Classification of hyperspectral remote sensing data based on DNA computing

JIAO Hongzan, ZHONG Yanfei, ZHANG Liangpei, LI Pingxiang

National Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Hubei Wuhan 430079, *China*

Abstract: Some initial investigations are conducted to employ DNA computing for hyperspectral remote sensing data classification. As a novel branch of computational intelligence, DNA computing expresses rich information of spectral features with DNA encoding, and acquires the most typical DNA encoding of each class by DNA modulating and controlling mechanism. For each pixel of the hyperspectral image, computing the distance between the pixel and the typical DNA sequence, finding the class property of the minimum distance, set the class property of each pixel as the minimum distance class. An experiment was performed to evaluate the performance of the proposed algorithm in comparison with other traditional image matching classification algorithms: binary cording, spectral angles and spectral derivative feature coding (SDFC). It is demonstrated that the proposed algorithm is superior to the three traditional hyperspectral data classification algorithms based on the experiment results.

Key words: DNA computation, hyperspectral, spectral matching, classification

CLC number: TP751.1 **Document code:** A

Citation format: Jiao H Z, Zhong Y F, Zhang L P and Li P X. 2010. Classification of hyperspectral remote sensing data based on DNA computing. *Journal of Remote Sensing.* **14**(5): 865—878

1 INTRODUCTION

Hyperspectral remote sensing is characterized by the integration of image and spectrum. Not only does it provide the spatial distribution of different classes, but also it reflects the significant spectral information. A hyperspectral signature provides significant spectral information for discrimination and classification due to hundreds of contiguous spectral bands. Based on the characteristic discrimination and classification method, spectral matching techniques for the hyperspectral remote sensing data, have been formed. Considerable attention has been given to developing spectral matching techniques, which allow remote sensing-derived spectrum to be compared with spectrums that were previously collected in the field or in the laboratory (Tong *et al*., 2006).

Encoding matching technique is a significant part of spectral matching techniques. A simple technique is binary spectral encoding matching (Mazer *et al*., 1988). Binary coding can be used to transform a hyperspectral reflectance spectrum into simple binary information. However, the simple binary coding causes the loss of the spectral curve signatures, so that we can't carry out discrimination and classification with high accuracy.

To address this problem, some general encoding approaches have been investigated for hyperspectral signature characterization. The encoding-based approach which encodes spectral signatures as codewords and spectral analysis is then conducted by using the Hamming distance as a spectral similarity measure. Three such methods are notable. One is called spectral analysis manager (SPAM) developed by Jia & Richards (1993), which encodes an L-dimensional signature as a (2L-2)-dimensional binary code word composed of the first L binary values used to encode the sign of the difference between a signature and its signature mean, and additional L-2 binary values used to encode the sign of the difference in spectral values between a band and its adjacent band. The SPAM binary coding was further extended to the so-called spectral feature-based binary coding (SFBC) by Qian *et al*. (1996), who introduced additional L-2 binary values to encode a signature as a (3L-4)-dimensional binary code word. The new added L-2 binary values are used to dictate whether the deviation of a spectral variation from the signature mean is greater than a prescribed threshold. Recently, a new signature coding method,

Received: 2009-07-15; **Accepted**: 2009-11-11

Foundation: Major State Basic Research Development Program (973 Program) of China (No. 2009CB723905), 863 High Technology Program of the People's Republic of China (No. 2009AA12Z114), and National Natural Science Foundation of China (No. 40901213, No. 40930532 and No. 40771139), Foundation for the Author of National Excellent Doctoral Dissertation of PR China (FANEDD), Research Fund for the Doctoral Program of Higher Education of China (No. 200804861058), Program for New Century Excellent Talents in University (No.NECT-10-0624), The Hubei Province Science Foundation under Grant (No. 2009CDB173), Foundation of National Laboratory of Pattern Recognition and Key Laboratory of Geo-informatics of State Bureau of Surveying and Mapping.

First author biography: JIAO Hongzan (1985—), PhD candidate. He received the B.S. degree in land resource management from Wuhan University in 2008. He is currently pursuing the Ph.D. degree in the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing at Wuhan University. His current research interests focus on multi- and hyper- spectral image processing, DNA computing, and pattern recognition.

Corresponding author: ZHONG Yanfei, E-mail: zhongyanfei@lmars.whu.edu.cn

referred to as spectral derivative feature coding (SDFC), has been presented by Chang *et al*. (2009), which improves both SPAM and SFBC in the sense of signature characterization. These traditional encoding methods have demonstrated some success in getting spectral signatures. However, when traditional encoding methods are used to discriminate and classify the hyperspectral data with instability of the spectrum, the results is not satisfactory.

DNA computing is a novel intelligent method, and the initial investigations and ideas about DNA computing are conducted by Professor Adleman, University of Southern California, USA (1994). DNA computing has been exploited successfully in pattern recognition applications, fuzzy control, and decision-making problems using DNA encoding and DNA population control mechanism (Lipton, 1995; Faulhammer *et al.*, 2000; Ren *et al.*, 2001; Chen *et al.*, 2003; Benenson *et al*., 2004). Based on the advantages of DNA computing model, a DNA computing pattern recognition system for hyperspectral data encoding matching classification was developed. The result of hyperspectral classification is satisfactory.

2 BASIC CONCEPT OF DNA COMPUTING

DNA computing, inspired by natural evolution, presents a transferring mechanism of organisms' genetic information (Adleman, 1994).

2.1 Transferring law of genetic information

In nature, the different species perform biodiversity, while the same species present biological similarity. The phenomenon is decided by the genetic material, Deoxyribonucleic acid, DNA. DNA is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms and some viruses. DNA is composed by 4 kinds of Nucleobases (adenine, A; guanine, G; cytosine, C; thymine, T). The permutations of the Nucleobases perform extremely abundant information. Though the biochemical reactions, the organisms transfer the genetic information, which is the basis phenomenon of organism.

It is the sequence of these four bases along the backbone that encodes information (Fig.1). This information is read using the genetic code, which specifies the sequence of the amino acids within proteins. The code is read by copying stretches of DNA into the related nucleic acid RNA, in a process called transcription.

In the mathematics learning field, the value of a complex computable function can be compound with a series of simple functions. From the viewpoint of DNA computing, the solutions of complex problems can be solved by the permutations of the Nucleobases. This is a similarity between the mathematics and biological intelligence (Ding *et al*., 2002).

2.2 Model, theory, method of DNA computing

Based on the background of biology, the initial idea of DNA computing and simulating the genetic mechanism of organism,

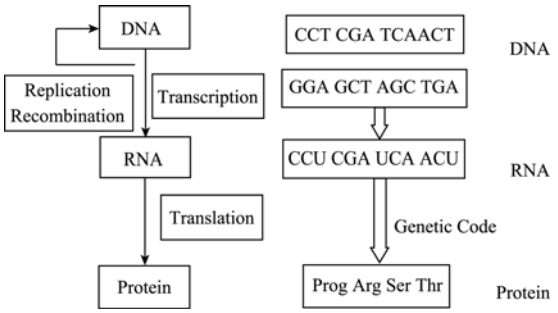

Fig. 1 Transformation process of biological genetic information

the information model of DNA computing is proposed. On the traditional computer, it combines DNA computing with artificial intelligence to use the DNA coding and the controlling mechanism of genetic information. The information model based on the DNA coding, deepens the research on DNA computing theory, and enlarges its applications. DNA computing model has been used in the artificial immune system (Ren & Ding, 2001), genetic algorithm (Shin *et al*., 2005)

2.2.1 DNA encoding and group initialization

DNA is composed by 4 kinds of Nucleobases, adenine (A), cytosine (C), guanine (G) and thymine (T). The initialization of DNA encode is the transformation from band values to DNA encode, {T, C, A, G}.

The fitness of each DNA individual measures by the diversity of amino acid sequences between training DNA sequence and typical DNA individual.

2.2.2 Gene operation of DNA computing

Recombination is a significant part of gene operation. Recombination allows chromosomes to exchange genetic information and produces new combinations of genes, which increases the efficiency of natural selection. The mechanism of gene operation can be considered as a process of optimization.

3 DNA COMPUTING CLASSIFICATIONS OF HYPERSPECTRAL DATA

Combined with the spectrum matching and DNA computing, the genetic information model is built up based on the DNA encoding and genetic information controlling mechanism. Therefore, a novel classification method, DNA computing classification for hyperspectral data, is proposed. The model transforms the abundant information of spectrum curve by DNA encoding, and controls and adjusts DNA encodes by DNA genetic operations, finaly get the most typical DNA code. It is an optimization process of the spectral matching classification.

3.1 DNA computing model initialization

DNA population parameter setting: Before performing the DNA models for the classification, some parameter of the DNA population must be stetted, including the quantity of individual in population, Crossover probability, Mutation probability and Loop termination condition.

Sample Selection: the samples of DNA computing model divided into 2 parts, the training samples and the validation sample. Referring to the corresponding high resolution image and ground-truth data, the samples can be acquired. The training samples and the validation samples, with fixed number, will be selected randomly from the ground-truth imagery. And then, the DNA initial individuals with fixed population can be selected randomly from the training samples.

