Identification Rice Varieties Based on K-means Clustering Algorithm and BP Neural Network

Shu-fang QIN, Chang-hua LIU* and Shen-ao HUANG
School of Mathematics & Computer Science, Wuhan Polytechnic University, Wuhan, China

*Corresponding author

Keywords: Identification, Rice varieties, K-means, BP neural network.

Abstract. BP neural network is characterized by adaptability and real-time learning, so it is used widely in classification. While the more complex samples classify, the lower the accuracy of BP neural network is. So a method was purposed to identify rice varieties that combined K-means clustering algorithm with BP neural network. The gray-scale mean value, aspect ratio and circularity, which were the three parameters that expressed rice, were extracted by image processing. The K-means clustering algorithm was used to classify the data based on the aforementioned three parameters and the classification result was entered into the BP neural network and trained to get the classifier. The overall results indicate that the method mentioned is more effective than using K-means clustering algorithm or BP neural network singly and the accuracy is up to 80%. Experimental results show that combination of K-means clustering algorithm and BP neural network is feasible for identifying rice varieties.

Introduction

Rice is one of the grain crops in China. With the improvement of life standard, the demand for rice varieties and rice quality become more and more high. So how to identify rice varieties rapidly and accurately has significant meaning to common eating. Traditional rice varieties identification usually adopts artificial method, which mainly use observation and feel after sampling, and which is time-consuming, laborious, and subjective, low accuracy. As used on satellite remote sensing and medical widely, machine vision is also imported in the agriculture. It is successful to identify rice varieties taking spectral information of near-infrared images and smell so on. However, high-cost and complex the device makes larger-scale utilization these research difficult[1~3]. In this paper, a efficient nondestructive identification method for rice varieties is purposed based on image processing and analysis technique, which only need simple and low-cost device.

K-means Clustering Algorithm

Clustering refers to gathering same or alike data and divides different data into several categories. Clustering is a technology that can find inherent connection on data. Clustering follows the idea that there is the maximum similarity in the same sample set, and there is the maximum diversity in different sets. K-means clustering algorithm is the most famous clustering algorithm and it is the most popular clustering algorithm due to its simplicity and effectiveness[4,5].

Supposing the sample set is \( X = \{X_i| i = 1, 2, \cdots, N\} \), considering \( K \) categories as \( C_j (j = 1, 2, \cdots, K) \). \( K \) objects are chose randomly from the sample set as initial clustering center. Euclidean distance between every sample and all clustering center is calculated:

\[
d(x_i, x_j) = \sqrt{(x_i - x_j)^T (x_i - x_j)}.
\]

Every object is distributed to the nearest category from the clustering center, new clustering center of every category is calculated:
\[ A_i = \frac{1}{N_i} \sum_{x \in C_j} x \]  

New clustering center and cluster category are calculated again until clustering center no longer changes or the number of iteration reaches the default maximal number of iteration.

K-means clustering algorithm have an good effect on clustering with less sample, but will cause misclassification with more category sample. In tradition, the initial clustering center is chose randomly which decides that the effect on clustering varies from clustering time to time. Meanwhile Euclidean distance equation will influence the effect on clustering due to treat all data indiscriminately.

**BP(Back Propagation) Neural Network**

Artificial neural network imitates human’s brain to achieve some function. BP neural network is put forward by Rumelha the McClelland and other scientists in 1986. BP neural network is a multi-layer feedforward network and trained by error back propagation algorithm, adaptability and real-time learning. Not only can BP neural network learn and store a large number of mapping relationship between input and output patterns, but also it can calculate the output in term of the input[6~8]. So it is widely used in pattern recognition and classification.

The structure of the three-layer BP neural network is shown in Figure 1 and it includes an input layer, a hidden layer and an output layer, which are composed of \( n \), \( p \) and \( q \) neurons respectively. Input of hidden layer neurons are calculated:

\[ b_j = f\left( \sum_{i=1}^{N} \alpha_{ij} x_i - \theta \right) \]  \hspace{1cm} (3)

Input of output layer neurons are calculated:

\[ c_i = f\left( \sum_{j=1}^{p} v_{ij} b_j - r_i \right) \]  \hspace{1cm} (4)

where, \( \alpha_{ij} \) represents connection weights between the \( i \)th input neuron and the \( j \)th hidden neuron ; \( v_{ij} \) represents connection weights between the \( j \)th hidden neuron and the \( t \)th output neuron; \( \theta \) and \( r_i \) represents corresponding neural threshold; \( f(x) \) represents neural activation function, which is calculated:

\[ f(x) = \frac{1}{1+e^{-x}} \]  \hspace{1cm} (5)
\(y_t\) represents expect output value of the \(t\)th neural network; \(c_t\) represents actual output value of the \(t\)th neural network. When total number of training sample is \(K\), the global output error of the network is calculated by the minimum variance and defined:

\[
\varepsilon = \frac{1}{K} \sum_{k=1}^{K} \sum_{t=1}^{T} (y_t - c_t)^2
\]

