spark framework adopts the Standalone running model to design and use one master node for job scheduling and two worker nodes for job execution. Each node was realized by a single VMware VM. The configuration information of each node was as shown in Table 3. In addition, during data processing, we used an in-house program to convert the spectrum data set format to Libsvm format for the input data.

3.4. Multivariate model for classification of apple chilling injury

ANN and SVM-based multivariate classification models were built for full-waveband and optimal-waveband data sets. We randomly chose the spectra of 70% of apple samples (167) to be the training set, and the spectra of the remaining 30% (73) of apple samples were used as the prediction set. The classification model and statistical classification results were established using the cloud computing platform. As the cloud platform would incur different time overheads in task allocation, memory allocation, and job scheduling during each computation, no fewer than five running tests were conducted for each classification model to record the running time and the accuracy of classification prediction. The average values were used to reflect the running time and accuracy of each classification model.

As can be seen in Table 4, the accuracy of the ANN model using the full-waveband data sets was 89.57%, 86.69%, and 72.52% for the raw spectral data set, first-derivative data set, and second-derivative data set, respectively. The raw spectral data set and the first-derivative data set showed similar accuracies, while that of the second-derivative data set was poorer. Similar to the full-waveband results, the classification results for the ANN model using the optimal-waveband data sets were 81.25%, 88.10%, and 66.40% for the raw spectral data set, first-derivative data set, and second-derivative data set, respectively. As can be seen from these results, because the spectra of injury severity levels 2 and 3 were very similar, the classification, training and prediction results for samples of these grades was not good, but the classification results for grades 0 and 3 were better.

SanKaran et al. (2011) used linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), K-nearest neighbor (KNN), and soft independent modeling of class analogy (SIMCA) classification models to analyze and classify citrus leaves infected with Huanglongbing using the derivative spectra. The results showed that the classification accuracy of the disease grade by second-derivative spectra was higher than that of the first-derivative spectra. We used an ANN model to classify the levels of chilling injury using the full and optimal wavebands of the second-derivative spectra, and the accuracies were poor. This may be because the ANN algorithm is sensitive to convergence tolerance; the second-derivative spectral data accuracy was 1E-6, where tolerance values that are too large will affect the classification accuracy, while values that are too small will lead to neural network computations that are difficult to converge.

Using the SVM classification model, as shown in Table 5, the classification accuracies of the full-waveband spectral data sets were 84.69%, 84.07%, and 85.66% for the raw spectral data set, first-derivative data set, and second-derivative data set, respectively. For the raw and first-derivative data sets, the classification results with the ANN model were slightly better than those with the SVM model. With the optimal band data sets, the SVM model classification results were 74.37%, 82.25%, and 80.72% for the raw spectral data set, first-derivative data set, and second-derivative data set, respectively. Thus, the accuracy of the ANN model was better than that of the SVM model using the optimal band data.

In addition, it can be seen that the classification accuracy of the second-derivative spectra using the SVM model was better than that of the ANN model. Thus, the SVM model is more suitable for predicting injury classifications based on higher-order derivative spectra than the ANN model. These results are consistent with the work of SanKaran et al. (2011).

Table 4
ANN classification and prediction results.

