ARTICLE IN PRESS

The Egyptian Journal of Remote Sensing and Space Sciences xxx (2017) xxx–xxx

HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

The Egyptian Journal of Remote Sensing and Space Sciences

journal homepage: www.sciencedirect.com

Research Paper

Using a Feature Subset Selection method and Support Vector Machine to address curse of dimensionality and redundancy in Hyperion hyperspectral data classification ☆

Amir Salimi [a,*], Mansour Ziaii [a], Ali Amiri [b], Mahdieh Hosseinjani Zadeh [c], Sadegh Karimpouli [d], Mostafa Moradkhani [b]

[a] Faculty of Mining, Petroleum and Geophysics, Shahrood University of Technology, Shahrood, Iran
[b] Computer Engineering Group, Faculty of Engineering, University of Zanjan, Zanjan, Iran
[c] Department of Ecology, Institute of Science and High Technology and Environmental Science, Graduate University of Advanced Technology, Kerman, Iran
[d] Mining Engineering Group, Faculty of Engineering, University of Zanjan, Zanjan, Iran

A R T I C L E I N F O

Article history:
Received 17 April 2016
Revised 25 January 2017
Accepted 15 February 2017
Available online xxxx

Keywords:
Curse of dimensionality
Feature Subset Selection
Hydrothermal alteration
Hyperspectral
SVM

A B S T R A C T

The curse of dimensionality resulted from insufficient training samples and redundancy is considered as an important problem in the supervised classification of hyperspectral data. This problem can be handled by Feature Subset Selection (FSS) methods and Support Vector Machine (SVM). The FSS methods can manage the redundancy by removing redundant spectral bands. Moreover, kernel based methods, especially SVM have a high ability to classify limited-sample data sets. This paper mainly aims to assess the capability of a FSS method and the SVM in curse of dimensional circumstances and to compare results with the Artificial Neural Network (ANN), when they are used to classify alteration zones of the Hyperion hyperspectral image acquired from the greatest Iranian porphyry copper complex. The results demonstrated that by decreasing training samples, the accuracy of SVM was just decreased 1.8% while the accuracy of ANN was highly reduced i.e. 14.01%. In addition, a hybrid FSS was applied to reduce the dimension of Hyperion. Accordingly, among the 165 useable spectral bands of Hyperion, 18 bands were only selected as the most important and informative bands. Although this dimensionality reduction could not intensively improve the performance of SVM, ANN revealed a significant improvement in the computational time and a slightly enhancement in the average accuracy. Therefore, SVM as a low-sensitive method respect to the size of training data set and feature space can be applied to classify the curse of dimensional problems. Also, the FSS methods can improve the performance of non-kernel based classifiers by eliminating redundant features.

1. Introduction

The remote sensing images are applied to identify and classify the earth surface objects. They provide the valuable information which can be enjoyed by varied fields of sciences. Geologists with the aim of lithological mapping (Amer et al., 2010; Kamel et al., 2016; Karimpouli et al., 2016) and ore minerals exploration also profits this technology. For example, the hydrothermal alteration mineral mapping through remote sensing have been widely and successfully performed for the exploration of various hydrothermal deposits (Amer et al., 2012; Hosseinjani Zadeh et al., 2014a, 2014b; Hosseinjani Zadeh and Tangestani, 2011; Pour and Hashim, 2011; Shahriari et al., 2015, 2013).

The developed hyperspectral sensors compared with the earlier multispectral types use the hundreds of channels to measure spectral data. Therefore, they prepare fine details and great volume of spectral information about materials of the earth surface that enable them to distinguish subtle spectral differences and to identify similar materials of the ground surface (Bioucas-Dias et al., 2013; Camps-Valls et al., 2014; Chang, 2007). Accordingly, the hyperspectral remote sensing has greatly promoted the potential of mineral mapping (Wang and Zheng, 2010) so that within the two last decades it has been an important tool to study minerals

Peer review under responsibility of National Authority for Remote Sensing and Space Sciences.

* Corresponding author.
E-mail address: amir.salimi.1360@gmail.com (A. Salimi).

http://dx.doi.org/10.1016/j.ejrs.2017.02.003
1110-9823/© 2017 National Authority for Remote Sensing and Space Sciences. Production and hosting by Elsevier B.V.

and rocks of the earth surface (Zhang and Peijun, 2014). The hyperspectral data have been utilized to get accurate information about hydrothermal alteration minerals (Bedini et al., 2009; Gersman et al., 2008; Hosseinjani Zadeh et al., 2014a), including an extensive range of the hydroxyl-bearing clay minerals (Van der Meer et al., 2012).

Due to the having of large amount of data, hyperspectral images can be considered as a high dimensional data set. The high dimensional data are composed of a large number of features (or spectral bands in remote sensing data) while many of them are irrelevant or redundant and increase just the size and complexity of the feature space. The high dimensionality of feature space can be led to the curse of dimensionality which is known as an ill-posed problem, because the patterns of curse dimensional feature spaces are not easily recognized (Gheyas and Smith, 2010). The Classification as an important pattern recognition technique is defined as a process which converts data into meaningful information (Mountrakis et al., 2011). The accuracy rate of classification can be negatively affected by the curse of dimensionality which is named the Hughes phenomenon (Alajlan et al., 2012; Camps-Valls et al., 2014; Medjahed et al., 2016; Pal and Foody, 2010; Plaza et al., 2009; Waske et al., 2009). The Hughes problem is usually observed due to insufficient training samples regarding the size of the feature space (Plaza et al., 2009). It means that the average accuracy can be decreased by both the inadequate training samples and the high number of features. This degradation is resulted from an incorrect estimation of statistical parameters of each class because of the low ratio of training samples to the number of features (Waske et al., 2009).

