Utilizing UPCA and SPCA in Unsupervised Classification Using Landsat TM data

Byung-Gul Lee and In Joon Kang[1)]

Associate Professor, Cheju National Univ. Dept. of Civil and Environmental Engineering
[1)] Professor, Pusan National Univ. Dept. of Civil Engineering

개요

본 연구는 무감독영상해석(Unsupervised Classification)에서 주성분 분석법(Principal Component Analysis)의 응용성을 연구하기 위하여, 주성분 분석법을 K-means, ISODATA 두가지 무감독분류법에 적용하였다. 적용대상지역은 제주도이다. 본 연구에서 주성분 분석 방법중에서 비정규형 주성분 분석방법 (Unstandardized PCA)과 정규형 주성분 분석방법(Standardized PCA) 두가지 경우로 나누어서 각각 연구하였다. 이를 위하여 제주도의 Landsat TM영상과 국토연구원에서 조사한 제주도 식생분류 조사자료와 현장조사 자료 그리고 1/25,000 수치지도를 이용하였다. 그리고 분석된 자료의 정확도를 평가하기 위하여 오차행렬(Error Matrix)을 도입하여 계산하였다. 우선 비정규형 주성분 분석법으로 구한 주성분 영상과 Landsat TM 원래 영상을 오차행렬을 이용하여 제주도의 식생 분류에 각각 적용하였다. 그 결과, K-means 무감독분류법에서는 Landsat TM 자료를 직접 이용한 경우에는 바다와 육상의 분류가 잘 되지 않았으며, 또한 전반적인 영상분류결과가 관측치와 많은 차이를 보였다. 그러나, 주성분 분석법으로 계산된 주성분 영상으로 K-means방법으로 분류 한 결과는 관측치와 잘 일치를 하였다. ISODATA의 경우, Landsat TM 원래영상을 계산하면, K-means으로 분류한 결과보다는 좋은 값을 나타냈으나, 주성분 분석법으로 구한 영상의 계산결과와 비교하면, 주성분 영상으로 구한 분류결과의 정확도가 약 15%정도 높게 나타났다. 정규형 주성분 분석법의 경우를 보면 K-means에서는 Landsat TM 원래 자료보다 우수한 결과를 보여주었으나, 비정규형 주성분 분석법으로 계산된 결과보다는 정확도가 다소 떨어지는 단점이 있었고, ISODATA의 경우도 Landsat TM원래 자료보다 약 7%정도의 높은 정확도를 보였으나, 비정규형 영상보다는 약8%정도 낮은 정확도를 보였다.

본 연구에서 주성분 분석법으로 계산된 결과에서 주목되는 것은, 주성분 분석법으로 구한 주성분 영상은 분류방법(K-means, ISODATA, artificial neural networks)에 따라 분류된 결과값이 비슷하게 나타난 반면, Landsat TM원래 자료는 분류방법에 따라 결과값이 많은 차이를 보여 주었다. 그리고 주성분 분석 방법 중에서도 비정규형 주성분 분석법(Unstandardized PCA)이 정규형 주성분 분석법(Standardized PCA)보다 영상분석에서 더 좋은 결과를 보여주는 것으로 나타났다.

주요어: 주성분분석법, 정규형 주성분분석, 비정규형 주성분분석, K-menas, ISODATA

1. Introduction

Principal component analysis (PCA) was used to improve image classification by three unsupervised classification techniques, the K-means, and the ISODATA.. To do this, I selected a Landsat TM scene of Jeju Island, Korea and proposed two methods for PCA: unstandardized PCA(UPCA) and standardized PCA(SPCA). The estimated the accuracy of the image classification of

Jeju area was computed by Error matrix. The error matrix was derived from three unsupervised classification methods. Error matrices indicated that classifications done on the first three principal components for UPCA and SPCA of the scene were more accurate than those done on the seven bands of TM data and that also the results of UPCA and SPCA were better than those of the raw Landsat TM data. The classification of TM data by the K-means algorithm was particularly poor at distinguishing different land covers on the island. For the ISODATA algorithm, the discrimination accuracy of the error matrix for the principal components of UPCA was approximately 15% and for the ones of SPCA was approximately 6% better than for the TM data. From the classification results, I also found that the principal component based classifications had characteristics independent of the unsupervised techniques (numerical algorithms) while the TM data based classifications were very dependent upon the techniques. This means that PCA data has uniform characteristics for image classification that are less affected by choice of classification scheme. And the PCA data were estimated statistically to find out relationship between PCA and raw Landsat TM data. In the results, we also found that UPCA results are better than SPCA since UPCA has wider range of digital number of an image. It means that the increased range of PCA image histogram values was an important role of image classification of Jeju Island. Finally, from the relationship of the probability density function, I derived a transformation function between PCA and Landsat TM data.