(6)

Connection weights is adapted in term of backing propagation gradient descent method:

1) The value of adjustment \(d^k_t\) is calculated by Eq.7, which is connection weights \(v_{ij}\) between hidden layer neurons and output layer neurons:

\[
d^k_t = (y^k_t - c_t) \cdot c_t (1 - c_t) (t = 1, 2, ..., q)
\]

(7)

New connection weights and threshold are calculated between next hidden layer neurons and output layer neurons using \(d^k_t\), \(v_{ij}\) and \(r_t\):

\[
v_{ij}(N) = v_{ij}(N-1) + \alpha \cdot d^k_t \cdot b_j
\]

(8)

\[
r(N) = r(N-1) + \alpha \cdot d^k_t
\]

(9)

2) The value of adjustment \(e^j_i\) of connection weights between input layer neurons and hidden layer neurons is calculated:

\[
e^j_i = \sum_{t=1}^{T} (d_t \cdot v_{ij}) \cdot b_j (1 - b_j) (j = 1, 2, ..., p)
\]

(10)

New connection weights and threshold is calculated between next input layer neurons and hidden layer neurons using \(e^j_i\), \(a^k_i\) and \(\omega_{ih}\):

\[
\omega_{ij}(N) = \beta \cdot e^j_i \cdot a^k_i + \omega_{ij}(N-1) (i = 1, 2, ..., n)
\]

(11)

\[
\theta(N) = \beta \cdot e^j_i + \theta(N-1)
\]

(12)

where, \(\alpha\) and \(\beta\) is coefficient of learning; \(N\) represents the \(N\)th learning; \(x^k_i\) is the \(K\)th sample inputted to network; \(i\) represents input value of \(i\)th neuron; \(n\), \(p\) and \(q\) represents the number of input layer neurons, hidden layer neurons and output layer neurons respectively.

Corresponding connection weights is adjusted using Eq.8 and Eq.11 after each sample learning until \(K\) sample learning finished, then the global output error function is decided whether it reaches the default convergence limit value or not. Error function reaching the limit value and the time of network training reached the maximum time of learning, network training will finish.

BP neural network is used widely in classification, but BP neural network still have a certain of limitations. The category of sample is too much and the number of sample is too less to classify sample accurately using BP neural network. BP neural network get into the local minimum of the error performance function easily during the training, which causes terrible result of training. To solve these problem, increasing number of neurons and network layer are traditional treasures, which also make the network more complexity.

**Combination of K-means Algorithm and BP Neural Network**

The classification method combing K-means clustering algorithm and BP neural network is purposed and the model is shown in Figure 2.
Images of 11 kinds of rice are inputted and three characteristic parameter are extracted, responding to gray-scale mean value, aspect ratio and circularity. Since rice mainly include three classes- indica rice, rice and glutinous rice, date are divided into three class using K-means clustering algorithm. Then data are inputted into BP neural network to be trained, then get the classifier. Testing data are inputted into classifier trained, then classification accuracy rate is outputted.

**Result and Analysis**

**Experiment Data**

Rice is divided into two type including indica rice and rice in plant scientific research[9], but rice always is divided into three type including indica rice, rice and glutinous rice in term of appearance and quality in real life.

This paper chooses 11 kinds of rice as research objects, including three kinds of indica rice from Hubei province, one kind of indica rice from Thailand, one kind of indica glutinous rice from Hubei province, five kinds of rice and one kind of Japonica glutinous rice from the north-east, according to International classification standard of rice. 50 grains of rice of every types sample--the total is 550--is chose as experiment sample, whose gray-scale mean value, aspect ratio and circularity three characteristic parameter are extracted using the selected feature parameter extraction method as input data. Indica rice is numbered from 1 to 200; rice is numbered from 201 to 450; glutinous rice is numbered from 451 to 550. Every type of rice is numbered and 11 labels represent 11 kinds of rice simple respectively. Among them from 1 to 4 represent indica rice; from 5 to 9 represent rice; 10 and 11 represent glutinous rice, which all serve as feature parameter extraction.

**Experiment Result**

550 grains of rice serve as input data for K-means clustering algorithm; 30 grains of every simple serve as training simple and the rest serve as testing simple for BP neural network. Classification accuracy rate of every kind of rice is counted and contrasted using K-means clustering algorithm, BP neural network and method purposed in this paper respectively to verify the performance of the proposed algorithm. The result is shown in Table.1.

