# Analysis of DNA molecular genetics features in infected styphnolobium japonicum

Hu Zhongyi

Ningbo City College of Vocational Technology,Ningbo,Zhejiang 315502 China

**Keywords:** DNA image; infected styphnolobium japonicum, identification rate

**Abstract.** During the study of DNA image recognition at the beginning of infected styphnolobium japonicum, the features of DNA profile are arbitrary. Both characteristic angle and direction change are disorderedly distributed. This study presents a new method to examine the features of early DNA profile based on gray-level co-occurrence matrix (GLCM) and fuzzy mean clustering. Different types of lesion characteristics are distinguished by learning sets and fuzzy means clustering methods. Multilayer cascade classifier is used to extract features of DNA image, and achieve segmentation feature recognition. Simulation results show that the improved method has the efficient ability to evaluate the DNA image with a high rate of infection feature recognition.

## Introduction

Infection is one of the factors causing the death of styphnolobium japonicum with external skin ulcers and gangrene. With the development of biological and plant medical imaging equipment, image diagnosis has important value for improving plant healthcare standards. Usually, the identification of diseases are through naked eye observation. Wrong analysis will be made due the lack of experience, fatigue etc. Therefore, it is of great importance to have the diagnosis based on computer technology [1, 2].

Recently, DNA profile feature is a new method for diagnosis of infected styphnolobium japonicum. Effective diagnosis can be made by abnormal DNA profile. At the initial stage of infection, special angles and directions exhibits like multi-level disorders. Traditional Gaussian model segmentation method for multi-division does not have the ability to learn. It is limited the identification of DNA lesions cannot be realized by only a single layer superimposed [3, 4, 5].

## Principle of DNA profile characteristic recognition for early infection

The early infection can be identified by the altered profile of DNA profile. Normal characters of DNA profile are regularly distributed, while infection causes altered features of texture image such as blurting, dryness, deepened color and deformation (Figure 1).

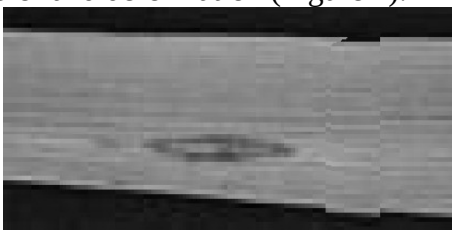

Figure 1 Altered DNA profile due to infection

Comprehensive description of DNA profile of the infected tree in the direction, intervals, and change range, is an effective way to portray infected texture information. GLCM shows the intervals (Δx,Δy) information of angle and texture. If the gray level of image is R, GLCM is R×R matrix described by M(Δx,Δy)(h,k). $m_{hk}$ stands for grayscale h. If s represents the target range A with a specific area of pixels on the set, the co-occurrence matrix CM as formula (1):

$$CM(g_1,g_2)=\frac{\#\left\{\left[(x_1,y_1),(x_2,y_2)\right]\in S\middle|f(x_1,y_1)=g\middle|\&f(x_2,y_2)=g_2\middle|\right\}}{\#S} \quad (1)$$

 

After the DNA profile GLCM matrix is got, a variety of statistical characteristics is defined. These characteristics including the moment of inertia constitute texture, energy, entropy, correlation, and uniformity, can be used to calculate the similar amount COR between DNA image and disease characteristics.

$$COR = \frac{1}{\sigma_x \sigma_y}\left[\sum_h \sum_k hkm_{hk} - \mu_x \mu_y\right] \qquad (2)$$

$\mu_x, \mu_y$ is the average of $m_x, m_y$, $\sigma_x, \sigma_y$ is the standard value of $h_j(x) = \begin{cases} 1, p_j f_j(x) < p_j \theta_j \\ 0, other \end{cases}$. $m_x = \sum_k m_{hk}$ is the elements summation of matrix M. A similar amount represents the similar degree between rows and columns of elements in the matrix, which is a measure of the gray linear relationship. According to similarity, threshold is set to complete lesion feature detection.