Full wavelengths								Optimal wavelengths						
Type	Grades	Prediction (601 input layers, 60 neurons, 4 output layers)						Grades	Prediction (5 input layers, 3 neurons, 4 output layers)					
		Level 0	Level 1	Level 2	Level 3	Accuracy	Duration		Level 0	Level 1	Level 2	Level 3	Accuracy	Duration
Raw	Level 0	15	0	0	0	100%	/	Level 0	15	0	0	0	100%	/
	Level 1	0	17	4	0	80.95%	/	Level 1	2	16	3	0	76.19%	/
	Level 2	0	2	14	1	82.35%	/	Level 2	0	4	10	3	58.82%	/
	Level 3	0	0	1	19	95.00%	/	Level 3	0	1	1	18	90.00%	/
	Total	15	19	19	20	89.57%	16.85 s	Total	17	21	14	21	81.25%	8.83 s
1st	Level 0	15	0	0	0	100%	/	Level 0	15	0	0	0	100%	/
	Level 1	2	16	3	0	76.19%	/	Level 1	0	17	2	2	80.95%	/
	Level 2	0	4	12	1	70.58%	/	Level 2	0	3	13	1	76.47%	/
	Level 3	0	0	0	20	100%	/	Level 3	0	1	0	19	95.00%	/
	Total	17	20	15	21	86.69%	15.34 s	Total	15	21	15	22	88.10%	9.21 s
2nd	Level 0	15	0	0	0	100%	/	Level 0	13	1	1	0	86.66%	/
	Level 1	0	12	5	4	57.14%	/	Level 1	4	13	2	2	61.90%	/
	Level 2	2	1	9	5	52.94%	/	Level 2	2	3	8	4	47.05%	/
	Level 3	1	2	1	16	80.00%	/	Level 3	1	2	3	14	70.00%	/
	Total	18	15	15	25	72.52%	15.49 s	Total	20	19	14	20	66.40%	9.62 s

Compared with the full-waveband classification results, the classification accuracy with the optimal-waveband data sets was lower. This is consistent with the conclusions of Cheng et al. (2015). The main reason for this is that the full-waveband data sets contains all of the spectrum data, whereas the optimal band data sets contain less information, even though they provide the most important information.

On the cloud computing platform, in terms of the time spent on model building, training, and prediction, the SVM algorithm was on average 17% faster than the ANN algorithm, so the SVM algorithm was more suitable for building the spectrum-based apple chilling injury classification model on the cloud computing platform than was the ANN algorithm.

3.5. Model for binary classification of apple chilling injury

Compared with multivariate classification models, binary classification models are preferred by many researchers. To compare the performance and differences between binary and multivariate classification models in the cloud computing environment, we used MLlib to realize decision tree and random forest classification algorithms in the Spark environment. Furthermore, based on the previous spectrum data sets, we combined three of the chilling injury categories (levels 1–3) into one category (level 1). Along with the chilling injury category corresponding to no injury (level 0), this gave two injury categories. In this way, the decision tree and random forest binary classification models could be used for modeling and analysis.

Using the full-waveband data sets, the injury classification accuracy of the decision tree model for the raw spectra was 96.55%, higher than that of the random forest model. However, for classification of the chilling injury levels using derivative spectra, the random forest model was superior. Similar to the full-waveband analysis, with the optimal-waveband data sets, the accuracy of the decision tree model was 98.27% for the raw spectra, which was also higher than that of the random forest model. For the classification of the chilling injury levels using derivative spectra with the optimal-waveband data sets, the random forest model was again better than the decision tree model.

These results showed that the random forest model was more suitable for classification modeling using derivative spectra compared with the decision tree model.

As can be seen from Tables 6 and 7, with both the decision tree and random forest models, the accuracy of the binary classification model was higher than that of the multivariate classification model. This result

Table 5
SVM classification and prediction results.

Full wavelengths								Optimal wavelengths						
Type	Grades	Prediction (Iteration times 100)						Grades	Prediction (Iteration times 20)					
		Level 0	Level 1	Level 2	Level 3	Accuracy	Duration		Level 0	Level 1	Level 2	Level 3	Accuracy	Duration
Raw	Level 0	15	0	0	0	100%	/	Level 0	15	0	0	0	100%	/
	Level 1	0	15	4	2	71.42%	/	Level 1	0	13	5	3	61.90%	/
	Level 2	0	2	14	1	82.35%	/	Level 2	0	2	12	3	70.58%	/
	Level 3	0	1	2	17	85%	/	Level 3	0	5	2	13	65.00%	/
	Total	15	18	20	20	84.69%	13.53 s	Total	15	20	19	19	74.37%	8.02 s
1st	Level 0	15	0	0	0	100%	/	Level 0	15	0	0	0	100%	/
	Level 1	1	18	2	0	85.71%	/	Level 1	2	14	1	4	66.67%	/
	Level 2	0	4	12	1	70.58%	/	Level 2	0	1	14	2	82.35%	/
	Level 3	0	2	2	16	80.00%	/	Level 3	1	2	1	16	80%	/
	Total	16	24	16	17	84.07%	13.40 s	Total	18	17	16	22	82.25%	8.66 s
2nd	Level 0	15	0	0	0	100%	/	Level 0	15	0	0	0	100%	/
	Level 1	0	16	3	2	76.19%	/	Level 1	1	15	3	2	71.42%	/
	Level 2	0	2	13	2	76.47%	/	Level 2	0	1	13	3	76.47%	/
	Level 3	0	1	1	18	90.00%	/	Level 3	0	2	3	15	75.00%	/
	Total	15	19	17	22	85.66%	12.67 s	Total	16	18	19	20	80.72%	8.81 s

