International Symposium on Applied Geoinformatics (ISAG2021)

PRINCIPAL COMPONENT ANALYSIS APPROACH IN HYPERSPECTRAL IMAGE CLASSIFICATION WITH MACHINE LEARNING METHODS

Saziye Ozge Atik[1], Muhammed Enes Atik[1]

[1] Istanbul Technical University, Faculty of Civil Engineering, Department of Geomatics, Istanbul, Turkey (donmezsaz@itu.edu.tr; atikm@itu.edu.tr); **ORCID 0000-0003-2876-040X**, **ORCID 0000-0003-2273-7751**

ABSTRACT

Hyperspectral data technology provides the opportunity to view the world with the help of hundreds of bands extending from ultraviolet (UV) to the infrared region. With the help of hyperspectral data, which is widely used in many environmental monitoring programs, geological minerals and ground observation studies, high detail level land cover and usage classes can be classified. In this study, using Indian Pines hyperspectral data with a large number of bands, it has been transformed into different sized datasets with the help of principal component analysis (PCA) to extract information efficiently. In these new dimensions, datasets are classified by support vector machines (SVM) and random forest (RF) machine learning algorithms. Finally, the classification results were compared quantitatively.

Introduction

Machine learning algorithms have been widely used for classification processes in recent years. Additionally, increasing the efficiency and ease of data processing, some dimension reduction methods are used to improve the analysis of feature information to be extracted from the data. PCA is a widely used size reduction method for extracting maximum information from the image. Therefore, various analyzes of hyperspectral images can be performed with size reduction methods (Agarwal, 2007). Using these techniques as a preprocessing step provides significant advantages (Rodormer and Jie, 2002). In this study, Indian Pines hyperspectral dataset was classified with support vector machines (SVM) and random forest (RF) using different band sizes that were selected by using PCA. The results of machine learning algorithms were compared quantitatively.

Materials and Methods

In the study, Indian Pines dataset (URL-1) was used, and the dataset includes public available hyperspectral scenes that were gathered by AVIRIS sensor. This dataset consists of 224 bands that are in the 0.4-2.5x10-6 m spectral range. The data has dimensions of 145 × 145 pixels and a spatial resolution of 20 m. Additionally, it contains 16 classes that mainly consist of land cover classes. In the data region, there are some agricultural classes as woodland, perennial vegetation. Also, the dataset includes a low density of settlements, minor roads, and some other structures. The image of the data set and the image of the reference map are shown in Figure 1. The sampling numbers of the classes differ, and the dataset is imbalanced.

Principal Component Analysis (PCA) (Cunningham, 2008) is a widely used dimension reduction technique. This technique is a way of identifying the pattern in the data and expressing it to highlight the similarities and differences of the data (Smith, 2002). new values of the data are generated with the PCA method. PCA is a powerful tool for analyzing high-dimensional data because the graphical representation of the high-dimensional data is problematic (Smith, 2002). The dimensions of hyperspectral images consist of hundreds of bands. Thus all the data process has difficulties for supervised classification. PCA uses statistical properties of hyperspectral bands to examine the correlation of bands with each other (Rodormer and Jie, 2002).

Support Vector Machines (SVM) (Cortes and Vapnik, 1995) is a machine learning approach used for both classification and regression. SVM creates a hyperplane that classifies the points in the data separately in an area of n features.

Random forest (Breiman, 2001) is defined as a non-parametric classification and regression algorithm. This algorithm calculates parameters for each part of the input space it divides into components. For generating a tree, two parameters are determined (Atik et al, 2021) Each independent tree in the random forest generates a prediction for the input class, and the class with the most votes is assigned as the input class. (Akar and Gungor, 2012).

This study reduced the number of bands in hyperspectral data by using PCA, a dimension reduction technique, as a preprocessing step. Classification tests were carried out at this stage by creating a data set with three different band numbers, 15, 35 and 55 bands. The hyperspectral image was classified by using SVM and RF using PCA bands. Several PCA bands samples were illustrated that were used in the experiments (Figure 2).

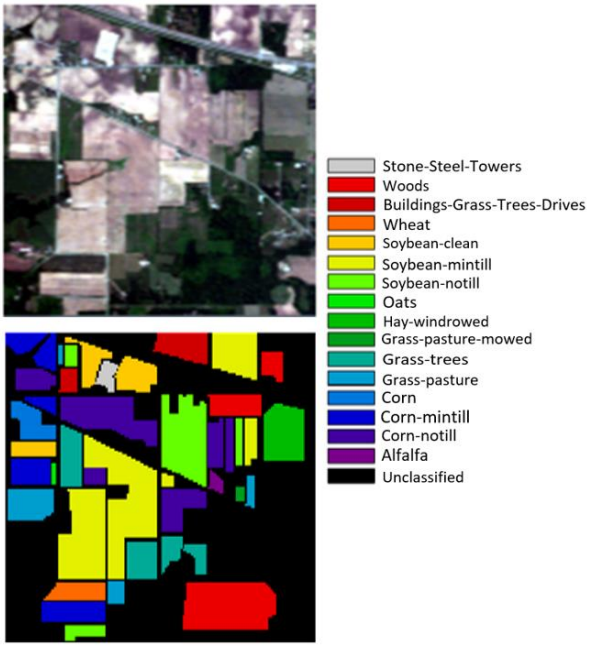

Figure 1 Original image(up), Ground truth (down)

