Comparative Studies on Feature Extraction Methods for Multispectral Remote Sensing Image Classification

Yanqin Tian and Ping Guo
Department of Computer Science
Beijing Normal University
Beijing, 100875, China
pguo@bnu.edu.cn

Michael R. Lyu
Department of Computer Science & Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong, China
lyu@cse.cuhk.edu.hk

Abstract – *Feature extraction of multispectral remote sensing image is an important task before classifying the image. When land areas are clustered into groups of similar land cover, one of the most important things is to extract the key features of a given image. Usually multispectral remote sensing images have many bands, and there may have been much redundancy information and it becomes difficult to extract the key features of the image. Therefore, it is necessary to study methods regarding how to extract the main features of the image effectively. In this paper, five methods are comparatively studied to reduce the multi-bands into lower dimensions in order to extract the most available features. These methods include the Euclid distance measurement* (*EDM*)*, the discrete measurement criteria function* (*DMCF*)*, the minimum differentiated entropy* (*MDE*)*, the probability distance criterion* (*PDC*)*, and the principle component analysis* (*PCA*) *method. The advantage and disadvantage of each method are evaluated by the classification results.*

Keywords: Multispectral Remote Sensing Image, Dimension Reduction Methods, Feature Extraction.

1 Introduction

The number of Earth observation satellites that are in operations is rising every year. These satellites carry a diverse spectrum of radar and optical sensor capital of accruing imageries which are applied in many fields such as generating classification maps. Before the classification, feature extraction is an important processing procedure. With extracted features, a classifier is built to recognize the interested objects in remote sensing image. There are two kinds of classification: supervised and unsupervised. In general, when we have little knowledge about given image, we have to adopt unsupervised classification techniques. Among the unsupervised methods, the finite mixture model analysis has many advantages [1][2] and it attracts many researchers' interest in image segmentation as well as other applications [3][4]; whereas in this paper we adopt a finite mixture model as a classifier. When building a classifier, we assume that the data in the feature space as a mixture of Gaussian probability density distribution, and the finite mixture model is used to cluster the extracted features. The expectation-maximization (EM) algorithm can be used to estimate the model parameters, and final Bayes decision is applied to classify these data in the feature space [5].

Gray value is an important characteristic for the analysis of various types of remote sensing images. It is believed that the gray value plays an important role in the visual systems for recognition and interpretation of given data. Furthermore, texture analysis is an important research field in remote sensing image processing, as the texture describes the attribution between a pixel and the other pixels around it [6]. Texture feature extraction must be considered based on a small region, not a single pixel. However, texture analysis method has shortcomings, such as the edge between different classes may be incorrectly classified. Therefore, gray value is adopted as the features of the image in this paper. There exist a number of dimension reduction methods in the literature; here we investigate five dimension reduction methods [7]. These methods are the Euclid distance measurement (EDM), the discrete measurement criteria function (DMCF), the minimum differentiated entropy method (MDE), the probability distance criterion (PDC), and the principle component analysis (PCA). We reduce the dimensions for the purpose that the feature information may not be redundant and the convergent speed of estimating the parameters of classifiers may be accelerated. Classification accuracy is used to assess these methods.

2 Dimension reduction methods

In this paper, we focus on comparative studying the five methods to reduce the dimensions. These dimension reduction methods are described in the following paragraphs. Although we can find the theoretic description of the first four methods in the reference [7], few researchers have applied these theories to real applications. When the first four methods are applied to analyze the multispectral remote sensing image, we suppose that the original image has *D* bands, and the bands reduced into *d* dimensions after data dimension reduction. We can define the original feature data vector as **y**, the transformed data vector as **x**, where $\mathbf{y} = [y_1, y_2, \cdots, y_D]^{\mathrm{T}}$, $\mathbf{x} = [x_1, x_2, \cdots, x_d]^{\mathrm{T}}$, and the transformation formula is:

$$\mathbf{x}=\mathbf{W}^{\mathrm{T}}\mathbf{y}\ . \tag{1}$$

W is the combination of the *d* dimension eigenvectors of a spectral matrix **T**, where the eigenvectors are corresponding to the first *d* maximum eigenvalues, and **W** is a *D*×*d* dimension matrix.