Table 6
Decision tree classification and prediction results.

Full wavelengths							Optimal wavelengths			
Type	Grades	Prediction (maxDepth20, maxBins 32) Grades Prediction					(maxDepth 6, maxBins 8)			
		Level 0	Level 1	Accuracy	Duration		Level 0	Level 1	Accuracy	Duration
Raw	Level 0	15	0	100%	/	Level 0	15	0	100%	/
	Level 1	4	54	93.10%	/	Level 1	2	56	96.55%	/
	Total	19	54	96.55%	13.82 s	Total	17	56	98.27%	7.02 s
1st	Level 0	14	1	93.33%	/	Level 0	13	2	86.66%	/
	Level 1	2	56	96.55%	/	Level 1	4	54	93.10%	/
	Total	16	57	94.94%	14.08 s	Total	17	56	89.88%	7.66 s
2nd	Level 0	14	1	93.33%	/	Level 0	12	3	80.00%	/
	Level 1	3	55	94.82%	/	Level 1	6	52	89.65%	/
	Total	17	56	94.07%	13.23 s	Total	18	55	84.82%	7.81 s

is consistent with the report of Sun et al. (2017). They used a similar model for classifying peach chilling injury using hyperspectrum data, and the results showed that the binary classification had higher accuracy than the multivariate classification. Similar results were also obtained in the classification model realized on our cloud computing platform. The binary classification model showed good performance using both the raw and derivative spectra. Similar to the multivariate classification model, the prediction accuracy with the optimal-waveband data sets was not as good as that with the full-waveband data sets, but the modeling costs of the binary classification were reduced.

3.6. Cloud platform performance evaluation

Armbrust et al. (2010) showed that multiple VMs can perform well within a cloud computing framework, but that network and disk sharing can be problematic. The computation time on a cloud computing platform depends on three parameters: the cluster communication time, task scheduling time, and classification modeling time (Yang et al., 2012).

To evaluate the computing performance of the spectrum-based apple chilling injury classification model using a Spark framework, we extended the raw spectral full-waveband data set to 50, 100, 200, 400, and 600 copies, so that the number of samples was 12,000, 24,000, 48,000, 96,000, and 144,000, respectively, and the data set sizes were 104 MB, 209 MB, 419 MB, 839 MB, and 1.22 GB, respectively. The cloud computing platform was run in local, one-node, and two-nodes modes, and we used four kinds of classification algorithm (ANN, SVM, DT, and RF) to test and record the running time.

By extending the spectral data set, we can see from Table 8 that the performance differed between the ANN algorithm and the SVM algorithm when running the classification models in all three modes. Both the ANN and SVM algorithms require extensive iterative computations for classification, training, and prediction. With the increase in size of the spectral data set and number of iterations, the time- and space-related complexity of the classification algorithm increase markedly. At this point, via the distributed parallel computing of the cloud platform and the memory calculations done by Spark, the operating efficiency of the classification algorithms can be obviously improved, where the advantage is clearer for larger amounts of data. In addition, by increasing the number of nodes in the cloud computing platform, the running time of the classification algorithm can be obviously reduced.