**DNA identification texture segmentation**

**DNA imagetraining of infected styphnolobium japonicum based on cascade classifier**

In order to solve the problem of the traditional method, a cascade classifier is used:

$$h_j(x) = \begin{cases} 1, p_j f_j(x) < p_j \theta_j \\ 0, other \end{cases} \qquad (3)$$

DNA cascade classifier is used to train molecular genetics features. Overall system for object recognition is set $F_{target}$. Different classifier maximum rate of misrecognition is $f_{\max}$. $\log_{f\max} F_{\max}$ strong splitters is analyzed. By this means, positive and negative samples set of training needs is got.

$for \quad t = 1,...,T$, i strong classifiers verify the negative samples, and filter the misstated selections which are incorporated into new negative samples.

**Identification of DNA lesions texture**

DNA image texture vector can distinguish different types of lesions. Learning set and fuzzy means clustering method was used. Squared error functions is a function of the cluster specification. $X = \{x_1, x_2, ..., x_n\}$ describes the gathering sets of pace lesions texture feature. Fuzzy mean is got by (4):

$$Z_m = \sum_{i=1}^{c} \sum_{j=1}^{N} (S + I)^m \| x_j - H_i \|^2 \qquad (4)$$

$\| \bullet \|$ stands for euclidean distance, and m represent fuzzy index ( $1 \le m < \infty$ ). By searching all texture feature vectors, a connected graph will be obtained $U = \{X_{ij}\}$. The requirements are as follows:

1) $X_{ij} \in [0,1]$, $i = 1,2,...,C$; $j = 1,2,...,N$;

2) $\sum_{i=1}^{C} X_{ij} = 1$, $j = 1,2,...,N$;

3) $\sum_{j=1}^{N} X_{ij} > 0$, $i = 1,2,...,C$。

By improving the iterative of connected graph in objective function, recognition function for lesion image can be obtained as shown by (5):

$$Xij = \frac{\left[\frac{1}{\| x_j - H_i \|^2}\right]^{\frac{1}{m-1}}}{\sum_{i=1}^{C}\left[\frac{1}{\| x_j - H_i \|^2}\right]^{\frac{1}{m-1}}} \quad (5)$$

$i = 1,2,...,C$ , $j = 1,2,...,N$ , The results are:

$$w_i = \frac{\sum_{i=1}^{N}(X_{ij})^m x_j}{\sum_{j=1}^{N}(X_{ij})^m} \quad (6)$$

$w_i$ represents the $i_{th}$ characteristic function of DNA lesions. Identification requirements are: Connected graph is got by searching the feature vector of DNA texture. Connected graph norm function is calculated to analyze whether there are same characteristic values of the same type of lesions. If it is consistent, the $i_{th}$ characteristic of lesions can be determined. The traditional Gaussian resolution method can only have examination by the color feature of image. Mix lesion characteristics cannot be analyzed. This study uses fuzzy recognition method to identify DNA imageand repeats iterative calculation several times. According to the texture feature vectors, different types of lesion characteristics are distinguished. Ultimately, effective prevention and treatment of infection will be realized.

In order to enhance the accuracy of clarification, spatial gradient information of molecular genetics features is used. The gradient of image f(x,y) is calculated. The corresponding horizontal and the vertical gradient values are:

$$G_x(x,y) = f(x,y+1) - f(x,y)$$
$$G_y(x,y) = f(x+1,y) - f(x,y) \quad (7)$$

Because the difference of DNA image among infected trees is large, the molecular genetics features are uniform for benign images with low gradient value. However, malignant images usually have messy texture and high gradient value. A thresholds should be set to summarize the number of pixels that have higher gradient value than threshold. It is defined as:

$$MUM1 = (abs(G_x(x,y)) > T)$$
$$NUM2 = (abs(G_y(x,y)) > T)$$
$$MUM = NUM1 + NUM2 \quad (8)$$

In this study, the threshold is set T=20. According to the principle of spatial filtering, the boundary filter can enhance the image texture to a great degree. In order to enhance the effects of texture division and disease recognition performance, the boundary function characteristics are used. Boundary function is defined as sum of absolute value of the difference among the pixel values with d distance. As (9) shows:

$$g_d(i,j) = \left|I(i,j) - I(i+d,j)\right| + \left|I(i,j) - I(-d,j)\right|$$
$$+ \left|I(i,j) - I(i,j+d)\right| + \left|I(i,j) - I(i,j-d)\right| \quad (9)$$