2.1 EDM method

In the method of EDM, **W** is the combination of the *d* dimension eigenvectors of matrix $S_w^{-1}S_b$. S_b is the discrete measurement matrix among different classes and S_w is the discrete measurement matrix in the same class [7]:

$$S_w=\sum_{i=1}^{c}P_iE_i\left[(y-\mu_i)(y-\mu_i)^T\right] \tag{2}$$

$$S_b=\sum_{i=1}^{c}P_i\left(\mu_i-\mu\right)\left(\mu_i-\mu\right)^T \tag{3}$$

where *c* is the number of all the classes, μ_i is the average vector of the *i*th class, μ is the average vector of all the vector data and P_i is the prior probability function of the corresponding *i*th class. E_i is the expectation between the vector data and the average vector of the *i*th class.

How to compute the transformation matrix **W** is illustrated with following numerical example.

For example, there are two classes, which have the same prior probability.

The corresponding mean vectors are:

$\mu_1=[1,3,-1]^T$, $\mu_2=[-1,-1,1]^T$.

The corresponding covariance matrices are:

$$\Sigma_1=\begin{bmatrix}4&1&0\\1&4&0\\0&0&1\end{bmatrix},\ \Sigma_2=\begin{bmatrix}2&1&0\\1&2&0\\0&0&1\end{bmatrix}$$

The mean vector μ is $\mu=\frac{1}{2}(\mu_1+\mu_2)=[0,1,0]^T$, the discrete measurement matrix among different classes S_b and the discrete measurement matrix in the same class S_w computed respectively as follows:

$$\begin{aligned}S_b&=\frac{1}{2}\sum_{i=1}^{2}(\mu_i-\mu)(\mu_i-\mu)^T\\&=\frac{1}{4}(\mu_1-\mu_2)(\mu_1-\mu_2)^T\\&=\begin{pmatrix}1&2&-1\\2&4&-2\\-1&-2&2\end{pmatrix}\end{aligned}$$

$$\begin{aligned}S_w^{-1}&=(S_w)^{-1}\\&=\left(\frac{1}{2}(\Sigma_1+\Sigma_2)\right)^{-1}\\&=\frac{1}{8}\begin{pmatrix}3&-1&0\\-1&3&0\\0&0&8\end{pmatrix}\end{aligned}$$

There is only one eigenvalue of $S_w^{-1}S_b$, so **W**=*w*. Then $S_w^{-1}S_bw=\lambda w$, or $\frac{1}{4}S_w^{-1}(\mu_1-\mu_2)(\mu_1-\mu_2)^T w=\lambda w$. In this equation, $\frac{1}{4}(\mu_1-\mu_2)^T w$ is a scale value, so $\mathbf{W}=w=S_w^{-1}(\mu_1-\mu_2)=\frac{1}{4}(1,5,-8)^{\mathrm{T}}$.

2.2 DMCF method

For the DMCF method, **W** is the eigenvector system of the matrix **T**. For the sum matrix between every two classes, **T** is described as the following equation:

$$\mathbf{T}=\sum_{i=1}^{c}\sum_{j=1}^{c}\left(\Sigma_i^{-1}+\Sigma_j^{-1}\right)M_{ij} \tag{4}$$

Σ_i, Σ_j are the covariance of the *i*th and *j*th classes, and the covariance can be described as

$$\begin{aligned}\Sigma_j&=E\left\{\left(y-\mu_j\right)\left(y-\mu_j\right)^T\right\}\\&=\frac{1}{l}\sum_{i=0}^{l}\left(y_i-\mu_j\right)\left(y_i-\mu_j\right)^T\end{aligned} \tag{5}$$

$M_{ij}=(\mu_j-\mu_i)(\mu_j-\mu_i)^T$; μ_i and μ_j are the mean vectors of the *i*th and *j*th classes, respectively.