3.2 DNA encoding

The Primary problem of DNA computing for hyperspectral data classification is the method of DNA coding. On the basis of the characteristic of hyperspectral data, with multi-dimensions, DNA code need to satisfy 3 requirements: (1) DNA code adopt the form of 4 value code, and the code words represent by {T, C, A , G}. The way of coding can retain much more details of spectral curves. (2) DNA code can get physical reflection and absorption features of spectrum. (3) DNA code performs the capability of noise tolerance, can weaken the interference of noise, which influence the accuracy of spectrum classification.

Combined with the spectrum coding method and the demand of DNA computing model, a new coding method, DNA coding for hyperspectral data, is proposed. DNA coding is composed by two parts, named DNA_A coding and DNA_B coding. DNA_A coding part measures overall trend of spectrum, by using the mean of spectrum values; and DNA_B coding part describes the diversification's details of spectrum by using the change of adjacent band values.

Details are as follows,

DNA_A part Firstly let T_0 be the spectral mean of a hyperspectral signature, and divide the spectral value into 2 intervals, $[x_{\min},T_0)$ and $[x_0,T_{\max})$. Secondly, try to compute the spectral means of 2 intervals respectively, T_l and T_r, therefore the spectral signature is divided into 4 intervals, $[x_{\min},T_l)$, $[T_l,T_0)$, $[T_0,T_r)$ and $[T_r,T_{\max})$, and then, coding the 4 intervals with 4 code words, $\{T,C,A,G\}$, described by Eq. (1). Each band value of hyperspectral signature of each pixel gets the code word, according to the intervals that the band value is in. When the number of band is L, the length of code words is also L.

$$S_i^{\mathrm{DNA\_A}}=\begin{cases}T & S_i\in[x_{\min},T_l)\\ C & S_i\in[T_l,T_0)\\ A & S_i\in[T_0,T_r)\\ G & S_i\in[T_r,x_{\max}]\end{cases} \tag{1}$$

DNA_B part It is developed to describe texture features based on gradient changes in gray levels of three successive adjacent pixels among these bands. More specifically, assume that $S=(s_1,s_2,...s_L)^T$ is a hyperspectral signature where L is the total number of spectral bands and the S_l is the l-th spectral band. Also, let Δ be a desired spectral value tolerance. There are four types of successive gradient changes in spectral values that can be described as follows:

$$\begin{aligned}&\text{Type 1 if } |s_i-s_{i-1}|\leqslant\Delta \text{ and } |s_{i+1}-s_i|\leqslant\Delta\\ &\text{Type 2 if } (|s_i-s_{i-1}|\leqslant\Delta \text{ and } |s_{i+1}-s_i|>\Delta)\\ &\qquad\text{or } (|s_i-s_{i-1}|>\Delta \text{ and } |s_{i+1}-s_i|\leqslant\Delta)\\ &\text{Type 3 if } (s_i-s_{i-1}<-\Delta \text{ and } s_{i+1}-s_i<-\Delta)\\ &\qquad\text{or } (s_i-s_{i-1}>\Delta \text{ and } s_{i+1}-s_i>\Delta)\\ &\text{Type 4 if } (s_i-s_{i-1}<-\Delta \text{ and } s_{i+1}-s_i>\Delta)\\ &\qquad\text{or } (s_i-s_{i-1}>\Delta \text{ and } s_{i+1}-s_i<-\Delta)\end{aligned} \tag{2}$$

Additionally, the Δ used in Eq. (2) can be set to

$$\Delta=(1/(L-1))\sum_{l=2}^{L}|r_i-r_{i-1}| \tag{3}$$

According to degrees of successive gradient changes in spectral values among three consecutive adjacent spectral bands, the spectral encoding will fall into four types. The graphic representations of these four types of gradient changes in spectral values are illustrated in Fig.2 for visualization to better understand why these four types of gradient changes can be more effective in characterizing spectral variability among three consecutive adjacent bands (Chang *et al*., 2009).

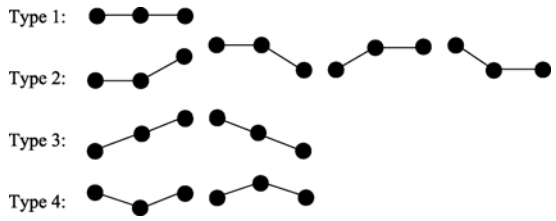

Fig. 2 Graphical representation of spectrum curve characteristics

Now we are ready to develop a coding technique to capture spectral texture feature change in three consecutive adjacent bands according to the four types of gradient changes in spectral variation described by Eq.(3). For $1<i<L$ we define DNA_B as

$$S_i^{\mathrm{DNA\_B}}=\begin{cases}T & \text{if } S_i \text{ is type 1}\\ C & \text{if } S_i \text{ is type 2}\\ A & \text{if } S_i \text{ is type 3}\\ G & \text{if } S_i \text{ is type 4}\end{cases} \tag{4}$$

By virtue of Eq. (1) and Eq. (4), we can encode the spectral value, S_l of the l-th spectral band as

$$\begin{aligned}S^{\mathrm{DNA}}=\{&S^{\mathrm{DNA\_A}},S^{\mathrm{DNA\_B}}\}=\{S_1^{\mathrm{DNA\_A}},S_2^{\mathrm{DNA\_A}},\\ &...,S_L^{\mathrm{DNA\_A}},S_2^{\mathrm{DNA\_B}},S_3^{\mathrm{DNA\_B}},...,S_{L-1}^{\mathrm{DNA\_B}}\}\end{aligned} \tag{5}$$

We can interpret Eq.(5) as a set of $2L$-2 quaternary code words.

DNA coding conducts a transformation from the spectral space to the DNA code space. The graphic representations of this transformation are illustrated in Fig.3 for visualization to better understand the process. Fig.3(a) represents original spectral data, with the L bands (L=80), and the spectral range is from 0.417 to 0.854μm; Fig.3(b) represents code words after DNA coding, with the length of $2L$-2.

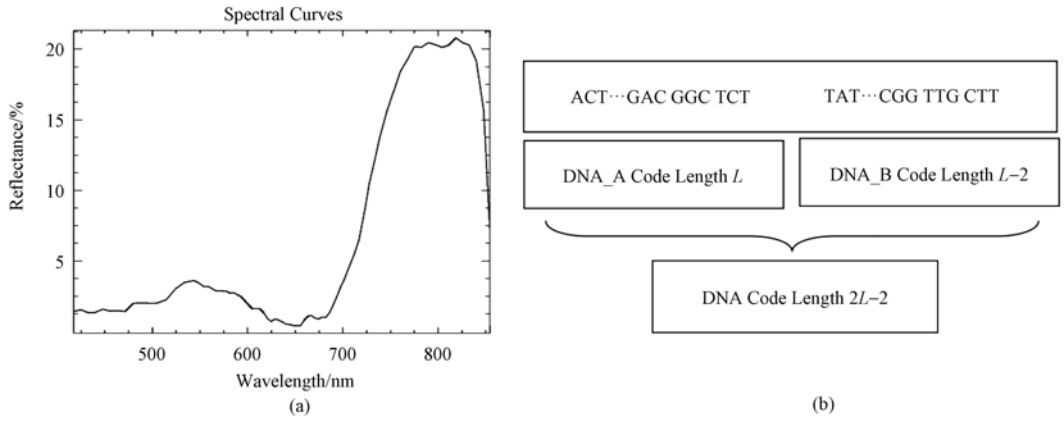

Fig. 3 DNA coding for spectral data
(a) Original spectral data; (b) DNA coding words

DNA encoding is conduced by quaternary coding, describes the absorption and reflection features and the overall tendency of the spectrum, and extracts the physical feature from the hyperspectral curves; the control and adjustment of spectral value tolerance Δ, reduce the influence of spectral noise, and make DNA coding more stable.

3.3 DNA controlling mechanism of spectrum matching for hyperspectral data

3.3.1 Transcription and translation

The genetic code is the set of rules by which information encoded in genetic material (DNA or RNA sequences) is translated into proteins (amino acid sequences) by living cells (Ren & Ding, 2001).

The operation of Transcription and Translation imitate the process of Protein formation. Using the DNA encoding method, we can translate the codons into the amino acids based on the corresponding relationship between the codons and the amino acids shown in Table 1. Then, the amino acids are translated into the design parameters of spectral signature. The translation process in Table 1 imitates the translation process from DNA to protein. Also, it is the basic framework for translating the codons into the amino acids.

