Chemometrics and Intelligent Laboratory Systems 169 (2017) 35–44

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Stability analysis of hyperspectral band selection algorithms based on neighborhood rough set theory for classification

Yao Liu [a], Junjie Yang [a,*], Yuehua Chen [b], Kezhu Tan [b], Liguo Wang [c], Xiaozhen Yan [d]

[a] School of Information Engineering, Lingnan Normal University, Zhanjiang 524048, China
[b] College of Electrical and Information, Northeast Agricultural University, Harbin 150030, China
[c] College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China
[d] School of Information and Electrical Engineering, Harbin Institute of Technology at WeiHai, WeiHai 264200, China

ARTICLE INFO

Keywords:
Hyperspectral image
Band selection
Stability
Neighborhood rough set

ABSTRACT

Band selection is a well-known approach for reducing the dimensionality of hyperspectral data. When the neighborhood rough set theory is used to select informative bands, different criteria of the band selection algorithms may lead to different optimal band subsets. Many studies have been analyzed the classification performance of band selection algorithms and have demonstrated that different algorithms are similar for classification. Therefore, rather than evaluating band selection algorithms using only classification accuracy, their stability should also be explored. The stability of an algorithm, which is quantified by the sensitivity of the algorithm to variations in the training set, is a topic of recent interest. Most studies on stability compare the band subsets chosen either from perturbation datasets by randomly removing methods or from perturbation datasets by cross validation methods. These methods either result in an unknown degree of overlap between perturbation datasets, or an invariable degree of overlap. In this work, we propose an adjustable degree of overlap method to construct perturbation datasets, which can set different levels of perturbation. By introducing the Jaccard index as a metric of stability, we explore the stability of six band selection algorithms based on the neighborhood rough set theory. We experimentally demonstrate that the level of perturbation, the degree of overlap, the size of the subset, and the size of the neighborhood affect stability. The results show that the maximal relevance minimal redundancy difference band selection algorithm has the greatest stability overall and better classification ability.

1. Introduction

Hyperspectral imaging has become an active research topic because it is widely used in many fields, such as remote sensing applications, agriculture and food quality evaluation, mineral exploration and environmental monitoring [1–3]. A hyperspectral image contains hundreds of bands with fine spectral resolution and possesses much richer spectral information than a multispectral image. Each pixel in a hyperspectral image contains a spectrum representing the light absorbing and scattering properties of the spatial region represented by that pixel. Hyperspectral imaging technology enables the visualization of the chemical composition of the sample. It is fast, non-destructive and does not require chemicals. However, the current hyperspectral imaging system is a laboratory-based system, which is not yet ready for implementation in process monitoring and real-time inspection due to its high dimensionality as well as time constraints for image acquisition and subsequent image analysis. Band selection is an effective method to reduce dimensionality, as it creates an optimal subset of bands by removing redundant bands and leaving only the most important bands [4]. Because the optimal subset is much smaller than the entire hyperspectral dataset, the computation time of building a model is greatly reduced. Applying band selection benefits the performance of the model and enhances the interpretability of the model [5]. On the basis of the selected optimal spectral bands, the high dimensional hyperspectral images can be reduced to form multispectral images, and then a rapid, stable and more accurate multispectral imaging system can be established for routine chemical analysis and ultimately can be implemented directly in industrial applications.

A variety of band selection algorithms have been developed and proven to be effective at improving accuracy for classification in many application domains [6]. Genetic algorithm (GA) was used as a band selection algorithm to help alleviate the problem of high dimensional

* Corresponding author.
E-mail address: jjyang2013@163.com (J. Yang).

http://dx.doi.org/10.1016/j.chemolab.2017.08.005
Received 9 May 2017; Received in revised form 11 August 2017; Accepted 16 August 2017
Available online 19 August 2017

complexity. It selected bands that relate to principal physio-chemical properties of plants and maintained the separation of species classes at a high level [7]. The successive projections algorithm (SPA) combined with uninformative variable elimination (UVE) were proposed as a band selection approach for multivariate calibration. The proposed method was applied to analyze nicotine and sugar contents in tobacco samples and was proven to be an efficient tool. Slightly better results were obtained compared with a full-spectral UVE method [8]. C. Ferrari et al. used interval partial least squares-discriminant analysis (iPLS-DA) as a band selection algorithm to discriminate bruised apples from sound apples. The results confirmed that a band selection algorithm is an effective tool to deal with hyperspectral datasets and can efficiently detect the presence of bruises [9]. Y. Dong et al. employed information gain to solve the problem of computational complexity and to regulate the size of the discernibility matrix. The proposed band selection algorithm was illustrated by an analysis of classification models for phenylalanine in plasma [10].

The rough set (RS) theory [11] proposed by Pawlak is a new paradigm to deal with uncertain, vague, and incomplete information. It has been applied to pattern recognition, rule extraction, and especially to attribute reduction [12] (also called feature selection). The essence of the RS theory of feature selection is to find a subset of features that keeps the discernibility of the original dataset and has no redundant features. The standard RS model could not be directly used for hyperspectral datasets because they are always real-valued. A better choice to solve the problem is applying the neighborhood rough set (NRS) theory [13], which is suitable for handling both numerical and categorical datasets.