Classification accuracy rate is calculated:

\[
rate = \frac{\text{sum}(la == labels)}{\text{sum}(la)} \times 100\%
\]

where \(la\) represents label of input data, \(labels\) represents label of output data, \(\text{sum}\) counts.
Table 1. Comparison result of classification effect on every kind of rice simple.

<table>
<thead>
<tr>
<th>type</th>
<th>number</th>
<th>K-means algorithm actual type</th>
<th>accuracy rate</th>
<th>BP neural network actual type</th>
<th>accuracy rate</th>
<th>method of this paper actual type</th>
<th>accuracy rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1~50</td>
<td>1, 2, 4</td>
<td>62%</td>
<td>1, 2, 3</td>
<td>92%</td>
<td>1</td>
<td>100%</td>
</tr>
<tr>
<td>2</td>
<td>51~100</td>
<td>1, 2, 3</td>
<td>15%</td>
<td>2, 3</td>
<td>30%</td>
<td>2, 3</td>
<td>40%</td>
</tr>
<tr>
<td>3</td>
<td>101~150</td>
<td>2, 3, 4</td>
<td>80%</td>
<td>3</td>
<td>100%</td>
<td>3</td>
<td>100%</td>
</tr>
<tr>
<td>4</td>
<td>151~200</td>
<td>1, 3, 4</td>
<td>42%</td>
<td>3, 4</td>
<td>5%</td>
<td>3, 4</td>
<td>35%</td>
</tr>
<tr>
<td>5</td>
<td>201~250</td>
<td>5</td>
<td>100%</td>
<td>4, 5, 6</td>
<td>54%</td>
<td>5</td>
<td>100%</td>
</tr>
<tr>
<td>6</td>
<td>251~300</td>
<td>6, 7, 9</td>
<td>80%</td>
<td>6, 7</td>
<td>15%</td>
<td>6, 7</td>
<td>62%</td>
</tr>
<tr>
<td>7</td>
<td>301~350</td>
<td>7, 9</td>
<td>84%</td>
<td>7</td>
<td>100%</td>
<td>6, 7</td>
<td>70%</td>
</tr>
<tr>
<td>8</td>
<td>351~400</td>
<td>8, 9</td>
<td>84%</td>
<td>7, 8</td>
<td>15%</td>
<td>7, 8</td>
<td>82%</td>
</tr>
<tr>
<td>9</td>
<td>401~450</td>
<td>6, 8, 9</td>
<td>64%</td>
<td>7, 9</td>
<td>15%</td>
<td>9</td>
<td>100%</td>
</tr>
<tr>
<td>10</td>
<td>451~500</td>
<td>10, 11</td>
<td>64%</td>
<td>10</td>
<td>100%</td>
<td>10</td>
<td>100%</td>
</tr>
<tr>
<td>11</td>
<td>501~550</td>
<td>11, 10</td>
<td>40%</td>
<td>11, 10</td>
<td>96%</td>
<td>11</td>
<td>100%</td>
</tr>
<tr>
<td>total classification accuracy rate</td>
<td>——</td>
<td>65%</td>
<td>——</td>
<td>56.55%</td>
<td>——</td>
<td>80.64%</td>
<td>——</td>
</tr>
</tbody>
</table>

It is clear that three main class, indica rice, rice and glutinous rice, are not distinguished in each other when 11 kinds of simple is classified using K-means algorithm. However classification is failure easily among 4 kinds of indica rice, among 5 kinds of rice and among 2 kinds of glutinous rice and most classification accuracy rate of every type of simple is lower than 80% and total accuracy rate is 65%. 4 kinds of indica rice is always regarded as 3th class and 5 kinds of rice is always regarded as 7th class during the classification using BP neural network. And accuracy rate of some simple is up to 100%, but accuracy rate of most simple are lower than 60% and total accuracy rate is 56.55%. While the classification result using method purposed in this paper shows that accuracy rate of some simple is lower than 35%, but most simple is up to 80% and total accuracy is 80.64%. Above all, the method of this paper can improve the classification accuracy obviously and has good robustness.

Summary

This paper reported a method combing traditional K-means algorithm and BP neural network for identification of rice, which overcame the limitation that more sample types and low classification accuracy. The result showed that some classification accuracy rate is low, but most classification accuracy rate is up to 80% and total classification accuracy reached 80.636363%. While total classification accuracy rate of K-means algorithm is 65% and of BP neural network is 56.55%. It can be seen that classification accuracy rate is the most high and have good robustness using method of this paper.

Acknowledgement

This research was in part supported by grands from the Hubei Provincial Department of Education Science and Technology Project(NO.B2015001)

References


256