Table 1 Translation from the DNA codons into the amino acids

1st base	2nd base				3rd base
	T	C	A	G	
T	Phe(0)	Ser(2)	Tyr(3)	Cys(4)	T
	Phe(0)	Ser(2)	Tyr(3)	Cys(4)	C
	Leu(1)	Ser(2)	Stop(9)	Stop(9)	A
	Leu(1)	Ser(2)	Stop(9)	Try(9)	G
C	Leu(1)	Pro(5)	His(6)	Arg(8)	T
	Leu(1)	Pro(5)	His(6)	Arg(8)	C
	Leu(1)	Pro(5)	Gln(7)	Arg(8)	A
	Leu(1)	Pro(5)	Gln(7)	Arg(8)	G
A	Ile(11)	Thr(12)	Asn(13)	Ser(2)	T
	Ile(11)	Thr(12)	Asn(13)	Ser(2)	C
	Met(10)	Thr(12)	Lys(14)	Arg(8)	A
	Met(10)	Thr(12)	Lys(14)	Arg(8)	G
G	Val(15)	Ala(16)	Asp(17)	Gly(19)	T
	Val(15)	Ala(16)	Asp(17)	Gly(19)	C
	Val(15)	Ala(16)	Glu(18)	Gly(19)	A
	Val(15)	Ala(16)	Glu(18)	Gly(19)	G

The major object of DNA translation is building the framework of DNA operation. By employing the fuzzy rules, the DNA encoding is translated into amino acids, and then translated into the design parameters; based on this translation, we can conduct the classification operation on the training samples and the image. Furthermore, the framework of DNA translation enhances the stability and tolerance ability of DNA computing model.

3.3.2 Fitness calculation

The fitness of each DNA individual measures by the diversity of amino acid sequences between training DNA sequence and typical DNA individual.

Assume that S^{DNA} and S^{Amino} as DNA codewords and amino acid sequence respectively:

$$S^{\mathrm{DNA}} \rightarrow S^{\mathrm{Amino}} = \{a_1, a_2, a_3, \ldots, a_T\} \tag{6}$$

Where $T = \lfloor (2 \times L - 2)/3 \rfloor$ and $a_i = 0,1,2,\cdots,19$.

Calculate the fitness of training DNA individuals in Eq. (7).

$$\mathrm{Fitness} = \sum_{i=0}^{N} | S_{\mathrm{Train\_}i}^{\mathrm{amino}} - S_{\mathrm{individual}}^{\mathrm{amino}} | = \sum_{i=0}^{N} \sum_{j=0}^{T} | a_{ij} - a_j^{'} | \tag{7}$$

Where $S_{\mathrm{Train\_i}}^{\mathrm{amino}} = \{a_{i1}, a_{i2}, \cdots, a_{iT}\}$, $S_{\mathrm{individual}}^{\mathrm{amino}} = \{a_1', a_2', \cdots, a_T'\}$, and $T = \lfloor (2 \times L - 2)/3 \rfloor$ And N is the number of training samples.

3.3.3 Genetic operators

Based on the new DNA encoding method, we develop the genetic operators in the DNA computing model. They are crossover, mutation and updating of DNA individuals.

Crossover operation: Crossover is a process of exchanging genetic information, which is important for the entire search process. Studies have shown that signature codes crossover seems to be a better method for crossover, and is often adopted in the traditional Gas. For the reason, we adopt signature codes

crossover operation in this model. An example is shown in Fig.4.

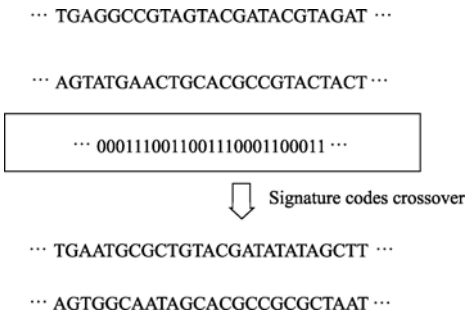

Fig. 4 Standard crossover diagram

Mutation operation: Any change in a DNA sequence is called a mutation. They are point mutations and frameshift mutation. Two kinds of point mutation exist in a DNA sequence, such as transition mutation and transversion mutation. In order to imitate the mutations of the biological DNA, we employ the same point mutation manners in the DNA computing model. During the process of practical implementation, we do the mutation operation as follows:

Generate a random number, *s*, in [0, 1] at random, the base at a selected place in a DNA individual will be mutated to:

1) T, if $s\in[0,0.25)$

2) C, if $s\in[0.25,0.5)$

3) A, if $s\in[0.5,0.75)$

4) G, if $s\in[0.75,1]$

Update of DNA generation: The fitness value of the DNA individual is calculated, and the individuals that have low fitness values are excluded from the generation sets. At the same time, the same amount of new individuals are generated and added into the new generation sets. And set the individual of the highest fitness value as the best individual.

3.3.4 Stopping condition

When the best individual's fitness value satisfies the threshold, which is set before this process, or the number of evolution generation reach the maximum number, the genetic operation will be stopped. Otherwise the operation will be operating, until the requirement is satisfied.

After conducting the gene operation, the typical DNA spectral encoding can be optimized.

3.4 Classification

For each pixel of hyperspectral image, computing the amino acids parameters distance between pixel and the typical DNA sequence, finding the class property of the minimum distance, set the class property of each pixel as the minimum distance class.

Fig. 5 gives the process of DNA computing classification.

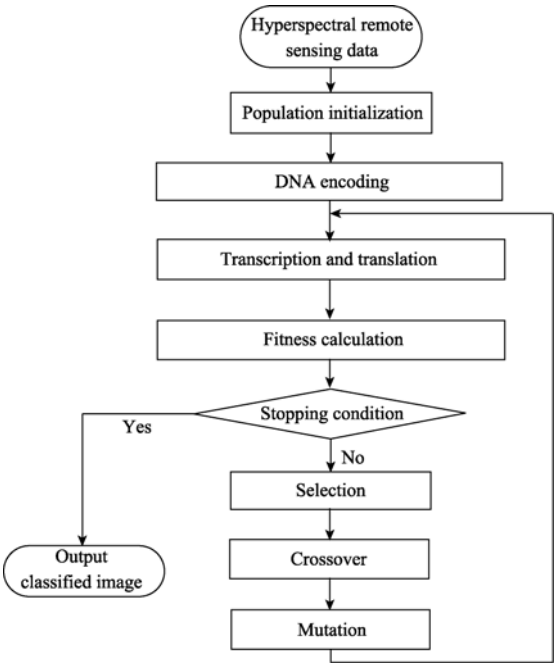

Fig. 5 Flowchart of DNA computing classification algorithm for the hyperspectral data

4 EXPERIMENTS AND ANALYSIS

4.1 Experiment data

In this experiment, the data is airborne imaging spectrometer (PHI) data, 80 bands taken from Xiaqiao test site which is a mixed agricultural area in China. Eighty bands of PHI image (340×390 pixels) were used in this experiment, and their spectral ranges were from 0.417—0.854 m. Fig. 6 shows the experimental PHI image. The observed image was expected to fall into six classes: road, corn, rubble grounds, vegetable, grassland and water. The reference of Spectral Curves for six classes is given in Fig. 7.

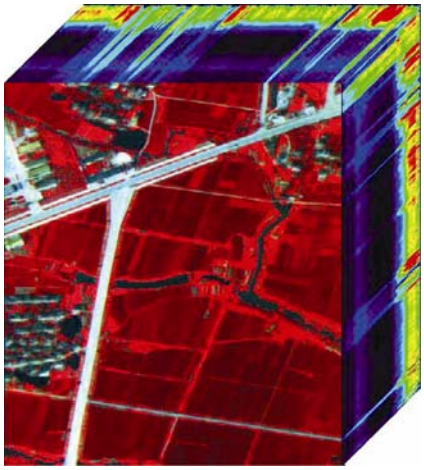

Fig. 6 Xiaqiao PHI image RGB(70,40,10)
Bands 80 (0.41—0.85μm)

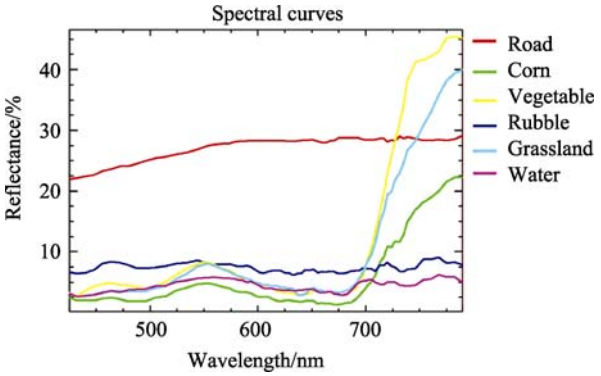

Fig. 7 Reference spectral curves for six classes

Before conducting the experience, we need not to perform pretreatment, such as radiometric correction, band selection and filtering.

4.2 Input data

In the experience, we select randomly 180 samples for the six classes, and the number of training samples is 80, while the number of validation is 100.

The values of parameters set in Experiment. The details are as follows: individual number in one generation as 20, cross-over probability as 0.9, mutation probability as 0.02, maximum generation as 200 and the threshold as 0.98.

4.3 Experiment result

Fig.8(d) illustrates the classification result using DNA computing. Fig.8(a)–(c) illustrates the classification results using binary coding, spectral angle matching and SDFC spectrum matching method.