2.3 MDE method

For the MDE method, **T** is the differentiated entropy. For two classes,

$$\mathbf{T} = V(p,q) + V(q,p) \tag{6}$$

where $V(p,q)$ is the relative entropy, the definition of which can be described as the following:

$$V(p,q) = -\sum p(y_i)\log\left[p(y_i)/q(y_i)\right] \tag{7}$$

$p(y_i)$, $q(y_i)$ are the distributed prior probability functions of the two classes. On the assumption that, for remote sensing image, the prior probability is the gray value when it is read by using the computer machine. And Equation 6 can be written as

$$\begin{aligned}\mathbf{T}(p,q) = & \\ & -\sum p(x_i)\log p(x_i) - \sum q(x_i)\log q(x_i) \\ & +\sum p(x_i)\log q(x_i) + \sum q(x_i)\log p(x_i)\end{aligned} \tag{8}$$

For more than two classes, **T** becomes

$$\mathbf{T} = \sum_{i=1}^{c}\sum_{j=1}^{c}\left(V(p_i,p_j) + V(p_j,p_i)\right). \tag{9}$$

where **T** means the summation of all the two different classes' relative entropy.

2.4 PDC method

For the method of PDC, generally **W** is the combination of the *d* dimension eigenvectors of the eigenvector system $\mathbf{T} = \sum_2^{-1}\sum_1$ where $\sum_1$, $\sum_2$ are the covariance matrices of the two classes, respectively. Here, system **T** is supposed as the following hypothesis: the mean vectors of every class are equal.

If the mean vectors are not equal and the covariance matrices are equal which are described as $\sum$, then the eigenvector system **T** can be described as the following equation:

$$\mathbf{T} = \sum^{-1}\left(\mu_2 - \mu_1\right) \tag{10}$$

For more than two classes, the eigenvector system **T** can be written as

$$\mathbf{T} = \sum_{i=1}^{c}\sum_{j=1}^{c}\sum_j^{-1}\sum_i \tag{11}$$

2.5 PCA method

For the method of PCA, we can refer to the definitions in reference paper [8]. And the transform formula is $\mathbf{x}=\mathbf{W}^T(\mathbf{y}-\mathbf{m})$; **W** is the combination of the *d* dimension eigenvectors of the covariance of the image, in which the corresponding eigenvalues are the maximal ones. And *W* is a $D \times d$ matrix and **m** is the data mean vector.

From the detailed description of each dimension reduction method, we can know that except the PCA method, all the other methods need to assign each pixel to a class label at first. However, usually we have little prior knowledge about of each pixel's class membership. In order to resolve this problem, we adopt the random sample method, which means we can first assign each pixel to a class randomly.

3 Experiments

In order to speedup the convergent rate while estimating the parameters, we use the gray histogram method to initialize the mean vectors and the covariance matrices. However, if an image contains many classes, the peaks of the histogram are not distinct from each other. It is very difficult to determine which classes the peaks in the histogram should belong to and to find proper parameters for initialization before applying the EM algorithm. In this case, only random initialization parameter method can be adopted.

How to judge whether the feature extraction methods are good or not? In this paper, under the same classification circumstance we assess the feature extraction methods by using the classification accuracy. For the same testing data, if the classification accuracy is the highest, we think this feature extraction method is the best.

The finite mixture model is adopted to analyze the multispectral remote sensing images and the Expectation-Maximization (EM) [9] algorithm is used to estimate the parameters. With this iterative EM algorithm, the mixture parameters can be estimated until the likelihood function reaches a local minimum value.

Redner [2] has proved that the EM algorithm was convergent and assured likelihood function could be close to a local minimum value. Perhaps there are many local minimum values for a given function. In this paper, the parameters are adopted when the local minimum values reach the smallest one.

With the pre-assigned classification region number *k*, the posterior probability can be described as: $P\left(j=1\middle|x_i\right)$, $P\left(j=2\middle|x_i\right),\cdots P\left(j=k\middle|x_i\right)$, we use Bayes decision

$j^* = \arg\max P\left(j \middle| x_i\right)$ to classify x_i into cluster j^*. This procedure is called Bayesian probabilistic classification.

The unsupervised classification method is adopted because we can get better results in the case where there is a lack of prior knowledge about remote sensing images.

The testing remote sensing images are from the database of platform Landsat-5, which was launched on March 1 in 1984 by USA, and the remote sensor was thematic mapper (TM). For the 6th band the resolution is 120 meters, and for other bands, the resolution is 30 meters. All the data are TM images of Beijing, China in 1996 and all the data can be classified at least two classes including water and other geographical objects. Then the original remote sensing data have 7 bands, and for better and easy clustering, only 3 bands are used after processing features.