In the NRS theory, for a classification task, the key step of the feature selection is to design effective criteria that measure the discriminating ability of a feature or a feature subset to distinguish different classes. Some criteria of feature selection algorithms based on the NRS theory have been proposed, such as distance criteria, dependency measure, consistency measure, and information measure [14,15]. Different criteria may lead to different optimal feature subsets. For a given dataset, there are usually multiple feature subsets, and they all can be employed for constructing multiple classifiers. It is necessary to use metrics to evaluate the feature selection algorithms. One metric which is widely used to evaluate an algorithm is classification accuracy. This metric demonstrates the ability of the selected features to distinguish the class of the data [16]. From previous studies, we find that there are several different feature subsets (either from the same feature selection algorithm or from different feature selection algorithms), that perform equally well in terms of classification accuracy [17]. This observation motivates people to seek other assessment metrics to evaluate the feature selection algorithms. The stability assessment of feature selection results satisfies this need and has drawn increasing attention. It is defined as the sensitivity of a feature selection algorithm to small perturbations in the training dataset [18]. The algorithms, which produce similar results even with changing data, are said to be stable, and they are preferred to those algorithms that produce inconsistent outputs [19]. The models built by stable feature subsets will produce more trustworthy results because any feature chosen by a stable selection algorithm is not chosen by luck. Few studies exist on the stability of feature selection algorithms. To the best of our knowledge, our work is the first to study the stability of band selection algorithms based on the NRS theory for hyperspectral datasets. It is important to note that improving the stability of feature selection results without considering classification accuracy is not expected, because domain experts are not interested in an algorithm that yields very stable feature subsets but leads to bad classification performance [20]. Thus, these two aspects always have to be investigated together.

In this work, we present a comprehensive comparison of the stability of six different band selection algorithms that apply different evaluation criteria of the NRS theory. We explore the variation of stability across band selection algorithms or within each algorithm under different experimental conditions. For an in-depth understanding of the stability issue, several factors are discussed that may affect stability. We empirically demonstrate the influence on stability of the level of perturbation, the degree of overlap between perturbation datasets, the size of the subset, and the size of the neighborhood. Note that the purpose of this study is to find the specific band selection algorithm with superior performance than others both in terms of stability and classification ability, and also to provide valuable insight for designing proper band selection algorithms. The remainder of this paper is organized as follows. Section 2 reviews the band selection algorithms based on the NRS theory and presents the evaluation methods of stability. In Section 3, the acquisition of a hyperspectral dataset and experimental design are described. Section 4 presents the experimental results and discussions of stability. We then conclude the paper in Section 5.

2. Preliminaries

2.1. Band selection algorithms based on the NRS theory

Let $IS = (U, C \cup D, V, f)$ be an information system for classification learning, where $U = \{x_1, x_2, \cdots, x_n\}$ is a nonempty and finite sample set called the universe, C is called condition features to characterize the samples, D is a set of decision label of the samples called decision features with $C \cap D = \phi$. Neighborhood relations have been proposed as an extended RS model in the information system. The neighborhood of the sample $x_i \in U$ in the subspace $B \subseteq C$ is defined as $\delta_B(x_i) = \{x_j | \Delta_B(x_i, x_j) \leq \delta\}$, where δ is the threshold.

On the hyperspectral datasets, assume that here are c classes of samples, let $D = \{d_1, d_2, \cdots, d_s\}$ denote the class labels of s samples, where $d_i = k, (i = 1, 2, \cdots, s)$ indicates the sample i being class k, where $k = 1, 2, \cdots, c$. Let $S = \{s_1, s_2, \cdots, s_n\}$ be a set of samples and $W = \{w_1, w_2, \cdots, w_m\}$ a set of bands. The corresponding hyperspectra - band expression matrix can be represented as $C = \{c_{ij} | i = 1, 2, \cdots, n; j = 1, 2, \cdots, m\}$, where c_{ij} is the expression of band j in sample i.

In the literature, band selection algorithms are widely used prior to the classification of hyperspectral datasets [21]. The task of band selection is to find an optimal band subset from all bands, which is most dependent on the target class and provides good classification results. The optimal band subset is always relative to a certain band evaluation criterion. A good criterion can select a subset that is unchanged with any other possible variation and can emphasize differences in discriminating classes [22]. Three evaluation criteria of the NRS theory, that is, dependency measure, consistency measure and information measure, are considered to evaluate the effectiveness of a band subset. Based on the three evaluation criteria, six band selection algorithms, namely dependency measure of the NRS theory (DMNRS), variable precision of the NRS theory (VPNRS) [23], consistency measure of the NRS theory (CMNRS) [21], neighborhood mutual information (NMI) [24], maximal relevance minimal redundancy difference (MRMRD) and maximal relevance minimal redundancy quotient (MRMRQ) [25], are used to investigate the stability of band selection algorithms.

The DMNRS algorithm evaluates the significance of a band by neighborhood dependency, which is introduced as follows.

Definition 1. *Given a neighborhood decision system (NDS) $(U, C \cup D, N)$, $B \subseteq C$, the neighborhood dependency of B with respect to D is defined as*

$$\gamma_B(D) = \frac{\left|\underline{N_B} D\right|}{|U|}, 0 \leq \gamma_B(D) \leq 1, \tag{1}$$

where $|\cdot|$ denotes the cardinality of a set, and $\underline{N_B} D$ denotes the lower approximations of D with respect to band subset B.

The change in dependency when a band is added to the subset, is a measure of the significance of the band. The higher the change in dependency, the more significant the band is.

Definition 2. *Given an NDS $(U, C \cup D, N)$, $B \subset C$, $a \in C - B$, the significance*