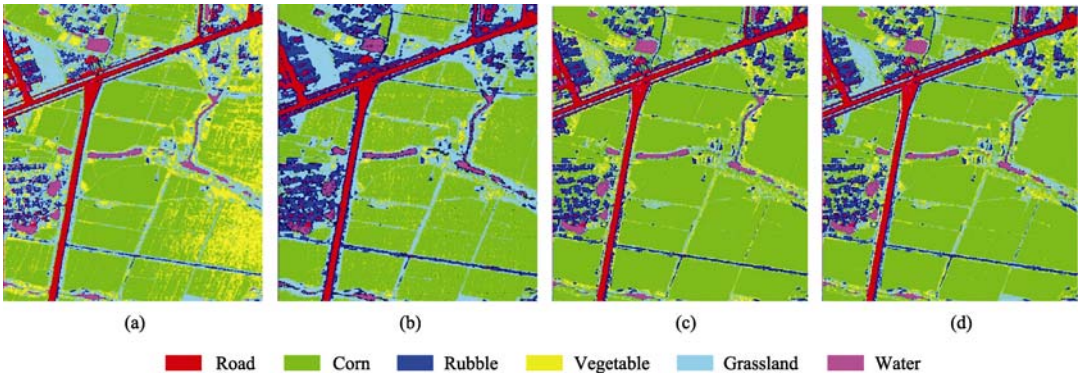

Fig. 8 Spectral matching classification images for Xiaqiao PHI image
(a) Binary encoding; (b) Spectral angel; (c) SDFC match; (d) DNA computing match

4.4 Accuracy comparison

The classification accuracy for the several classifiers is given in Table 2 and Table 3 including producer's accuracy, user's accuracy, overall accuracy and Kappa coefficient of agreement based on the confusion matrixes (Foody, 2002).

Table 2 Comparison of four classification methods using producer's and user's accuracy

/%

	Class	Road	Corn	Rubble	Vegetable	Grassland	Water
Binary encoding matching	Producer's accuracy	97	51	65	86	75	100
	User's accuracy	89.81	92.73	97.01	54.44	77.32	91.74
Spectral angel mapper	Producer's accuracy	100	76	91	66	93	73
	User's accuracy	91.74	95	86.67	68.75	75.61	92.41
SDFC matching	Producer's accuracy	95	99	87	68	45	93
	User's accuracy	100	72.79	87.88	61.26	76.27	93
DNA computing matching	Producer's accuracy	95	100	87	83	82	94
	User's accuracy	100	81.30	90.63	87.37	91.11	93.07

Table 3 Comparison of four classification methods using over all accuracy and Kappa coefficient

Method / Accuracy	Binary encoding matching	Spectral angel mapper	SDFC matching	DNA computing matching
Overall accuracy/%	79.00	83.17	81.17	90.17
Kappa coefficient	0.7480	0.7985	0.7740	0.8820

As shown in Table 2 and Table 3, the DNA computing classifier produces better classification results than traditional classifiers. The details are as follows: DNA computing improved overall classification accuracy from 79.00% to 90.17%, an improvement by 21.17% and Kappa coefficient from 0.748 to 0.882, improving 0.134. Based on the above, we can make a conclusion that DNA computing classifier is the good classifier applied with hyperspectral remote sensing image spectral matching classification. The reason of DNA computing classifier's advantage is that it is not only obtain the more details about the spectral curves, but also the DNA computing model can avoid the interference from spectral noise and make the matching process optimal.

5 CONCLUSION

In this paper, some initial investigations are conducted to employ DNA computing for hyperspectral remote sensing data classification. As a novel branch of computational intelligence, DNA computing express rich information of spectral features with DNA encoding, and acquire the most typical DNA encoding of each class by DNA modulating and controlling mechanism. For each pixel of hyperspectral image, computing the distance between pixel and the typical DNA sequence, finding the class property of the minimum distance, set the class property of each pixel as the minimum distance class. The main idea of DNA computing is the transformation from the spectral signature space to DNA codes space, just like Wavelet and Fourier transformation; and the optimization process by DNA gene operation.

The object of DNA computing matching classification is hyperspectral data with reflectance curves. DNA computing classifier is not only obtain the more details about the spectral curves, but also the DNA computing model can avoid the interference from spectral noise and make the matching process optimal. Because of these, we needn't to conduct pretreatment, such as radiometric correction, band selection and filtering.

The Xiaqiao PHI hyperspectral data experiment was performed to evaluate the performance of the proposed algorithm. It is demonstrated that the proposed algorithm is superior to the three traditional hyperspectral data classification algorithms based on the experiment results, and its overall accuracy and Kappa coefficient reach 90.17% and 0.8820, respectively.

The genetic operations, used in the DNA computing model, are some traditional biological intelligence operation, such as selection, crossover and mutation. Because the aim of this paper is examine the adaptability of DNA computing for hyperspectral data. We will research new gene operation like inversion and separation in the future work.

REFERENCES

Adleman L. 1994. Molecular computation of solutions to combinatorial problems. *Science*, **266**(5187): 1021—1024

Benenson Y, Gil B and Ben-Dor U. 2004. An autonomous molecular computer for logical control of gene expression. *Nature*, **429**: 423—429

Chang C I, Chakravarty S Chen H and Ouyang Y. 2009.Spectral derivative feature coding for hyperspectral signature analysis. *Pattern Recognition*, **42**: 395—408

Chen J, Li H, Sun K and Kim B. 2003. How will bio-informatics impact signal processing. *IEEE Signal Processing Magazine*, **20**: 16—26

Ding Y S, Shao S H and Ren L H. 2002. DNA Computing and Soft Computing. Beijing: Science Press

Faulhammer D, Cukras A R, Lipton R J and Landweber L F.2000. Molecular computation: RNA solutions to chess problems. *ProcNatl Acad. Science*, **97**(4): 1385—1389

Foody G M. 2002. Status of land cover classification accuracy assessment. *Remote Sensing of Environment*. **80**: 185—201

Jia X and Richards J A. 1993. Binary coding of imaging spectrometer data for fast spectral matching and classification. *Remote Sensing of Environment*, **43**: 47—53

Lipton R J. 1995. DNA solution of hard computational problem. *Science*, **268**: 542—545

Mazer AS, Martin M, Lee M and Solomon J E. 1988. Image processing software for imaging spectrometry analysis. *Remote Sensing of Environment*. **24**(1): 201—210

Qian S, Hollinger A B, Williams D and Manak D. 1996. Fast three-dimensional data compression of hyperspectral imagery using vector quantization with spectral-feature-based binary coding. *Optical Engineering*. **35**(11): 3242—3249

Ren L H and Ding Y S. 2001. Design of fuzzy control system by a new DNA-based immune genetic algorithm. The 10th IEEE International Conference Fuzzy System. Melbourne, Australia, Dec. 2-5, 244—247

Shin S Y, Lee I H, Kim D and Zhang B T. 2005. Multi-objective evolutionary optimization of DNA sequences for reliable DNA computing. IEEE Transactions on Evolutionary Computation. **9**(21): 143—158

Tong Q X, Zhang B and Zhen L F. 2006.Hyperspectral Remote Sensing. Beijing: Science Press

高光谱遥感数据的 DNA 计算分类

焦洪赞, 钟燕飞, 张良培, 李平湘

武汉大学 测绘遥感信息工程国家重点实验室, 湖北 武汉 430079

摘 要: 提出了一种基于 DNA 计算的高光谱遥感数据光谱匹配分类新方法。该方法利用 DNA 编码提取各类地物光谱所携带的物理吸收与反射特征信息, 将地物光谱特征转换为 DNA 编码空间特征, 通过 DNA 计算基因操作寻找各类地物最典型的 DNA 信息链。在此基础上, 利用 DNA 计算原理建立一系列模糊规则, 对高光谱数据进行光谱匹配分类。通过与传统的光谱匹配算法(二值编码, 光谱角, 光谱差分特征编码)的分类结果进行比较, 证明该算法分类精度优于传统高光谱数据的光谱匹配分类方法, 具有实用价值。

关键词: DNA 计算, 高光谱, 光谱匹配, 分类

中图分类号: TP751.1 **文献标志码:** A

引用格式: 焦洪赞, 钟燕飞, 张良培, 李平湘. 2010. 高光谱遥感数据的 DNA 计算分类. 遥感学报, **14**(5): 865—878
Jiao H Z, Zhong Y F, Zhang L P and Li P X. 2010. Classification of hyperspectral remote sensing data based on DNA computing. *Journal of Remote Sensing*. **14**(5): 865—878

1 引 言

高光谱遥感具有图谱合一的特点, 在提供地物空间分布的同时获取反映地物物性的光谱信息。由于高光谱遥感数据具有高维的光谱特征, 使得利用光谱曲线进行地物识别成为可能, 也形成了高光谱遥感数据特有的识别方法——光谱匹配技术。在光谱匹配技术的基础上, 形成了面向高光谱数据特点的分类算法, 其中基于地物的光谱反射或发射曲线的分类识别方法最具特色(Tong 等, 2006)。编码匹配分类识别是光谱匹配分类的一种重要形式, 常用的编码匹配分类识别方法是光谱二值编码(Mazer, 1988)。

