Genetic Feature Selection for Texture Classification

PAN Li  ZHENG Hong  ZHANG Zuxun  ZHANG Jianqing

ABSTRACT  This paper presents a novel approach to feature subset selection using genetic algorithms. This approach has the ability to accommodate multiple criteria such as the accuracy and cost of classification into the process of feature selection and finds the effective feature subset for texture classification. On the basis of the effective feature subset selected, a method is described to extract the objects which are higher than their surroundings, such as trees or forest, in the color aerial images. The methodology presented in this paper is illustrated by its application to the problem of trees extraction from aerial images.

KEYWORDS  genetic algorithms; feature selection; texture classification; fuzzy c-mean

CLC NUMBER  TP751

Introduction

Texture feature selection is the most important and difficult task in the field of image texture analysis and classification. Feature selection is the problem of choosing a small subset of features that is necessary and sufficient to describe target concept. The importance of feature selection is due to the potential for speeding up the processes of both concept learning and classifying objects, reducing the cost of classification, and improving the quality of classification. Feature selection has long been the focus of researchers of many fields such as pattern recognition, image understanding and machine learning. In general, the existent methods can be classified into two categories: ① The filter or open-loop approach[2] do not consider the effect of selected features on a whole processing algorithm performance. ② The wrapper or closed-loop methods[2] are based on feature selection as a wrapper around a classifying algorithm relying on which the relevant attributes are determined. Due to considering the classification result of each feature subset, the wrapper approaches have proved to be more effective than the filters. The approach to feature subset selection proposed in this paper is an instance of wrapper approach. It utilizes genetic algorithms to find out effective feature subset. The feature subsets are evaluated by computing the generalization classifying accuracy.

The work presented here was motivated by our experience in using conventional feature selection algorithms for difficult image understanding problems involving texture classification. In the case there can be a lot of features and complex interactions among the features. On such problems conventional features selection algorithms can be used, but there are two problems. One is that it is difficult for conventional features selection algorithms to find out the global optimal solution or the effective feature subset because the problem searching the minimum feature subset is a difficult optimization problem. The other is that some methods need special conditions. In order to solve these problems, this paper employs the genetic algorithm to implement feature selection. Some recent attempts in applying GAs...
for feature selection are reported\textsuperscript{2,3}. Our algorithm was inspired by the method proposed by Jihoon Yang\textsuperscript{2}, which offered an approach to feature subset selection for neural network pattern classifier of some medical cases. In this paper, our application is in the domain of aerial images. The genetic algorithms are used to find out the effective texture feature subset. Based on the selected effective feature subset, we present the algorithm to extract the objects which are higher than their surrounding in color aerial images. The methodology presented in this paper is illustrated by its application to the problem of trees extraction from aerial images.

1 Genetic feature selection

Feature selection techniques generally involve both a search algorithm and a criterion function. This paper employs genetic algorithm to search the effective feature subset and the generalization accuracy of classification is taken as the criterion function, namely the fitness function.

The origin of genetic algorithms (GAs) is attributed to Holland’s work on cellular automata. The range of application of GAs includes such diverse areas as job shop scheduling, training neural nets, image analysis and classification. The main issues in applying GAs to any problem are selecting an appropriate representation, an adequate evaluation function, and defining a suitable genetic operation. These issues will be described as follows.

1.1 Encoding scheme

Now a set of individuals are defined in a population generated during \(t\) generation cycles. \(P(t) = \{I_k | k = 1, 2, \ldots, m\} \), where \(m\) is the number of individuals or the population size. The size affects both the ultimate performance and the efficiency of GAs. Each individual is generated by some encoded form known as a chromosome. Here a chromosome is the vector consisting of a single bit for each feature, with a 1 indicating that the feature participates in classification, and a 0 indicating that it is omitted.

1.2 Crossover and mutation

The operations, namely, crossover and mutation, are then following the encoding scheme. A crossover position is randomly chosen in the encoded string. The crossover operation applied to two parents produce offsprings. Mutation is carried out by selection a value 0 or 1 on a randomly selected positions of a parent string.

1.3 Fitness function

The fitness function is used to evaluate the goodness of a chromosome (solution). In this study, the fitness function has to combine two different criteria the accuracy of the classification function realized by the fuzzy c-mean method and the cost of performing classification. The accuracy of the classification function can be estimated by calculating the correct rate of each reduct decision table using the fuzzy c-mean method. The measures of the cost of classification suggest the number of condition attributes in the decision table needed for classification. Here, we choose a relatively simple form of a 2-criteria fitness function defined as follows:

\[
\text{Fitness}(i) = \frac{\text{correct rate}(i)}{1 + \lambda \times \text{num}(i)}
\]

where \(\text{Fitness}(i)\) is the fitness of the feature subset represented by individual \(i\), and \(\text{correct rate}(i)\) is the test accuracy of the fuzzy c-mean classifier using the feature subset represented by \(i\). Because the cost of classification is in proportion to the number of the selected features, \(\text{num}(i)\) is set to the number of the condition attributes or features of subset represented by individual \(i\). In addition, the parameter \(\lambda\) is the weight of the number of features, and in this paper \(\lambda = 0.01\). Obviously, the greater the value of fitness is, the better the performance of the selected feature subset is. The following contents describe that how the effective feature subset is used in texture classification.