The average gradient value with certain distance d is set as molecular genetics features:

$$g(d) = \frac{\sum_{i=0}^{M}\sum_{j=0}^{N} g_d(i,j)}{M * N} \quad (10)$$

Thus, the density and contrast of boundary separately are:

$$EdgeDensity = \left|\left\{(x,y)\middle|g(x,y) \geq Th\right\}\right|$$
$$EdgeContrast = average_{x,y}(g(x,y)) \quad (11)$$

**Results and analysis**

In order to verify the effectiveness of the proposed method, the corresponding experiments are carried out. 280 images of infected DNA image are collected (150 is the routine pathological images, and 130 is unconventional.). The 280 samples are randomly divided into 5 group. Every group contains 30 malignant images and 26 benign images. 4 groups are used as training set, and the rest 1 group is as test set. The examination is executed for 5 times until all the groups have completed test. Final classification accuracy of the overall data is examined.

In order to analyze the performance of this algorithm, sampling classification accuracy, sensitivity and specificity are to analyze the performance of different algorithms. The corresponding definition is as follows:

$$accuracy = \frac{tp + tn}{tp + fn + fp + tn} \quad (12)$$

$$sensitivity = \frac{tp}{tp + fn} \quad (13)$$

$$specificity = \frac{tn}{fp + tn} \quad (14)$$

tp is used to describe the accurately detected number of infected DNA texture. fn represents the false detection. fp represents the false number of benign image. tn represents benign image is accurately detected.

According to the above definition, the results of Table1 and 2 can be got. We can see that the correlations among the collected DNA image will lead to negative effects on classification results. In this study, after the DNA molecular genetics features are chosen, an optimal subset is classified. The results show that the accuracy of 3 kinds of classifiers is increased. This suggests that collected DNA subset has an enhanced ability to distinguish malignant and benign images. Besides, our algorithm has higher classification accuracy than others.

Table 1 Feature classification by other algorithm

| algorithm | accuracy% | sensitivity% | Specificity% |
|---|---|---|---|
| K-average | 83.46 | 74.52 | 92.42 |
| SVM | 85.62 | 79.18 | 93.26 |
| ours | 94.21 | 94.56 | 93.86 |

Table 2 Feature classification by our algorithm

| algorithm | accuracy% | sensitivity% | Specificity% |
|---|---|---|---|
| K-average | 92.58 | 90.07 | 94.06 |
| SVM | 94.57 | 94.6 | 97.01 |
| ours | 95.28 | 97.16 | 93.45 |

Receiver operating characteristic curve (ROC) can determine the performance of DNA image classification by different methods. ROC analysis can generate specific curve for classifier sensitivity. The curve can measure the overall classification performance. ROC curves for 3 types of classifiers optimized by our algorithm or others are demonstrated by figure 2 and 3. The area formed by curve and horizontal axis is used to describe the average performance among entire cost of classification system.