In this paper, the effects of using the unstandardized and the standardized Principal Component Analysis (PCA) with unsupervised classification results were considered. To do this, unsupervised classification results of raw Landsat TM data are compared with those of the unstandardized and the standardized principal components data, respectively. Three unsupervised techniques, K-means and ISODATA were used. The study area is Jeju Island, located off South Coast of Korea peninsula. The accuracy estimation of the PCA were performed by Error matrices(Confusion matrices) based on the three techniques. Finally, I conclude with a summary and a discussion of the limitations and the applicability of PCA as applied to unsupervised classification in Jeju Island.

2. Data and Method

In this study, I used the observation and the satellite data to estimate the calculated results of image classification. In remote sensing, the results of the classification of the satellite date usually have been compared to the observation data such as aerial photography and field observation using GPS or other survey instrument. For Jeju Island, the Korea Research Institute for Human Settlements (KRIHS) already classified the Island based on aerial photography, Landsat data, and field observation.

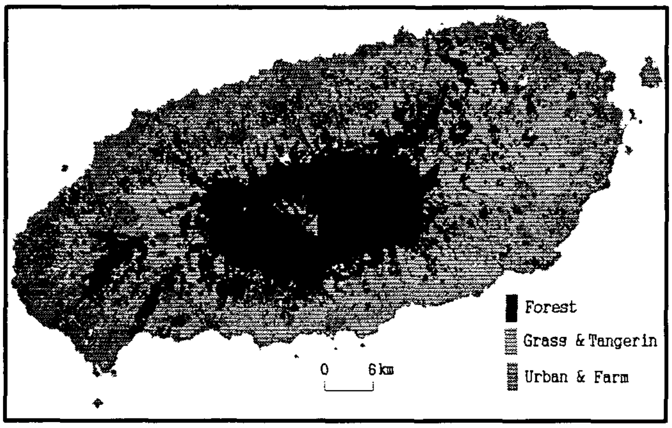

Figure 1. The reference classification based on Korea Research Institute for Human Settlements(KRIHS) in 1997.

To perform PCA, I need to diagonalize the covariance matrix of the remote sensing data. The covariance function *Var* between band *i* and *j* can be calculated as(Carr & Matanawi, 1999).

$$Var(k, l) = \frac{1}{NN}[\sum_{i=0}^{N}\sum_{j=0}^{N}(P_k(i,j) - P_{km})(P_l(i,j) - P_{lm}) \text{ --------------(1)}$$

in which N is the total number of pixels, k and l are band numbers. $P_k(i,j)$ and $P_l(i,j)$ are the pixel values (digital number) column j of row i in bands k and l. P_{km} and P_{lm} are mean of band k and l, respectively. In Landsat TM data, the covariance matrix will be 7 X 7 matrix since it has seven bands. The matrix is symmetric and a positive definite matrix if we use Landsat TM data. Using the matrix, the principal component can be calculated.

Diagonalizing the covariance matrix from Eq. (1) is the equivalent to solving standard eigenvalue problem as following:

$$Ax = \lambda x \text{ --------------------------------------(2)}$$

In Eq. (2), A is the covariance matrix, λ is the scalar of characteristic values, x is the matrix of eigenvectors.

From Eq.(2), the principal component image can be calculated as follows(Jensen, 1996)

$$OB_{i,j,p} = \sum_{k=1}^{n} a_{kp} IB_{i,j,k} \text{ ----------------------------------(3)}$$

in which k is band, i and j are pixel of row and column. k is band number, p is the principal component, and n is number of bands. *OB* is the new brightness value (digital number) at position *(i,j)* in principal component *p, IB* is the old brightness value and a_{kp} is the eigenvector in Eq.(3).