2 Texture classification based on the selected feature subset

We have applied feature selection for the pur-
The pose of texture classification in color aerial images. In this paper, trees and forests are interesting objects. Trees and forests are a kind of natural scenes which are not structured and cannot be represented easily by regular rules. In contrast to artificial objects such as houses and roads, trees or forests do not obey strict position rules. In addition, the appearance of trees and forests can vary greatly based on the geographic area, the season, the current weather conditions, or the past weather conditions. Hence, texture and color features are important cues for trees extraction in color aerial images. In practice, if only texture and color features are used, the results of trees extraction are inaccurate. There are two reasons: (1) The non-tree objects have the similar texture features to trees; (2) Usually different tree type has different texture features. But, it is the fact that trees are the objects that are higher than their surroundings. Our method combines texture features, color and height information to overcome these disadvantages. In recent years, some approaches to extract the objects height information like digital terrain model (DTM) have been proposed. A technique similar to ours was discussed in Reference [5], which deals with realistic and thus more complex scenes. But it needs high resolution aerial images and DTM. In this paper, low resolution aerial images and digital elevation model (DEM) are used in trees extraction. First, according to the DEMs, the original color aerial images are segmented into the high and low objects. High objects include trees, houses, bridges and so on. In order to refine trees or forest, the high objects are classified by fuzzy c-mean based on the effective feature subset. The procedure is described as follows.

2.1 Distinguishement between high and low objects

We start with DEM data automatically generated by the digital photogrammetry system -Virtuozo.

First, DEM data is mapped into images, which is called DEM images. According to the DEM images, high and low objects are distinguished.
Given: The samples and their feature values.

Step 1: All the values of features are normalized to the range from 0 to 1.

Step 2: The samples is classified by the effective feature subset.

Calculating the fuzzy membership of the feature values.

\[ U_i = \left( \frac{\sum_j D_k}{D_a} \right)^{-1} \forall i, k \quad \forall j \sum_k U_i X_k \forall i \quad (2) \]

where \( D_a \) is some measure of similarity between \( v_i \) and \( x_k \) or the attribute vectors and the cluster centers of each region; \( v = (v_1, v_2, \cdots, v_c) \) are geometric cluster prototypes; \( U \) denotes the fuzzy membership matrix of pixel block \( k \) in cluster \( i \); \( c \) denotes the number of cluster.

\[ U = [u_{ik}], 1 \leq i \leq c, 1 \leq k \leq n \quad (3) \]

Determine the feature value membership:

\[ J_m(U, v, X) = \sum_{i=1}^{c} \sum_{k=1}^{n} (\mu_k)^m D_a \quad (4) \]

where the real number \( m \in [0, \infty) \) is a weighting exponent on each fuzzy membership. As \( J_m \) is iteratively minimized, \( v_i \) became more stable. Iteration terminated when \( u_{ik}^{(a)} - u_{ik}^{(a-1)} < \beta \) or the maximum number of iterations is reached. Here, \( \alpha \) is the number of iteration and \( \beta \) is predefined tolerance.

Step 3: Compute the accuracy according to the category of samples.

\[ A = \frac{\eta}{\text{sum}} \quad (5) \]

where \( A \) is the accuracy of classification; \( \eta \) is the sample number of the samples classified correctly; \( \text{sum} \) is the sample sum.

3 Experimental results

In experiment, 12 color aerial images are used. The photography scale of color aerial images is 1 : 8 000. The principle focal is 152, 987 mm. The scanning resolution is 96 \( \mu \)m. The photo size is 23 cm \( \times \) 23 cm. A total of 200 samples are selected, which include density trees, sparse trees, houses, roads, grass, river and ground. First, the color aerial images samples in the RGB space are converted into in the HIS (hue, saturation, and intensity) space. In the intensity of the color aerial images samples, a total of 11 texture features per pattern (pixel) are computed. In addition, five color features are used, which include hue, saturation and normalized color \( (I_1, I_2 \) and \( I_3) \). Table 1 lists all features used in our experiments.