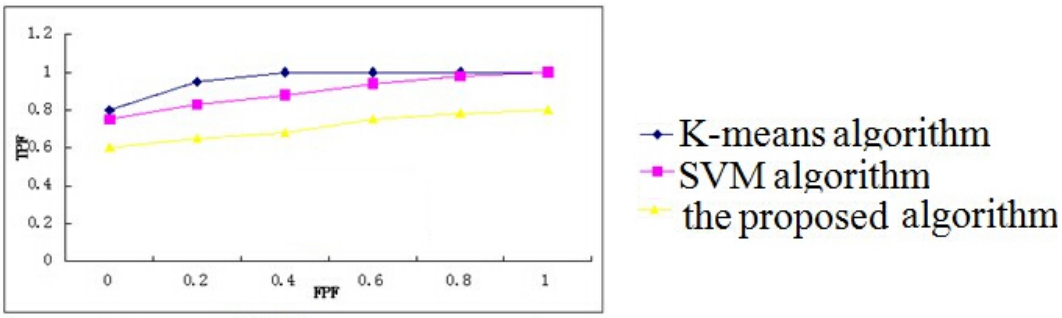


Figure 2 The ROC curve of classifiers not optimized by our algorithm

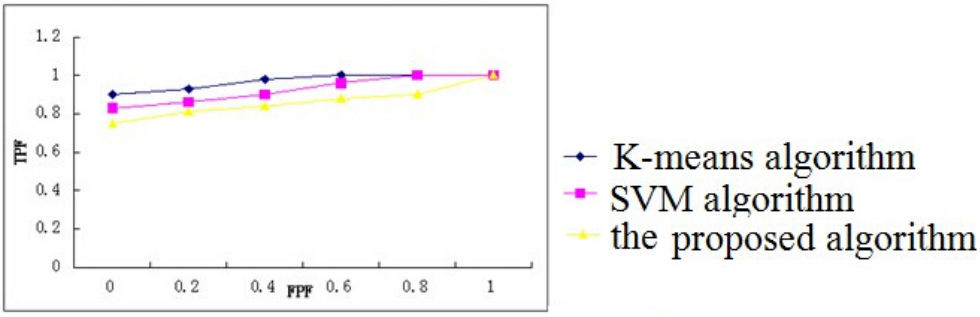


Figure 3 The ROC curve of classifiers optimized by our algorithm

In figure 2, the curve formed areas calculated by algorithms including K-average, SVM and our method are 0. 8211, 0. 8472, 0. 9282 respectively. In figure 3, the curve areas by K-average, SVM and our method are 0. 917, 0. 9475, 0. 9416 respectively. This indicates that the features optimized by our methods possess better ability to distinguish between benign and malignant properties.

The algorithm provided in this article optimizes 4 types of DNA profile vectors, which is used to classify the target images. In order to verity the effectiveness of the proposed method, vector machines are used to have a classification. Time consumption of the 3 types of methods analyzing texture image is showed in Table 3. Analysis shows that, for the same classifiers of test set and the training set, features analyzed by our algorithm can better describe the difference between benign and malignant lesions. Especially when the four features are classified together, the classification accuracy and sensitivity are enhanced significantly. Also, the algorithm has the characteristics of high acquisition speed.

Table 3 The classification for lesion characteristics（%）

| types | Classification accuracy | sensitivity | specificity |
|---|---|---|---|
| cracking | 83.45 | 67 | 100 |
| edema | 88.37 | 73 | 100 |
| fester | 89.75 | 88.09 | 92.45 |
| necrosis | 87.27 | 88.44 | 91.25 |

## Conclusions

This article presents a new prediction method based on DNA molecular genetics features evolution for infected styphnolobium japonicum. GLCM collects the features of DNA image of diseased trees. Boundary function feature enhances the texture division and recognition of infection. Learning set and fuzzy means distinguish different types of lesion characteristics. Cascade classifier is used to train DNA molecular genetics features. By this means, the effective prevention and treatment of infected styphnolobium japonicum can be realized. The results indicate that the optimized features by the method have the preferred characteristics of evaluating benign and malignant images. Moreover, this method has high speed of texture acquisition, and high accuracy of identification of infected styphnolobium japonicum.

## References

[1] A Clausid, Zhao Yong-ping. Rapid co-occurrence texture feature extraction using a hybrid data structure [J]. Computers& Geosci-ences, 2002, 28(6): 763-774.
[2] Dong Jian-wei, Huang Rong-bo, Ma Jian-hua. Application of multi-band wavelwts to medical image retrieval [J]. Chinese medical equipment, 2007, 28(1):5-10.
[3] Wang Hui-ming, Shi Ping. Methods to extract images molecular genetics features [J]. Journal of communication university of China science and technology, 2006, (13): 49-52.
[4] Ye Qing, Huang Yan-lei. Non-uniform distribution intrusion detection research and simulation of the model [J]. Bulletin of science and technology, 2013, 29(8): 169-171.
[5] Zhang Xue-gong. Introduction to statistical learning theory and support vector machines [J]. Acta automatica sinica, 2000, 26(1):32-42.