In previous section, the principal components calculated using the covariance matrix are referred to as unstandardized PCA and those calculated using the correlation matrix are referred to as standardized PCA. To calculate the rotation, we can start with either a variance-covariance matrix or a correlation matrix.

If one standardizes the data and calculates a variance-covariance matrix, then the result will be the same as a correlation matrix. Those that wish to practice their algebra can prove this by deriving the formula for the variance-covariance matrix and the correlation matrix calculated on raw data and then the variance-covariance matrix calculated on standardized data.

The standardized PCA is also an alternative method of computing a PC rotation is to derive the transformation matrix on the eigenvectors of the correlation matrix instead of the covariance matrix. The correlation matrix is equivalent to a covariance matrix for an image where each band has been standardized to zero mean and unit variance.

While it is less common to use this approach for normal remote sensing data sets, there are special situations when this method is preferable.

The correlation matrix can be calculated as

$$Cor(k,l) = \frac{Var(k,l)}{\sqrt{\sigma_k}\sqrt{\sigma_l}} \text{ ----------------------------------(4)}$$

where $\sigma_k = \sum_{i=1}^{n}(IB_{ik} - \mu_k)^2 \qquad \sigma_l = \sum_{i=1}^{n}(IB_{il} - \mu_l)^2$--------------------(5)

in which n is the total number of pixels, k and l are band numbers. IB_{ik} and IB_{il} are the pixel values of images in bands k and l. $\mu_k$and μ_l are means of band *k*th and *l*th images. σ_k and σ_l are variance of band *k*th and *l*th images, respectively.

With standardized PCA, the eigenvectors are computed from the correlation matrix. The characteristic of standardized PCA is to force each band to have equal weight in the derivation of the new component images(Eastman & Fulk, 1993).

3. Results and Discussion

In this study, I used the observation and the satellite data to estimate the calculated results of image classification. In remote sensing, the results of the classification of the satellite date usually have been compared to the observation data such as aerial photography and field observation using GPS or other survey instrument. For Jeju Island, the Korea Research Institute for Human Settlements (KRIHS) already classified the Island based on aerial photography, Landsat data, and field observation.

Unsatnadardized PCA(UPCA) and Standardized PCA(PCA) were applied to produce inputs to K-means and ISODATA unsupervised classifications for Jeju Island, respectively. The two PCAs(UPCA and SPCA) data would appear to be an excellent tool for the analysis of unsupervised classification of Landsat TM data. This research results are firstly implemented in the field of unsupervised classification based on remote sensing field.

The following summarize the conclusions achieved from this study:

First, the results in the error matrices of UPCA showed that the PCA data produced classification characterized by approximately 75 % correspondence built the KRIHS reference data. Using raw TM data produced less than 61%. The K-mean algorithm applied to the TM data produced confusion among the land cover classifications, while the PCA data clearly classified the Island as three parts. The classified image made by using ISODATA on the principal components was similar to the reference data and consistently a little better than that of the K-means algorithm. The results of using the K-means and the ISODATA algorithm on the PC data were almost same, while the results of applying the two algorithms to the raw TM data were much different.

Second, SPCA also produced the same results of those of UPCA although the accuracy of SPCA is less than that of UPCA. It is also found that the PCA data had independent characteristics not affected by an classification algorithm technique, whereas the classifications of the raw data were greatly affected by the algorithm employed. Another advantage of the PC data for classification was that fewer bands(the three PC images) were used for classifying than were used to classify the original TM image(seven TM bands). The results can be judged as PC transformed data is strongly affected by image enhancement effect. The reason of good results based on PCA data is strongly related with image enhancement effects.

References

Carr, James R. and Matanawi, K.(1999) Correspondence analysis for principal components transformation of multispectral and hyperspectral digital images, Photogrammetric Engineering and Remote Sensing. 65(8), 909-914.

Eastman, J.R. and Fulk, M.(1993) Long sequence time series evaluation using standardized principal components, Photogrammetric Engineering and Remote Sensing, 59(6), 991-996.

Jensen, J.R.(1996) Introductory digital image processing, Prentice-Hall, Englewood Cliffs, New Jersey, 316 p.