通过二值编码的方式提高了高光谱数据处理效率, 减少高光谱遥感数据的处理量, 解决海量数据冗余带来的处理问题。但是, 二值编码处理过程中会丢失细节光谱信息, 难以实现高精度匹配。针对该问题, 许多学者提出了光谱编码匹配的改进方法。Jia 和 Richards (1993)提出了一种基于光谱分析的二值光谱编码(SPAM)方法; Qian 等(1996)提出了一种基于光谱特征的编码(SFBC)方法; Chang 等(2009)提出了一种光谱差分特征的光谱编码(SDFC)方法。这些方法在一定程度上提高了光谱匹配的精度, 但由于光谱编码匹配, 无法利用统计学模型以及相邻像元间的相关信息, 如果同种地物光谱曲线的差异超过一定的范围, 传统的光谱编码匹配方法的精度将难以满足分类识别应用的需求。

DNA 计算是一种新型的智能化方法, 是 1994 年由美国南加州大学的 Adleman 教授在 Science 上提出的, 从而开创了 DNA 计算机的新纪元。DNA 计算利用 DNA 编码表示复杂知识或系统, 具有进一步地分析和模仿遗传信息调控系统的自生成、自组织功能, 在进化中能够获取和更新知识, 已经在模式识别、模糊控制、决策问题等工程领域得到广泛地应用(Lipton, 1995; Faulhammer 等, 2000; Ren 等, 2001; Chen 等, 2003; Benenson 等,2004)。基于 DNA

收稿日期: 2009-07-15; 修订日期: 2009-11-11

基金项目: 国家 973 计划资助项目(编号: 2009CB723905), 国家 863 计划资助项目(编号: 2009AA12Z114), 国家自然科学基金资助项目(编号: 40901213,40930532,40771139), 教育部博士点新教师基金(编号: 200804861058), 全国博士学位论文作者专项资金资助项目, 教育部新世纪优秀人才支持计划(编号: NECT-10-0624), 湖北省自然科学基金(编号: 2009CDB173), 模式识别国家重点实验室开放基金和地理空间信息工程国家测绘局重点实验室开放课题。

第一作者简介: 焦洪赞(1985—), 男, 2008 年毕业于武汉大学资源与环境学院, 现为武汉大学测绘遥感信息工程国家重点实验室硕士研究生, 主要研究方向为遥感图像处理, 模式识别, 人工智能和 DNA 计算。

通讯作者: 钟燕飞, E-mail: zhongyanfei@lmars.whu.edu.cn。

计算的优点，本文将DNA计算的思想引入光谱编码匹配算法中，按照优化原理将已有的地物光谱数据转化为相应的DNA链参数，建立在分子水平上的基于DNA编码机理和DNA控制机理的遗传信息模型，提出一种基于 DNA 计算的高光谱遥感数据光谱匹配方法。其核心思想是：首先利用DNA编码提取光谱所携带的地物信息，然后通过DNA操作从样本光谱中训练得到各类地物最典型光谱编码，建立地物DNA编码库，通过DNA计算原理构建的模糊规则，实现各类地物的DNA光谱匹配分类识别。

2 DNA 计算基本概念

DNA 计算的开创性计算思想来源于生物体遗传信息的传递规律(Adleman,1994)。

2.1 生物遗传信息传递规律

在自然界中，生物体表现出不同物种的多样性和相同物种的相似性，这是由生物体中的遗传物质脱氧核糖核酸(deoxyribonucleic acid, DNA)决定的。DNA 中有 4 种碱基，即腺嘌呤(adenine, A)、鸟嘌呤(guanine, G)、胞嘧啶(cytosine, C)和胸腺嘧啶(thymine, T)，各个碱基之间的不同组合就构成了异常丰富的信息。DNA 包含大量的遗传密码，通过生化反应传递遗传信息，这个过程是生命现象的基本特征之一。

生物遗传信息传递过程如图 1。生物体通过DNA 序列简单的操作，实现遗传信息的传递与表达。生物体所具有的复杂的结构实际上是编码在DNA序列中的原始信息经过简单的处理得到的。在数学领域，求一个含变量的可计算函数的值也可以通过求一系列含变量的简单函数的复合来实现。这是生物智能与数学过程的一个重要的共性，也是DNA计算的出发点。因此, DNA计算的本质就是利用大量不同的核酸分子杂交，产生类似某种数学处理的一种组合的结果，并对其进行筛选的过程。(Ding 等, 2002)。

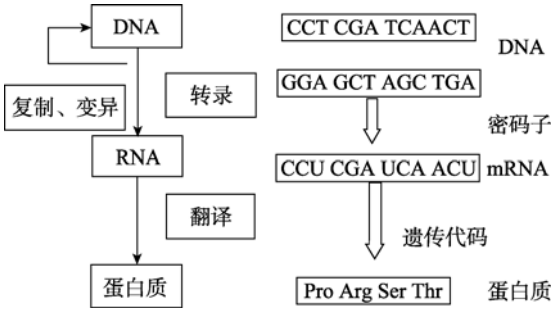

图 1 生物遗传信息传递过程

2.2 DNA 计算的模型、理论与方法

根据 DNA 的生物背景知识，充分发挥 DNA 计算的开创性思想，模拟生物的遗传机制和基因协调机理，提出了基于 DNA 计算的信息模型。在传统计算机上，利用 DNA 编码表示复杂知识或系统，分析和模仿遗传信息调控系统的自生成、自组织功能，可以将 DNA 计算与人工智能系统集成起来。基于DNA 编码的信息模型将会加深智能系统的理论研究，拓宽其应用范围。DNA 计算模型已被用于人工免疫系统(Ren & Ding, 2001)、遗传算法(Shin 等,2005)等生物智能计算中。DNA 计算模型由两部分构成：DNA 编码和 DNA 基因操作。

2.2.1 DNA 编码及种群初始化

DNA 链由四种碱基 A,T,C,G 的结合体构成，可以表示多个基因。DNA 链初始化，即是通过字符集{A,T,C,G}编码形成染色体，即 DNA 链。DNA 编码是一个关键点环节，DNA 链的长短将直接影响问题求解的精度和收敛速度。

DNA 种群中每个 DNA 链的密码子转译为氨基酸相应的参数值，并按照氨基酸的之间的差异性作为 DNA 种群适应度评价函数。若评价函数值高，表示该 DNA 个体具有较高的适应度。

2.2.2 DNA 基因操作

通过DNA基因操作能够产生新一代DNA种群，再进行适应度评价，并选择种群中优势个体进入下一代种群，如此循环、迭代使群体中个体的适应度和平均适应度不断提高，直到最优个体的适应度达到某一限值或最优个体的适应度和种群的平均适应度不再提高，则迭代过程收敛。

DNA计算模型从DNA编码的初始DNA种群出发，模拟人类遗传进化过程，最后选择优秀的种群和 DNA 个体，满足求解问题的要求。

3 高光谱遥感数据的 DNA 计算分类方法

结合光谱匹配与DNA计算的特点，构建在分子水平上的基于 DNA 编码与调控机理的遗传信息模型，提出高光谱遥感数据的 DNA 计算分类方法。利用 DNA 编码模型表达地物光谱丰富的重要特征信息，通过 DNA 操作对 DNA 编码的调节与控制，获取最典型地物编码，实现光谱编码匹配的最优化。

3.1 DNA 计算模型初始化

(1) DNA 种群参数选择：DNA 模型需要设定

DNA 种群个体数量, 交叉概率, 变异概率, 循环终止条件。

(2) 样本选择: DNA 模型的光谱匹配样本分为两部分, 包括训练样本与验证样本。首先, 参照相应高分辨率影像和地面真实数据, 获得各个类别的地物样本, 然后从中随机选取一定数量的像元分别作为训练样本和测试样本。从训练样本中选择一定数量的光谱, 作为 DNA 初始种群光谱。

3.2 光谱特征 DNA 编码

DNA 计算应用于高光谱遥感数据匹配时,首先要解决的问题是如何将光谱曲线利用 DNA 编码进行表达。根据高光谱遥感数据维数多, 数据量大的特点, DNA 编码需要满足以下 3 个条件: (1)符合 DNA 模型的四值编码方式, {A,G,C,T}, 保留更多光谱细节信息; (2)提取典型光谱的物理反射特征; (3)具备容差性能, 减弱噪声对光谱识别分类的干扰。

针对DNA模型的要求, 结合现有的光谱编码方式, 提出了一种 DNA 编码方法。DNA 编码由“DNA_A”编码和“DNA_B”编码两个部分构成, 其中 DNA_A 部分编码通过光谱曲线均值信息衡量光谱整体趋势; DNA_B 部分编码通过相邻光谱值的差异描述光谱细节特征信息。具体编码方式如下:

DNA_A 部分　首先对像元光谱向量取平均值, 得到阈值 T_0, 将像元属性值划分为$[x_{\min},T_0]$和$[x_0,T_{\max}]$两个区间, 确定两个区间的像元; 分别对两个区间的光谱值取均值, 得到两个新阈值 T_l 和 T_r, 最终形成四个区间$[x_{\min},T_l)$, $[T_l,T_0)$, $[T_0,T_r)$和$[T_r,T_{\max})$, 分别对应 T,C,A,G 4 个编码方式, 如公式(1)。对每个像元的光谱向量, 根据其各波段属性值所处的区间, 分别赋值相应的编码。当光谱波段数为 L 时, DNA_A 编码长度为 L。

$$S_i^{\text{DNA\_A}}=\begin{cases}T & S_i\in[x_{\min},T_l)\\ C & S_i\in[T_l,T_0)\\ A & S_i\in[T_0,T_r)\\ G & S_i\in[T_r,x_{\max}]\end{cases} \tag{1}$$