This problem as well as many others challenging problems of hyperspectral data classification have attracted the interest of data mining and machine learning fields (Bioucas-Dias et al., 2013; Camps-Valls et al., 2014). The two solutions to overcome Hughes problem can be: 1) the preparing of sufficient training samples in the training phase of the classification and 2) the decreasing of feature space size by the elimination of redundant features. The two most promising data mining approaches to meet requirements of each situation are, respectively: a) using supervised kernel-based methods which are inherently robust against the high dimensional problems even with a limited training data set (Camps-Valls et al., 2014; Chang, 2007; Mountrakis et al., 2011; Pal and Foody, 2010; Waske et al., 2009) and b) using feature reduction methods (Alajlan et al., 2012; Camps-Valls et al., 2014), because many of the primary spectral bands of hyperspectral data have been demonstrated to be redundant and they are too many to be used for classification (Bioucas-Dias et al., 2013; Landgrebe, 2003).

The kernel-based classifiers solve a linear problem after mapping data from an original input space to a higher dimension feature space. Amongst the most widely used kernel-based methods, Support Vector Machine (SVM) has been extensively applied in the wide variety of applications to analyze the hyperspectral data (Camps-Valls et al., 2014; Plaza et al., 2009). This method has shown the outperformed results with respect to other classification methods. Inherent specifications of the SVM enable it to achieve a high classification accuracy and a good generalization capability even with the small-sized training data sets where a low classification accuracy and a poor generalization are expected (Camps-Valls et al., 2014; Chang, 2007).

The feature reduction methods which are implemented as an important pre-processing step in the high dimensional data analysis, select or produce an efficient subset of the features from the original data set. They should be carried out to handle the curse of dimensionality and to remove irrelevant, redundant and noisy features. The feature reduction methods can increase the speed and generalization of classification, the predictive accuracy and the ability of understanding classification rules (Gheyas and Smith, 2010; Jimenez and Landgrebe, 1997). These methods try to find a small subset of features among the original features which is almost adequate to observe all information behind the data (Fukunaga, 1990). They are grouped in the two main categories: feature extraction and feature selection algorithms. The former alters the original data set to define a new subset of the features containing a vast range of information of the original data set. In the latter, a subset of the original features is identified and selected which covers the most useful information of the highly correlated and redundant features (Pal and Foody, 2010).

At the regional scale of the mineral exploration, the classification of hydrothermal alteration zones is a vital step because of the spatial relationships of deposits and the corresponding alteration zones. Usually, a ground survey is needed to gather the valid data as a training data set of the supervised classification. This work is time consuming and costly, because prospecting areas are not easily accessible and, therefore, the sampling of adequate instances is impossible. Due to this reason as well as the probable redundancy of hyperspectral data, the classification of alteration zones using the hyperspectral data will cause the curse of dimensionality. In the curse of dimensional circumstance, the performance of classification by the conventional classifiers may be compromised, hence the appropriate methods are required to address this problem (Camps-Valls et al., 2014). Therefore, the main objective of this study is the reducing of negative effects of the curse of dimensionality in a hyperspectral classification problem by means of the SVM and a hybrid Feature Subset Selection (FSS) technique. To evaluate the performance of these methods, we will compare the obtained results with the Artificial Neural Network (ANN).

The study area is the most important Cu porphyry type deposit of Iran- Sarcheshmeh- and its surrounding areas, Darrehzar and Sereidun. Aforementioned methods are used to classify the main hydrothermal alteration minerals of the study area using the Hyperion hyperspectral data.

2. Geological setting of the study area

The study area is located in the Iranian Uromiyeh-Dokhtar magmatic belt (Fig. 1a). This belt by diagonal stretching in Iran contains an extensive porphyry copper mineralization including the world class mines such as Sarcheshmeh and Sungun and many other sub economic ore bodies (Boomeri et al., 2010; Waterman and Hamilton, 1975). The largest porphyry copper mine in Iran-Sarcheshmeh- and two other porphyry deposits, Darrehzar and Sereidun, are the study areas of this paper (Fig. 1b) which are located in the southern part of this magmatic belt.

The huge Sarcheshmeh deposit with the 1200 Mt, Cu (with average grade of 0.69%) and Mo (with average grade of 0.03%) occurred in a diorite to granodiorite stock (Aftabi and Atapour, 2010; Waterman and Hamilton, 1975). This deposit is known as a typical porphyry Cu deposit regarding its alteration types, mineralization style, size, ore grade and tectonic setting (Boomeri et al., 2010; Waterman and Hamilton, 1975). Its various hydrothermal alterations from center to outward are the same with the alterations of typical porphyry Cu deposits containing potassic, biotitic, phyllic, argillic, and propylitic, respectively (Aftabi and Atapour, 2010; Waterman and Hamilton, 1975). Spatially, the arrangement of alterations from center to outside of the deposit is potassic, potassic affected by phyllic, strongly phyllic, and propylitic alterations (Shafiei and Shahabpour, 2012).

The Sereidun district, in the east vicinity of the Sarcheshmeh, contains the various alterations including propylitic, phyllic, argillic, advanced argillic and silicic. Spatially, the phyllic alteration