In this paper five simple multispectral remote sensing images are adopted as the testing data, the original remote sensing images are shown in Figure 1, and the classification accuracies can be seen in Table 1. Figure 2 and Figure 3 are the graphic display of the accuracies of all the feature extraction methods investigated in this work.

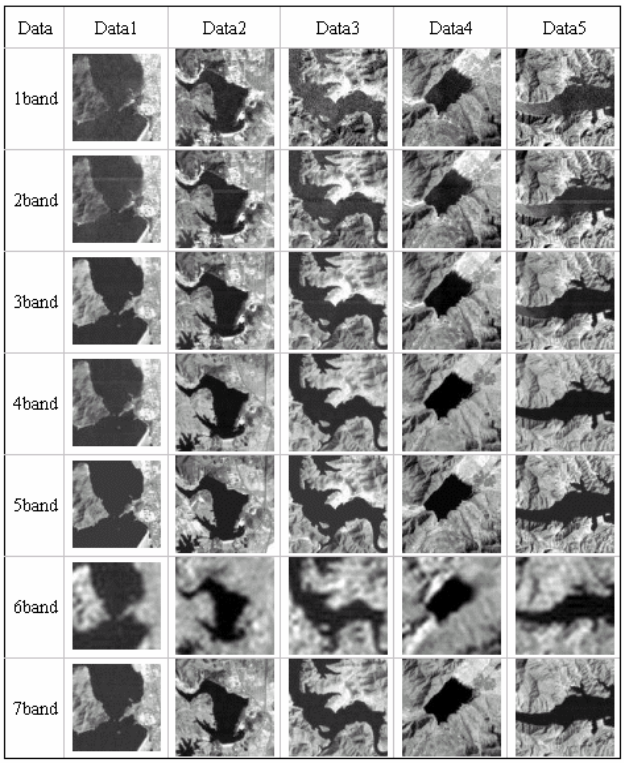

Figure 1. The original remote sensing images

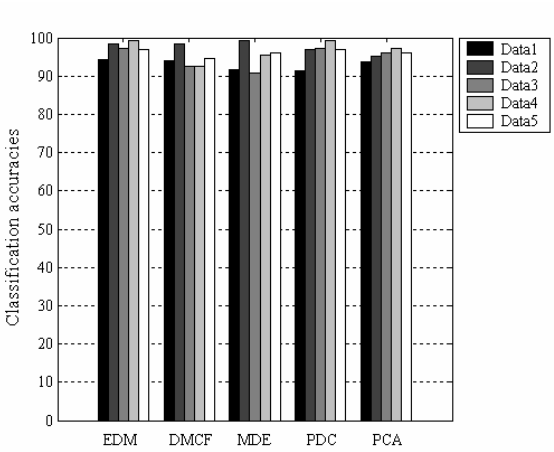

Figure 2. Comparison of different methods

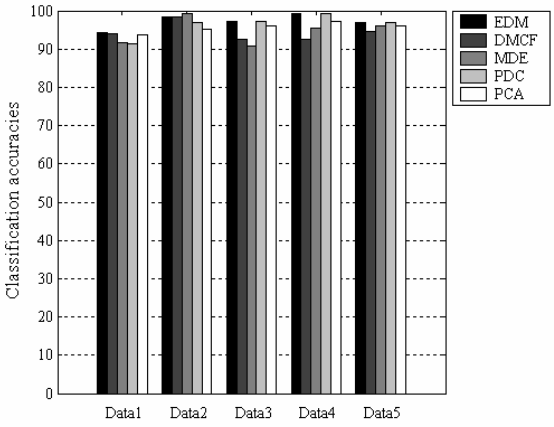

Figure 3. Comparison of different data sets

Table 1. Classification accuracies

Methods	EDM	DMCF	MDE	PDC	PCA
Data1	94.17%	94.04%	91.74%	91.23%	93.80%
Data2	98.31%	98.37%	99.19%	96.83%	95.28%
Data3	97.13%	92.45%	90.68%	97.13%	96.04%
Data4	99.33%	92.54%	95.43%	99.33%	97.14%
Data5	96.89%	94.62%	96.18%	96.89%	96.14%

Table 2. Rank of the feature extraction methods

Methods	EDM	DMCF	MDE	PDC	PCA
Data1	1	2	4	5	3
Data2	3	2	1	4	5
Data3	1	4	5	1	3
Data4	1	5	4	1	3
Data5	1	5	3	1	4
Average	1.4	3.6	3.4	2.4	3.6

From figure 2, we can know that:

For all of the dimension reduction methods, the classification accuracies are higher than 90% regardless of data set, which validate the effectiveness of investigated methods.