| No. | Feature | Model |
|-----|---------|-------|
| 1   | Mean    | Local statistics |
| 2   | Standard deviation | Local statistics |
| 3   | Skewness | Local statistics |
| 4   | Kurtosis | Local statistics |
| 5   | Contrast | CLCM |
| 6   | Entropy | CLCM |
| 7   | Inertia | CLCM |
| 8   | Energy(E5L5) | LTT |
| 9   | Energy(E5R5) | LTT |
| 10  | Energy(E5S5) | LTT |
| 11  | Energy(L5S5) | LTT |
| 12  | I1      | RGBSCM |
| 13  | I2      | RGBSCM |
| 14  | I3      | RGBSCM |
| 15  | hue     | HIS |
| 16  | saturation | HIS |

In addition, the parameters used in the genetic algorithm are listed as follows:

- Population size: 50
- Number of generation: 500
- Probability of crossover: 0.9
- Probability of mutation: 0.5

The original color aerial images used in the experiment of trees extraction are shown in Figs. 1(A), 1(B), 1(C), 1(D). First, according to DEM, the original color aerial images are segmented into high and low objects. Then the effective feature subset is selected. In high objects, trees are refined by fuzzy c-means with the effective feature set. The results are shown in Figs. 1(a), 1(b), 1(c), 1(d) and the statistical data are listed in Table 2. The best rate is 96.7% and the worst rate is 88.3%. The main reasons for faulty classification are:

(1) In initial segmentation, the low trees are classified into low objects.

(2) Because all shadows can not be eliminated by using DEM, some shadows cause some faulty results.
Notes: White area denotes trees, and black area denotes non-trees.

**Fig. 1** Original color aerial images ((A), (B), (C) and (D)) and results of classification (a), (b), (c) and (d)

**Table 2** Statistical results of trees extraction in the color aerial images shown in Fig. 1
(The total area of trees are computed in the image coordinate system)

| No.   | Size/pixel | Total area of trees/mm² | False/mm²  | Rate/(%) |
|-------|------------|-------------------------|------------|----------|
| Fig. 1(a) | 203 × 186   | 2,948,160               | 112,320    | 96.1     |
| Fig. 1(b) | 260 × 240   | 987,936                 | 117,024    | 88.2     |
| Fig. 1(c) | 413 × 296   | 6,798,048               | 326,592    | 95.1     |
| Fig. 1(d) | 436 × 491   | 6,805,344               | 711,936    | 89.5     |

(Continued on Page 173)
where". This is its shortcoming. But the authors still consider that using this mode to develop software on Web is a good solution.

REFERENCES

1 Dong H B, Zhang H F (2001) Developing Web database using Servlet. Computer Application Research, 2(2):100-102 (in Chinese)
2 Wang M (2001) The development of large-scale seamless image database (GeoImageDB) and the feasible research on measurable virtual reality, [Ph. D dissertation]. Wuhan, Wuhan University. (in Chinese)
3 Wang M, Gong J Y, Li D R (2001) Spatial seamless data organization of large scale image database. Geomatics and Information Science of Wuhan University, 26(5):419-424 (in Chinese)
4 Wang Y F, Yang C J, Hao P W et al. (2001) Use image based JPEG2000 in Web GIS. Journal of Remote Sensing, 5(5):388-395 (in Chinese)
5 Yang C W, Li Q, Cheng J C (2000) Research and implementation on Web publication of remote sensing images. Journal of Remote Sensing 4(1), 71-75 (in Chinese)

4 Conclusions

This paper describes an approach to find out the effective feature subset using genetic algorithm. Our application is in the domain of aerial images. The values of texture features are continuous data. To select effective feature subset, the fitness function is designed in this paper. Besides, on the basic of the effective feature subset selected, this paper presents an approach to extract the objects which are higher than their surroundings, such as trees or forests. The experiment results show that the feature subset selected and the method of classification are effective and practical. Although we have dealt with color aerial images, the technique is extensible to remote sensing images.

REFERENCES

1 Liu H, Setiono R (1996) A probabilistic approach to feature selection-a filter solution. Machine Learning: LCML’96, Bari, Italy
2 Yang J H, Honavar V (2001) Feature subset selection using a genetic algorithm. Journal of Intelligent Information Systems, 16: 215-228
3 Vafaie H, Jong K D (1993) Robust feature selection algorithms. International Conference on Tools with AI, Boston, Massachusetts.
4 Dang T, Jamet O, Maitre H (1994) Applying perceptual grouping and surface models to the detection and stereo reconstruction of building in aerial imagery. International Archives of Photogrammetry and Remote Sensing, Atlanta, USA.
5 Haala N (1994) Detection of building by fusion of range and image data. International Archives of Photogrammetry and Remote Sensing, Atlanta, USA.
6 Eckstein W, Steger C (1996) Fusion of digital terrain model and texture for object extraction. 2nd Airborne Remote Sensing Conference, San Francisco.
7 Pan L (2001) The Study on forest area recognition from color aerial Image and its application to automatic aerial triangulation, [Ph. D dissertation]. Wuhan, Wuhan University. (in Chinese)