Compared with the multivariate classification model, the binary classification model is simpler and requires less iterative computations, so that the cost of building a classification model is reduced. As cluster runs, communication between nodes, and task scheduling in Spark all require a certain amount of time, when we run the binary classification model and the amount of data is small, the Spark parallel computing does not take appreciably less time versus the local single mode. Thus, when the size of the data set is small and the classification model is relatively simple, running the classification algorithm on a cloud platform does not confer any advantage. However, as the amount of data increases, the number of computational nodes and complexity of the algorithm also increase, so that the advantages of parallel processing and distributed storage on the cloud computing platform become more obvious.

To compare the desktop PC and the cloud computing platform for their performance in the classification of chilling injury spectra, we realized the ANN, SVM, DT, and RF algorithms using the neural network toolbox and Libsvm toolbox provided by MATLAB 2012b software on the desktop with a common configuration (Inter i3 processor, 4 GB memory, and 500 GB hard disk), and used extended data for testing.

In cases with a high data volume, it may be necessary to import the

Table 7
Random forest classification and prediction results.

Full wavelengths							Optimal wavelengths			
Type	Grades	Prediction (NumTree 20)					Grades	Prediction (NumTree 5)		
		Level 0	Level 1	Accuracy	Duration		Level 0	Level 1	Accuracy	Duration
Raw	Level 0	14	1	93.33%	/	Level 0	15	0	100%	/
	Level 1	1	57	98.27%	/	Level 1	5	53	91.37%	/
	Total	15	58	95.80%	16.75 s	Total	20	53	95.68%	8.22 s
1st	Level 0	15	0	100%	/	Level 0	13	2	86.66%	/
	Level 1	4	54	93.10%	/	Level 1	3	55	94.82%	/
	Total	19	54	96.55%	17.62 s	Total	16	57	90.74%	8.07 s
2nd	Level 0	15	0	100%	/	Level 0	14	1	93.33%	/
	Level 1	3	55	94.82%	/	Level 1	4	54	93.10%	/
	Total	18	55	97.41%	17.23 s	Total	18	55	93.21%	8.34 s

Table 8
Cloud computing platform running time of the four kinds of classification model.

Mode	Time (unit: s)				Mode	Time (unit: s)			
ANN	Size	Local	1 Node	2 Nodes	DT	Size	Local	1 Node	2 Nodes
	104 MB	83.9	76.1	64.4		104 MB	23.8	25.2	27.3
	209 MB	160.0	134.1	98.3		209 MB	35.3	31.64	30.7
	419 MB	374.8	216.0	171.8		419 MB	51.2	48.2	43.5
	839 MB	640.6	513.2	334.8		839 MB	90.3	78.4	56.8
	1.22 GB	932.0	763.1	554.8		1.22 GB	132.5	98.1	74.6
SVM	104 MB	86.3	78.8	64.3	RF	104 MB	25.2	28.8	29.7
	209 MB	118.1	97.6	85.2		209 MB	38.8	37.7	32.0
	419 MB	145.3	110.8	96.5		419 MB	52.4	50.1	47.2
	839 MB	261.7	210.3	161.2		839 MB	94.9	78.8	64.3
	1.22 GB	285.7	242.4	213.2		1.22 GB	144.8	107.0	79.6

Table 9
Desktop PC running time for the four types of classification model.

Model	Size	Time (s)	Model	Size	Time (s)
ANN	104 MB	458.1	DT	104 MB	127.4
	209 MB	773.4		209 MB	293.6
	419 MB	Out of memory		419 MB	Out of memory
	839 MB	Out of memory		839 MB	Out of memory
	1.22 GB	Out of memory		1.22 GB	Out of memory
SVM	104 MB	342.4	RF	104 MB	288.2
	209 MB	528.1		209 MB	471.7
	419 MB	Out of memory		419 MB	Out of memory
	839 MB	Out of memory		839 MB	Out of memory
	1.22 GB	Out of memory		1.22 GB	Out of memory

data in segments when using MATLAB on a single computer. In addition, as shown in Table 9, during training and testing of a large data set, owing to restrictions of computer performance and memory size, there may be memory overflow and failure to complete classification training when the data volume is larger than 300 MB during classification modeling. Therefore, a standard desktop computer is not suitable for data analysis and data mining of massive spectrum data, and the operation efficiency is much lower than that of the cloud computing platform.