DNA_B 部分　假定一条光谱向量为, $S=(s_1,s_2,\cdots s_L)^{\mathrm{T}}$, L 代表光谱曲线的整体波段数目, S_l 代表第 l 个光谱波段, 设定光谱偏差的容限值Δ。若相邻两个波段的差值的绝对值$|s_i-s_{i-1}|>\Delta$, 则认为这两个相邻光谱变化显著, 否则光谱反射率值变化在限定的范围之内。并且, 若 $s_i-s_{i-1}>0$, 则光谱反射率显著上升; 若 $s_i-s_{i-1}<0$, 则光谱反射率显著下降。通过设定光谱偏差的容限值Δ, 能够减少噪声对光谱编码的干扰, 使得编码提取光谱本质的细节特征信息。本文中设定的光谱偏差容限值Δ由公式(2)确定。

$$\Delta=(1/(L-1))\sum_{l=2}^{L}|r_i-r_{i-1}| \tag{2}$$

式中, L 表示光谱的波段数码, r_i 为第 i 波段的光谱反射率值。

按照光谱反射率值变化的显著性程度将 3 个连续波段的光谱曲线变化情况归纳为如图 2 所示的 4 种情况(Chang 等,2009):

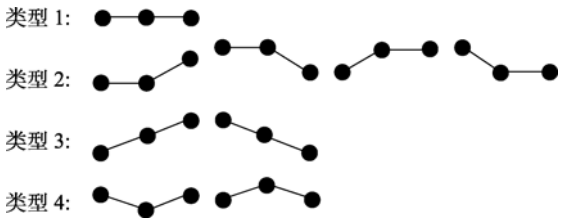

图 2　光谱曲线特征形态示意图

其数学表达如公式(3):

$$\begin{aligned}
&\text{类型 1：如果 } |s_i-s_{i-1}|\leqslant\Delta \text{ 与 } |s_{i+1}-s_i|\leqslant\Delta\\
&\text{类型 2：如果 } (|s_i-s_{i-1}|\leqslant\Delta \text{ 与 } |s_{i+1}-s_i|>\Delta)\\
&\qquad\text{或 } (|s_i-s_{i-1}|>\Delta \text{ 与 } |s_{i+1}-s_i|\leqslant\Delta)\\
&\text{类型 3：如果 } (s_i-s_{i-1}<-\Delta \text{ 与 } s_{i+1}-s_i<-\Delta)\\
&\qquad\text{或 } (s_i-s_{i-1}>\Delta \text{ 与 } s_{i+1}-s_i>\Delta)\\
&\text{类型 4：如果 } (s_i-s_{i-1}<-\Delta \text{ 与 } s_{i+1}-s_i>\Delta)\\
&\qquad\text{或 } (s_i-s_{i-1}>\Delta \text{ 与 } s_{i+1}-s_i<-\Delta)
\end{aligned} \tag{3}$$

根据以上条件, 将光谱变化类型转换到DNA编码值 T,C,A,G 上, 并将其赋予 3 个连续波段的中间波段的编码值。

$$S_i^{\text{DNA\_B}}=\begin{cases}\text{T} & \text{如果 } S_i \text{ 是 类型 1}\\ \text{C} & \text{如果 } S_i \text{ 是 类型 2}\\ \text{A} & \text{如果 } S_i \text{ 是 类型 3}\\ \text{G} & \text{如果 } S_i \text{ 是 类型 4}\end{cases} \tag{4}$$

因此, 对于长度为 L 的光谱向量, 其 DNA_B 编码长度为 $L-2$。

综上所述, 光谱 DNA 编码可以表示为 S^{DNA}, 其长度为 $2L-2$ 位。

$$\begin{aligned}S^{\text{DNA}}=\{&S^{\text{DNA\_A}},S^{\text{DNA\_B}}\}=\{S_1^{\text{DNA\_A}},S_2^{\text{DNA\_A}},\\&\ldots,S_L^{\text{DNA\_A}},S_2^{\text{DNA\_B}},S_3^{\text{DNA\_B}},\ldots,S_{L-1}^{\text{DNA\_B}}\}\end{aligned} \tag{5}$$

DNA 编码完成了高光谱数据由光谱空间至 DNA 编码空间的转换。图 3 显示了光谱数据 DNA 编码的一个简单示例, 图 3 (a)表示原始光谱数据, 波段 $L=80$, 其成像波段范围为 0.417—0.854μm, 图 3(b)表示编码后的光谱数据, 编码长度为 $2L-2$。

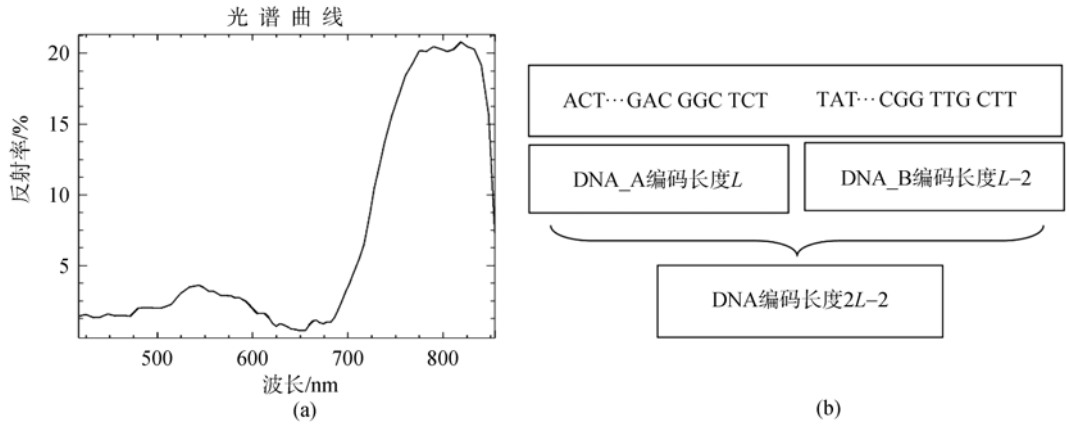

图 3 光谱数据 DNA 编码示意图

(a) 原始光谱数据; (b) DNA 编码后光谱数据

DNA 编码方式采用四值编码形式, 记录了光谱曲线吸收、反射特征以及光谱曲线的整体趋势, 提取了光谱曲线的物理特征; 通过对光谱偏差容限值 Δ的调控, 可以减弱噪声对编码的影响, 使得 DNA 编码更为稳定。

3.3 高光谱数据光谱匹配 DNA 基因调控机制

3.3.1 DNA 转译

DNA 链的转译是模拟从生物 DNA 到蛋白质形成的转译过程(Ren & Ding, 2001), 即先由 DNA 转录, 拼接成 *m*RNA。再将 *m*RNA 中由 3 个连续碱基组成的密码子, 对应为氨基酸, 64 种密码子对应 20 种氨基酸, 20 种氨基酸对应[0, 19]区间中的某一个数。转译规则如表 1。这种转换的主要目的是通过引入 DNA 链密码子转译框架, 构建 DNA 链模糊匹配规则, 将原始光谱特征 DNA 编码转换为氨基酸参数链; 在蛋白质层次上进行样本训练和分类操作。

表 1 DNA 链的密码子转译成参数值的基本框架

第 1 个碱基	第 2 个碱基				第 3 个碱基
	T	C	A	G	
T	Phe(0)	Ser(2)	Tyr(3)	Cys(4)	T
	Phe(0)	Ser(2)	Tyr(3)	Cys(4)	C
	Leu(1)	Ser(2)	Stop(9)	Stop(9)	A
	Leu(1)	Ser(2)	Stop(9)	Try(9)	G
C	Leu(1)	Pro(5)	His(6)	Arg(8)	T
	Leu(1)	Pro(5)	His(6)	Arg(8)	C
	Leu(1)	Pro(5)	Gln(7)	Arg(8)	A
	Leu(1)	Pro(5)	Gln(7)	Arg(8)	G
A	Ile(11)	Thr(12)	Asn(13)	Ser(2)	T
	Ile(11)	Thr(12)	Asn(13)	Ser(2)	C
	Met(10)	Thr(12)	Lys(14)	Arg(8)	A
	Met(10)	Thr(12)	Lys(14)	Arg(8)	G
G	Val(15)	Ala(16)	Asp(17)	Gly(19)	T
	Val(15)	Ala(16)	Asp(17)	Gly(19)	C
	Val(15)	Ala(16)	Glu(18)	Gly(19)	A
	Val(15)	Ala(16)	Glu(18)	Gly(19)	G

DNA 转译机制进一步加强了编码系统的稳定性和容差性能。

3.3.2 适应度计算

通过衡量 DNA 种群个体与 DNA 操作训练样本之间氨基酸参数的绝对实际距离之和, 获得种群中个体适应度。

DNA 种群与样本光谱进行 DNA 编码, 形成 DNA 编码链。然后, 在 DNA 链密码子转译参数表的框架下, 将 DNA 链 S^{DNA} 转译为氨基酸参数链 S^{Amino}, 如公式(6)。

$$S^{\mathrm{DNA}} \to S^{\mathrm{Amino}} = \{a_1, a_2, a_3, \ldots, a_T\}, \tag{6}$$