From the results we can find that the features of Data4 are obviously better extracted when using the EDM, DMCF, MDE, PDC and PCA methods. In other words, for the same feature extraction methods, the effectiveness is data dependent. In the experiments, using the Data2 and Data4 can get higher classification accuracy than other data set.

By analyzing Figure 3, Table 1 and Table 2, we can get following points:

For Data1, the features extracted with the EDM method are suitable to cluster, and the classification accuracy is 94.17%; For Data2, the MDE method is the best and the classification accuracy is as high as 99.19%; For Data3, the EDM or PDC method are the same in feature extraction and the classification accuracy can reach to 97.13%. For Data4, the classification accuracy is 99.33% based on the EDM or PDC method. Finally for Data5, 96.89% classification accuracy is obtained based on EDM or PDC feature extraction method.

In summary, the EDM and PDC methods are better than the other dimension reduction methods for all data sets used in this paper. For the same testing image, with the EDM and PDC methods, especially with the EDM method, we can get better features for classification. When extracting the features of multispectral remote sensing images, the EDM and PDC methods should be chosen firstly.

In practical applications, if obtained remote sensing image has a similar data structure to that of Data3, Data4 or Data5, when extracting the key features of the image, it is suggested that the EDM or PDC method should be considered firstly. If the data structure is similar to that of Data1, first of all we should choose the EDM method. But if the data structure is similar to that of Data2, the MDE method is the best one for the purpose of extracting the key features.

4 Conclusions

In this paper, five feature extraction methods are comparatively studied. The results may be different with different data sets, but as a whole the EDM and PDC methods are better while extracting the key features of the multispectral remote sensing images. Occasionally, MDE is also a better method to extract the key features, and the DMCF and PCA methods are the worst ones among all of the five feature extraction methods. Therefore, when classifying the remote sensing images, we suggest that the EDM or PDC method should be used to extract the features in order to obtain higher classification accuracy.

Acknowledgement

The research work described in this paper was fully supported by a grant from the National Natural Science Foundation of China (Project No. 60275002) and a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK4182/03E).

References

[1] A. P. Dempster, N. M. Laird, D. B. Rubin, “Maximum-likelihood from incomplete data via the EM algorithm”, *J. Royal Statist. Society (B)*, Vol. 39, No. 1, pp. 1-38, 1997.

[2] R. A. Redner, H. F. Walker, “Mixture densities, maximum likelihood and the EM algorithm”, *SIAM Review*, Vol. 26, No. 2, pp. 195-239, 1984.

[3] P. Santago, H. D. Gage, “Statistical models of partial volume effect”, *IEEE Trans. Image Processing*, Vol. 4, No. 11, pp. 1531-1540, 1995.

[4] S. Sanjay-Gopal, T. J. Hebert, “Bayesian pixel classification using spatially variant finite mixtures and the generalized EM algorithm”, *IEEE Trans. Image Processing*, Vol. 7, No. 7, pp. 1014-1028, 1998.

[5] P. Guo, H. Lu, “A Study on Bayesian Probabilistic Image Automatic Segmentation”, *Acta Optica Sinica*, Vol. 22, No. 12, pp. 1479-1483, 2002.

[6] B. anjunath, “Texture Features for Browsing and Retrieval of Image Data”, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 8, pp. 837-842, 1996.

[7] Z. Bian, X. Zhang, *Pattern Recognition*, Tsinghua University Press, Beijing, 2002.

[8] S. Lee, H. C. Kim, D. Kim, and YoungSik Choi, “Face Retrieval Using 1st- and 2nd-order PCA Mixture Model”, *Lecture Notes in Computer Science*, Vol. 2668, pp. 391-400, 2003.

[9] E. M. Mohamed, P.W. Robert, D. Ridder, V. Atalay, “Texture Segmentation Using the Mixtures of Principal Component Analyzers”, *Lecture Notes in Computer Science*, Vol. 2869, pp. 505-512, 2003.