4. Conclusion

In this study, we used a cloud computing platform based on the Spark framework and MLlib machine learning library to conduct a VNIR spectrum-based classification analysis of apple chilling injury. On the cloud computing platform, a multivariate classification model was used for the analysis; the results showed that the ANN classification model offered more accurate performance than the SVM classification model. For higher-derivative spectra, the SVM model was more accurate than the ANN model. Similarly, binary classification models were used for analysis; the random forest model had higher accuracy than the decision tree model for derivative spectra, but the decision tree model showed greater time efficiency. As it has fewer classification eigenvectors, the binary classification model is more suitable for selecting an optimal break point, which can reduce loss of computation function, thus increasing the accuracy of the classification. In addition, binary classification involves a simple principle that does not require iterative computations, so it requires much less computation time versus the ANN and SVM models. Five optimum wavebands (479, 676, 723, 810, and 982 nm) were chosen for the classification modeling, and the ANN and random forest algorithms showed higher accuracies when using these optimal bands. By choosing the optimum wavebands of the spectrum for injury classification, we can effectively reduce the volume of spectrum data and improve the classification efficiency.

Extending the capacity of the apple chilling injury spectra data sets, four classification models were run on the cloud computing platform and a desktop computer. When using the multivariate classification model to classify and make predictions for large spectral data sets, cloud computing technology based on Spark can significantly improve the efficiency of the classification model. Desktop computers are unable to model and classify large amounts of spectral data because of the processor and memory limitations. With continuous increases in the amount of spectral agriculture data, and the number of applications for spectral imaging and hyperspectral technology, spectral data mining and pattern recognition based on cloud computing represents a new direction for research and development in the fields of agriculture.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (Nos. 61640020, 61671244), the Agricultural Innovation Program of Jiangsu, China [Nos. CX(13)3054, CX(14)2114, and CX(16)1006], and the Key Research and Development Program of Jiangsu (BE2016368-1).

References

Apache Software Foundation, 2017. Apache Spark™ is a fast and general engine for large-scale data processing, USA Available at: < http://spark.apache.org/ > (Accessed: 1 February 2017).
Armbrust, M., Fox, A., Griffith, R., 2010. A view of cloud computing. Commun. ACM 53 (4), 50–58.
Bertone, E., Venturello, A., Leardi, R., Geobaldo, F., 2012. Prediction of the optimum harvest time of scarlet apples using DR-UV–vis and NIR spectroscopy. Postharvest Biol. Technol. 69, 15–23.
Cai, Y., Yu, M.L., Xing, H.J., 2010. Effects of low temperature conditioning on chilling injury and quality of cold-stored juicy peach fruit. Trans. CSAE 26 (6), 334–338.
Cheng, J.H., Sun, D.W., Pu, H.B., Zhu, Z.W., 2015. Development of hyperspectral imaging coupled with chemometric analysis to monitor K value for evaluation of chemical spoilage in fish fillets. Food Chem. 185, 245–253.
ElMasrya, G., Wang, N., Vigneault, C., 2009. Detecting chilling injury in Red Delicious apple using hyperspectral imaging and neural networks. Postharvest Biol. Technol. 52, 1–8.
Fan, G.Q., Zha, J.W., Du, R., et al., 2009. Determination of soluble solids and firmness of apples by Vis/NIR transmittance. J. Food Eng. 93, 416–420.
Ignat, T., Lurie, S., Nyasordzi, J., Ostrovsk, Y.V., Egozi, H., Hoffman, A., 2014. Forecast of apple internal quality indices at harvest and during storage by VIS-NIR spectroscopy. Food Bioprocess Technol. 10 (7), 2951–2961.
Jayaraman, P.P., Yavari, A., Dimitrios, G., 2016. Internet of things platform for smart farming: experiences and lessons learnt. Sensors 16 (11), 2–17.
Leisso, R.S., Buchanan, D.A., Lee, J., 2015. Chilling-related cell damage of apple (Malus × domestica Borkh) fruit cortical tissue impacts antioxidant, lipid and phenolic metabolism. Physiol. Plantarum 153 (2), 204–220.
Lokers, R., Knapen, R., Janssen, S., 2016. Analysis of Big Data technologies for use in agro-environmental science. Environ. Modell. Software 84, 494–504.
Lumpkin, C., Fellman, J.K., Rudell, D.R., 2014. 'Scarlett Spur Red Delicious' apple volatile production accompanying physiological disorder development during low pO_2 controlled atmosphere storage. Agric. Food Chem. 62 (7), 1741–1754.
Luo, X., Takahashi, T., Kyo, K., 2012. Wavelength selection in VIS/NIR spectra for detection of bruises on apples by ROC analysis. J. Food Eng. 109 (3), 457–466.
Mark, N.M., Alexei, E.S., Anatoly, A., Gitelson, 2003. Reflectance spectral features and non-destructive estimation of chlorophyll, carotenoid and anthocyanin content in apple fruit. Postharvest Biol. Technol. 27, 197–211.
Nagy, A., Riczu, P., Tamas, J., 2016. Spectral evaluation of apple fruit ripening and pigment contentalteration. Sci. Horticulturae 201, 256–264.