其中, $T=\lfloor(2\times L-2)/3\rfloor$, a_i=0,1,2,⋯,19。

由公式(7)计算 DNA 种群个体适应度。

$$\mathrm{Fitness} = \sum_{i=0}^{N} |\, S_{\mathrm{Train}\_i}^{\mathrm{amino}} - S_{\mathrm{individual}}^{\mathrm{amino}} \,| = \sum_{i=0}^{N}\sum_{j=0}^{T} |\, a_{ij} - a'_j \,| \tag{7}$$

其中, $S_{\mathrm{Train}\_i}^{\mathrm{amino}} = \{a_{i1}, a_{i2,} \ldots, a_{iT}\}$; $S_{\mathrm{individual}}^{\mathrm{amino}} = \{a'_1, a'_2, \ldots, a'_T\}$; N为训练样本个数; $T=\lfloor(2\times L-2)/3\rfloor$。

3.3.3 遗传操作

按照与 DNA 种群个体适应度相应的概率, 从 DNA 训练样本中选取部分个体, 组成新 DNA 种群。对 DNA 种群进行交叉和变异操作, 重新计算种群适应度, 并将新种群中适应度最高的个体标记为最佳 DNA 个体。

交叉算子是 DNA 遗传操作的核心, 体现了自然界信息交换的思想。本文采用的交叉算子是标准交叉算子。标准交叉中, 其后代个体是基于一个随机产生的交叉特征码, 对父代进行操作而得到的。若某一位置上交叉特征码为 0, 则其后代的碱基不变; 反正, 其后代的碱基由双亲互换得到, 也就是说所产生的后代个体是父代个体碱基序列的混合, 如图 4。

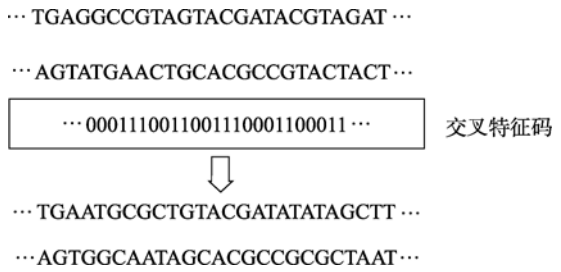

图 4 标准交叉示意图

本文中的变异算子只考虑碱基的替换，即染色体中某个碱基或多个碱基从一种状态突变为另一种状态。碱基的变异有两种方式：一种是同类型碱基转换变异，如胞嘧啶 T 变成胸腺嘧啶 C；另一种是异类型碱基颠换变异，如胞嘧啶 T 变成腺嘌呤 A。变异算子由变异特征码实现。

交叉、变异算子的特征码由事先设定的交叉、变异概率随机生成。

3.3.4 循环与循环终止

种群最佳 DNA 个体适应度达到设定阈值或循环次数达到最大循环次数。如不满足循环终止条件，则继续进行 DNA 操作。

通过 DNA 基因遗传操作获取各类别典型 DNA 光谱编码链。

3.4 分类

对于高光谱影像中的每一像元，分别计算该像素 DNA 与各个类别典型 DNA 链之间氨基酸参数距离，并将最小距离的类别属性赋予该像元。

由以上分析可知，高光谱遥感数据的DNA计算分类方法流程，如图 5。

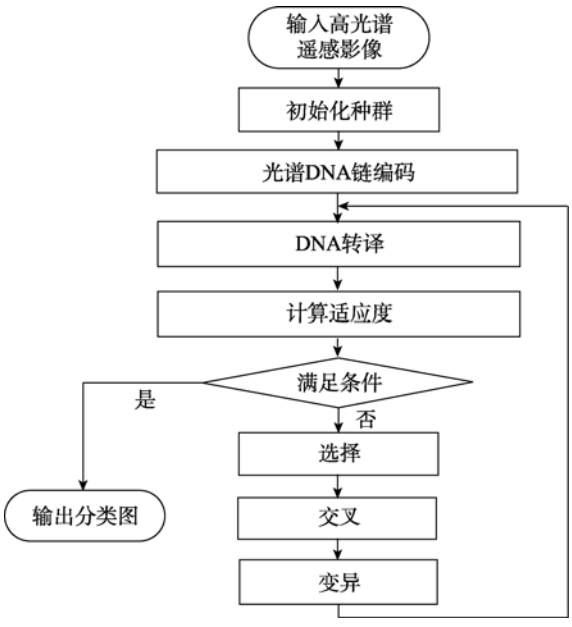

图 5 高光谱遥感数据的 DNA 计算分类方法流程图

4 实验结果

4.1 实验数据

研究区域是常州市夏桥，采用的是国产 PHI(推扫式光谱成像仪)遥感影像(340×390 像元)，波段数为 80 波段，该区域是一个混合的农业区域，其成像波段范围为 0.417—0.854μm。图 6 显示了该区域的高光谱影像立方体图。通过实际调查后，预期将该影像分为六类，具体为道路、农作物、碎石地、菜地、荒草地和水体。各类地物的光谱曲线如图 7。

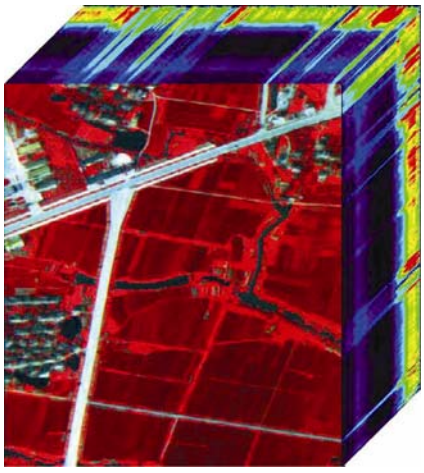

图 6 夏桥 PHI 影像立方体 RGB(70,40,10)波段数 80 (0.41—0.85μm)

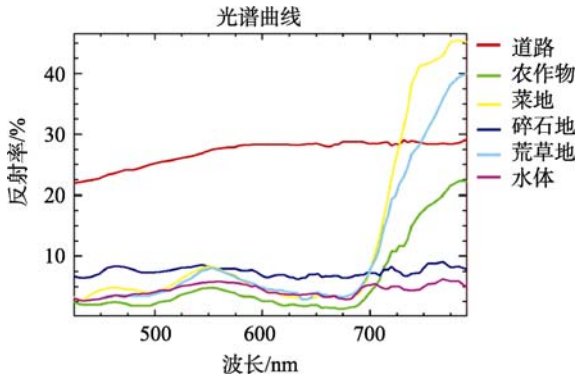

图 7 6 种地物参考光谱的光谱曲线

由于 DNA 计算光谱匹配分类算法能够提取地物光谱物理反射特性，并具备光谱容错性能，因此在进行分类实验前，不对高光谱影像进行辐射校正、波段选择和滤波处理。

4.2 数据输入

4.2.1 实验样本输入

本实验中，针对 6 类地物，分别选取 180 个样本光谱，其中随机选择 80 个作为 DNA 操作训练样本，

100 个作为分类后验证样本。

4.2.2 算法参数输入

本文中设定的种群个体数量为 20, 交叉概率为 0.9, 变异概率为 0.02, 迭代次数为 200 次, 循环终止阈值设定为 0.98。

4.3 分类结果

利用本文提出的基于 DNA 计算的高光谱遥感数据光谱匹配分类算法对夏桥影像进行分类,得到的分类结果为图 8(d)。为了证明这种分类算法的有效性, 采用传统的二进制编码光谱匹配分类、光谱角光谱匹配分类、SDFC 编码光谱匹配分类与此算法进行比较, 得到的分类结果分别为图 8(a), 图 8(b)和图 8(c)。

从图 8 的分类结果看, 采用二进制编码光谱匹配分类时错分现象比较严重, 图像右侧区域的农作物类别错分成菜地。这主要由于图像辐射强度不均匀, 右侧辐射较强; 同时, 二值编码损失大量光谱细节信息, 无法提取地物物理反射特征, 导致农作物与菜地的光谱编码混淆, 如图 8(a)。

光谱角匹配分类图像中出现大量的椒盐噪声, 将大部分菜地划分为荒草地类别, 存在未分类现象, 经判读和调查, 未分类的像元与实地情况不符, 如图 8(b)。SDFC 编码光谱匹配算法没有很好的区分荒草地和菜地这两种光谱形态类似的类别, 并存在椒盐噪声, 如图 8(c)。这两种光谱匹配的方式, 对光谱噪声较为敏感, 因此分类结果出现了椒盐噪声和错分现象。

基于 DNA 计算的光谱匹配分类算法对高光谱数据进行分类时, 错分现象明显减少, 菜地、荒草地和农作物得到了正确的分类。从人工目视判读上说明本文方法的光谱匹配分类算法效果优于传统的光谱匹配分类方法。