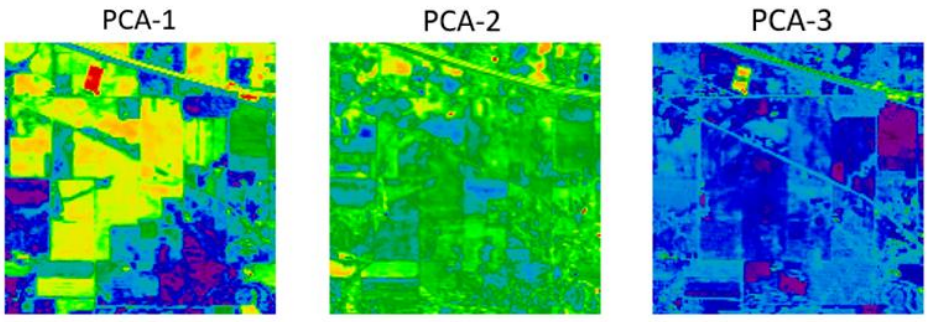

Figure 2 Samples of PCA bands

Results and Discussion

The SVM algorithm produced higher classification accuracy with 89.32% in the 55-band dataset, 88.63% in the 35-band dataset, and 85.22% in the 15-band dataset. In the RF method, the lowest classification accuracy among the experiments was performed on the 55-band dataset as 73.95%. On the other hand, the highest accuracy of the RF algorithm was obtained as 75.17% in the 15-band data set.

Table 1 Results of experiments as overall accuracy (The values are given as %)

Number of bands	SVM	RF
15 Bands	85.22	75.17
35 Bands	88.63	73.95
55 Bands	89.32	74.05

In this case, the best classification results of SVM method was obtained in the classification with the 55-band dataset. In contrast, in the RF algorithm, the highest accuracy was obtained in the 15-band dataset. It is shown that classification results as Land Use and Land Cover (LULC) maps for three bands combinations. (Figure 3).

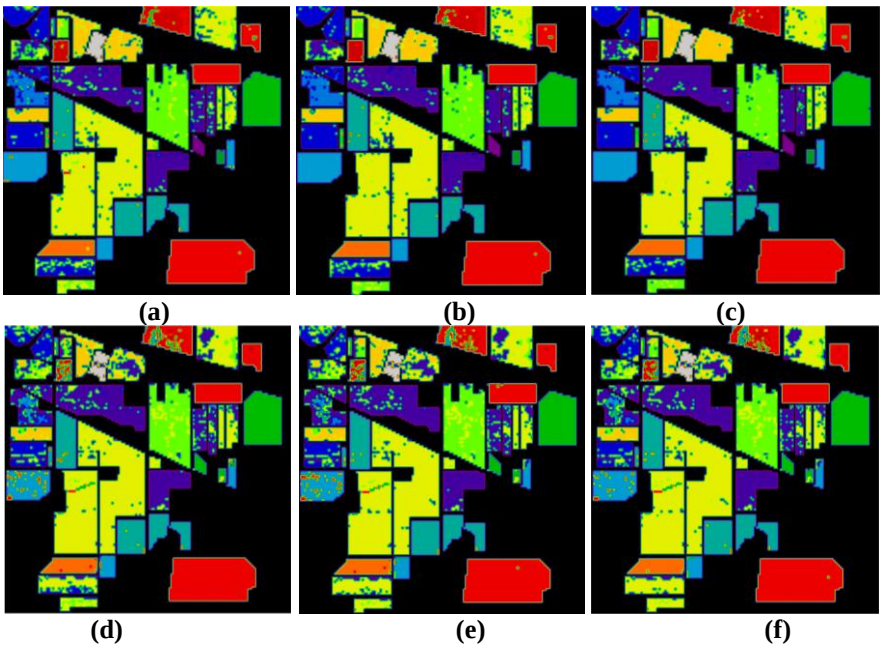

Figure 3 a, b and c as SVM results by order for 15,35,55 Bands, c,d and f as RF results by order for 15,35,55 Bands

Conclusion

According to these results, different machine learning algorithms have different classification accuracies in the hyperspectral image analysis when PCA approach was conducted. Furthermore, it has been seen that the size difference of the data set and the effect of the feature information on the classification accuracy according to the machine learning algorithm are not at the same rate. Therefore, in this case, while determining the ideal data size, tests should be done according to the machine learning algorithm.

Keywords: *Hyperspectral Image, Land Use and Land Cover, Principal Component Analysis, Machine Learning*

References

Agarwal, A., El-Ghazawi, T., El-Askary, H., Le-Moigne, J. (2007). Efficient Hierarchical-PCA Dimension Reduction for Hyperspectral Imagery, Signal Processing and Information Technology 2007 IEEE International Symposium on, pp. 353-356.

Akar, O., Gungor, O. (2012). Classification of multispectral images using Random Forest algorithm, Journal of Geodesy and Geoinformation, 1(2), 105-112

Atik, M. E., Duran, Z., & Seker, D. Z. (2021). Machine Learning-Based Supervised Classification of Point Clouds Using Multiscale Geometric Features. ISPRS International Journal of Geo-Information, 10(3), 187.

Breiman, L. (2001). Random forests, Machine learning, 45(1), 5-32

Cortes, C., Vapnik, V. (1995). Support-vector networks, Machine learning, 20(3), 273-297.

Cunningham, P. (2008). Dimension reduction. In Machine learning techniques for multimedia, pp. 91-112. Springer, Berlin, Heidelberg.

Rodarmel, C., Jie S. (2002). Principal component analysis for hyperspectral image classification, Surveying and Land Information Science 62.2, 115-122.

Smith. L. I. (2002). A tutorial on principal components analysis, Technical Report OUCS-2002-12, pp. 1-26.

URL-1 The Indian Pines hyperspectral dataset, Purdue University, Retrieved March 2, 2021, from https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html