Nguyen, V.Q., Nguyen, S.N., Kim, K., 2017. Design of a platform for collecting and analyzing agricultural big data. J. Digital Contents Soc. 18 (1), 149–158.
Radu, C., Apostol, E., Leordeanu, C., 2016. Integrated cloud framework for farm management. In: 10th International Conference on Complex, Intelligent, and Software Intensive Systems, Fukuoka, Japan, 06–08 JUL. pp. 302–307.
Sankaran, S., Mishra, A., Maja, J.M., Ehsani, R., 2011. Visible-near infrared spectroscopy for detection of Huanglongbing in citrus orchards. Comput. Electron. Agric. 77 (2), 127–134.
Sun, Y., Gu, X.Z., Sun, K., 2017. Hyperspectral reflectance imaging combined with chemometrics and successive projections algorithm for chilling injury classification in peaches. LWT-Food Sci. Technol. 75, 557–564.
Torres, C.A., León, L., Javier, S.C., 2016. Spectral fingerprints during sun injury development on the tree in Granny Smith apples: a potential non-destructive prediction tool during the growing season. Sci. Horticulturae 209, 165–172.
Turton, B., 1992. Novel variant of the Savitzky-Golay filter for spectroscopic applications. Measur. Sci. Technol. 3 (9), 858–863.
Val, J., Fernandez, V., Lopez, P., 2010. Low oxygen treatment prior to cold storage decreases the incidence of bitter pit in 'Golden Reinders' apples. J. Sci. Food Agric. 90 (3), 536–540.
Watkins, C., Nock, J.F., 2004. Smart Fresh TM (1-MCP)—the good and bad as we head into the 2004 season. New York Fruit Quart. 12 (3), 1–26.
Watkins, C., Liu, F.W., 2010. Temperature and carbon dioxide interactions on quality of controlled atmosphere-stored 'Empire' apples. HortScience 45 (11), 1707–1712.
Woodard, J., 2016. Big data and Ag-analytics an open source, open data platform for agricultural & environmental finance, insurance, and risk. Agric. Finance Rev. 76 (1), 15–26.
Vatari, S., Bakshi, A., Thakur, T, 2016. Green House by using IOT and Cloud computing. In: IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology (RTEICT), Bengaluru, India, 20–21 May, pp. 246–250.
United States Department of Agriculture (USDA) Foreign agricultural service, 2016. Fresh Deciduous Fruit: World Markets and Trade (Apples, Grapes, & Pears), USA. Available at: < https://apps.fas.usda.gov/psdonline/circulars/fruit.pdf > . (Accessed: 1 March 2017).
Yang, H.H., Du, L.L., Li, L.Q., 2012. Parallel PLS algorithm using map reduce and its application in spectral modeling. Spectrosc. Spectral Anal. 32 (9), 2399–2403.
Zaharia, M., Xin, R.S., Wendell, P., 2016. Apache spark: a unified engine for big data processing. Commun. ACM 59 (11), 56–65.