4.4 精度比较

为了进一步地验证 DNA 计算光谱匹配的遥感图像分类算法的有效性, 将本文该光谱匹配分类方法与传统光谱匹配方法(二值编码匹配方法、光谱角匹配方法、SDFC 编码光谱匹配方法)进行分类精度的定量比较。比较方法采用常用的分类精度比较方法(Foody, 2002) , 即通过混淆矩阵计算各类别生产者精度、消费者精度、总精度和 Kappa 系数。

表 2 对 4 种分类方法生产者精度和消费者精度进行了比较, 表 3 对 4 种分类方法的总精度和 Kappa 系数进行了比较。由表 2 可见, 二值编码匹配、光谱角匹配、SDFC 编码光谱匹配各类别的生产者精度和消费者精度表现得不稳定, 如 SDFC 编码光谱匹配荒草地的生产者精度为 45%, 二值编码光谱匹配菜地分类消费者精度为 54.44%。与之相比, DNA 模型光谱匹配分类的生产和消费者精度较为稳定, 所有精度值大于 80%; 由表 3 可见, 4 种分类方法的总精度由传统方法的 79.00%, 83.17%, 81.17%提高

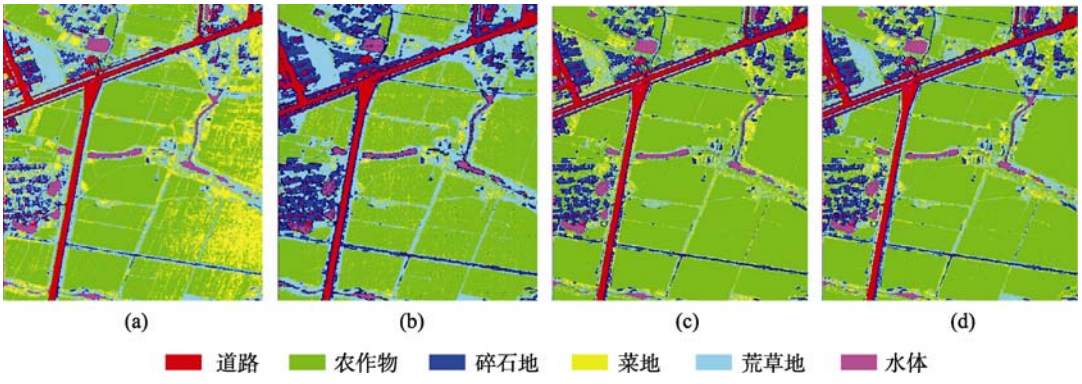

图 8 4 种方法分类结果比较

(a) 二值编码光谱匹配; (b) 光谱角光谱匹配; (c) SDFC 编码光谱匹配; (d) DNA 计算光谱匹配

表 2 4 种分类方法生产者精度和消费者精度比较

/%

	类别	道路	农作物	碎石地	菜地	荒草地	河流
二值编码光谱匹配分类	生产者精度	97	51	65	86	75	100
	消费者精度	89.81	92.73	97.01	54.44	77.32	91.74
光谱角光谱匹配分类	生产者精度	100	76	91	66	93	73
	消费者精度	91.74	95	86.67	68.75	75.61	92.41
SDFC 编码光谱匹配分类	生产者精度	95	99	87	68	45	93
	消费者精度	100	72.79	87.88	61.26	76.27	93
DNA 计算光谱匹配分类	生产者精度	95	100	87	83	82	94
	消费者精度	100	81.30	90.63	87.37	91.11	93.07

表 3 4 种分类方法总精度和 Kappa 系数的比较

方法 精度	二值编码 光谱匹配	光谱角 光谱匹配	SDFC 编码 光谱匹配	DNA 计算模型 光谱匹配
总精度/%	79.00	83.17	81.17	90.17
Kappa 系数	0.7480	0.7985	0.7740	0.8820

到 DNA 计算光谱匹配分类方法的 90.17%, Kappa 系数也有较大提高, 由光谱角匹配分类方法的 0.7480 提高到 DNA 计算光谱匹配分类方法的 0.8820。

从表 2 和表 3 中的统计数据可以得出基于 DNA 计算的光谱匹配分类算法的分类正确率要明显优于传统的光谱匹配方法。这是因为传统的基于光谱匹配的分类算法, 无法利用统计模型和相邻像元间的相关信息, 因此这些方法对于误差是敏感的。DNA 编码方式具有自组织和自学习的能力, 不仅能较好的保留光谱曲线特征信息, 而且通过 DNA 模型的优化过程, 能够寻找到最典型的光谱 DNA 编码, 实现优化的光谱匹配。

5 结论与讨论

(1) 本文提出了一种基于 DNA 计算的高光谱数据光谱编码匹配算法。该方法利用 DNA 编码{A,G,C,T}提取各个类别光谱曲线的典型特征, 实现地物光谱特征的有效表达; 利用 DNA 基因操作自组织、自适应的智能性能, 优化典型光谱匹配 DNA 编码, 实现地物光谱的较高精度识别与分类。

本文的主要研究思想是将高光谱遥感数据按照 DNA 构成原理进行重组、转化, 并不是真正要对实际地物进行 DNA 测量, 而是如同利用小波、傅里叶变换等数学工具可进行数据的空域与频域转换一样, 在这里是按照优化原理将已有的地物光谱数据转化为相应的 DNA 链参数, 使得传统的对原有地物光谱数据进行的直接处理转化为在 DNA 层次上的分析与处理, 从而利用 DNA 计算的原理、优势进行高光谱遥感数据的信息处理。通过高光谱遥感数据实验证明, 基于 DNA 计算的光谱匹配方法减少了光谱差异性对光谱匹配分类的影响, 提高了光谱匹配分类精度。

(2) DNA 计算光谱匹配分类识别方法针对的是高光谱遥感图像, 其图像光谱曲线一般为光谱地面视反射率曲线。由于 DNA 编码提取的是光谱曲线表现出来的物理反射特征, 因此同类地物表现为相同或相近的编码, 进行匹配分类前不需要对高光谱图像进行辐射校正; 同时, DNA 编码与 DNA 转译机制使得基于 DNA 计算的光谱匹配分类具备容错性能, 匹配分类前也不需要对高光谱遥感图像进行波段选择和滤波操作。

(3) 通过夏桥 PHI 高光谱数据实验看出, 本文提出的光谱匹配精度要高于传统的光谱匹配方法, 其分类精度和 kappa 系数分别为 90.17%和 88.20%。

(4) 在对高光谱数据进行 DNA 编码后仅利用了生物计算里传统的 DNA 遗传操作, 如交叉和变异, 其目的主要是检验 DNA 编码对于高光谱数据分类的适应性, 而对于 DNA 计算特有的基因遗传操作, 包括倒位、分离等将结合 DNA 编码在下一步研究中展开。

REFERENCES

Adleman L. 1994. Molecular computation of solutions to combinatorial problems. *Science*, **266**(5187): 1021—1024

Benenson Y, Gil B and Ben-Dor U. 2004. An autonomous molecular computer for logical control of gene expression. *Nature*, **429**: 423—429

Chang C I, Chakravarty S, Chen H and Ouyang Y. 2009. Spectral derivative feature coding for hyperspectral signature analysis. *Pattern Recognition*, **42**: 395—408

Chen J, Li H, Sun K and Kim B. 2003. How will bio-informatics impact signal processing. *IEEE Signal Processing Magazine*, **20**: 16—26

Ding Y S, Shao S H and Ren L H. 2002. DNA Computing and Soft Computing. Beijing: Science Press

Faulhammer D, Cukras A R, Lipton R J and Landweber L F.2000. Molecular computation: RNA solutions to chess problems. *ProcNatl Acad. Science*, **97**(4): 1385—1389

Foody G M. 2002. Status of land cover classification accuracy assessment. *Remote Sensing of Environment*. **80**: 185—201

Jia X and Richards J A. 1993. Binary coding of imaging spectrometer data for fast spectral matching and classification. *Remote Sensing of Environment*, **43**: 47—53

Lipton R J. 1995. DNA solution of hard computational problem. *Science*, **268**: 542—545

Mazer AS, Martin M, Lee M and Solomon J E. 1988. Image processing software for imaging spectrometry analysis. *Remote Sensing of Environment*. **24**(1): 201—210

Qian S, Hollinger A B, Williams D and Manak D. 1996. Fast three-dimensional data compression of hyperspectral imagery using vector quantization with spectral-feature-based binary coding. *Optical Engineering*, **35**(11): 3242—3249

Ren L H and Ding Y S. 2001. Design of fuzzy control system by a new DNA-based immune genetic algorithm. The 10th IEEE International Conference Fuzzy System. Melbourne, Australia, Dec. 2-5, 244—247

Shin S Y, Lee I H, Kim D and Zhang B T. 2005. Multi-objective evolutionary optimization of DNA sequences for reliable DNA computing. IEEE Transactions on Evolutionary Computation. **9**(21): 143—158

Tong Q X, Zhang B and Zhen L F. 2006.Hyperspectral Remote Sensing. Beijing: Science Press

附中文参考文献

丁永生, 邵世煌, 任立红. 2002. DNA 计算与软计算, 科学出版社

童庆禧, 张兵, 郑兰芬. 2006. 高光谱遥感——原理、技术与应用, 科学出版社