Remote Sensing of Environment 158 (2015) 478–491

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Improved time series land cover classification by missing-observation-adaptive nonlinear dimensionality reduction

L. Yan *, D.P. Roy

Geospatial Science Center of Excellence, South Dakota State University Brookings, SD 57007, USA

A R T I C L E   I N F O

Article history:
Received 11 June 2014
Received in revised form 12 November 2014
Accepted 15 November 2014
Available online 12 December 2014

Keywords:
Landsat
Land cover classification
Time series
Nonlinear dimensionality reduction
WELD

A B S T R A C T

Dimensionality reduction (DR) is a widely used technique to address the curse of dimensionality when high-dimensional remotely sensed data, such as multi-temporal or hyperspectral imagery, are analyzed. Nonlinear DR algorithms, also referred to as manifold learning algorithms, have been successfully applied to hyperspectral data and provide improved performance compared with linear DR algorithms. However, DR algorithms cannot handle missing data that are common in multi-temporal imagery. In this paper, the Laplacian Eigenmaps (LE) nonlinear DR algorithm was refined for application to multi-temporal satellite data with large proportions of missing data. Refined LE algorithms were applied to 52-week Landsat time series for three study areas in Texas, Kansas and South Dakota that have different amounts of missing data and land cover complexity. A series of random forest classifications were conducted on the refined LE DR bands using varying proportions of training data provided by the United States Department of Agriculture (USDA) National Agricultural Statistics Service (NASS) Cropland Data Layer (CDL); these classification results were compared with conventional metrics-based random forest classifications. Experimental results show that compared with the metrics approach, higher per-class and overall classification accuracies were obtained using the refined LE DR bands of multispectral reflectance time series, and the number of training samples required to achieve a given degree of classification accuracy was also reduced. The approach of applying the refined LE to multispectral reflectance time series is promising in that it is automated and provides dimensionality-reduced data with desirable classification properties. The implications of this research and possibilities for future algorithm development and application are discussed.

1. Introduction

There is an established heritage for the application of dimensionality reduction (DR) techniques to multispectral satellite time series prior to land cover classification. DR techniques transform high-dimensional data into data with lower dimensions that ideally maximize the information content and minimize noise (Bellman, 2003; Hughes, 1968). For example, the principal component analysis (PCA) DR technique has been applied to satellite time series to provide new dimensionality-reduced bands for land cover and land cover change classification (Collins & Woodcock, 1996; Murthy, Raju, & Badrinath, 2003; Small, 2012; Townshend, Justice, & Kalb, 1987; Zhong & Wang, 2006). PCA performs a linear mapping of the data to a lower dimensional space so that the variance of the transformed data is maximized. A number of other DR techniques based on linear transformations have been developed including linear discriminant analysis (LDA) (Fisher, 1936; Martinez & Kak, 2001), projection pursuit (PP) (Friedman & Tukey, 1974; Jimenez & Landgrebe, 1999), minimum noise fraction (MNF) (Green, Berman, Switzer, & Craig, 1988), independent component analysis (ICA) (Hyvärinen, 1999; Hyvärinen & Oja, 2000), and spatial–spectral eigenvector derivation (SSEVD) (Rogge, Bachmanna, Rivard, Nielsen, & Feng, 2014). However, linear DR techniques do not accommodate the intrinsic nonlinear characteristics of optical wavelength remotely sensed data, whereby the remotely sensed contribution of the observed scene components are not linearly proportional to their surface areas, and that are particularly apparent for vegetation. Nonlinearity is introduced by multiple scattering between different scene components that varies as a function of the wavelength, the viewing and illumination geometry, and the three dimensional structure of the scene components (Schaaf et al., 2002; Somers et al., 2009). Spectral indices based on ratios of reflectance, such as the normalized difference vegetation index (NDVI), are not linear functions of reflectance and so also have nonlinear characteristics (Verstraete & Pinty, 1996), and, for example, their use in spectral linear unmixing is not recommended (Busetto, Meroni, & Colombo, 2008; Settle & Campbell, 1998). The reflectance contribution of most scene components change temporally and this may also introduce nonlinearity when time series data are considered. Nonlinear DR techniques have been used by the remote sensing community

* Corresponding author.
E-mail addresses: lin.yan@sdstate.edu (L. Yan), david.roy@sdstate.edu (D.P. Roy).

http://dx.doi.org/10.1016/j.rse.2014.11.024

predominantly for application to single date hyperspectral data, and have been found to provide improved performance for classification, target discrimination and end-member extraction, when compared with linear DR techniques (Bachmann, Ainsworth, & Fusina, 2005; Feilhauer, Faude, & Schmidtlein, 2011; Han & Goodenough, 2005; Zhang, Zhang, Tao, & Huang, 2013). A few studies have applied nonlinear DR techniques to single date multispectral satellite data (Journaux, Foucherot, & Gouton, 2006). However, nonlinear DR techniques have not been applied to multispectral satellite time series.

An established limitation of linear and nonlinear DR techniques is that they cannot handle missing data (Hubert, Rousseeuw, & Verboven, 2002). This is particularly problematic for satellite time series application, as missing data occurs frequently due to cloud obscuration (Ju & Roy, 2008) and also sometimes due to issues associated with the satellite acquisition, satellite to ground station data transmission, and data production errors (Roy, Lewis, Schaaf, Devadiga, & Boschetti, 2006; Roy et al., 2002). The impacts of missing data may be minimized in coarse spatial resolution time series by processing them into reduced temporal resolution time series. This has been achieved either by application of per-pixel temporal compositing procedures that select a best pixel observation every *n*-days (Holben, 1986) or by inversion of *n*-days of observations against a model of the surface bidirectional reflectance distribution function (BRDF) to estimate reflectance at consistent viewing and illumination angles (Schaaf et al., 2002). These techniques are less appropriate, however, for application to time series with lower temporal resolution, such as Landsat. Per-pixel temporal compositing procedures are difficult to implement reliably with Landsat data because of the low cloud-free observation frequency relative to surface changes (Roy et al., 2010; White et al., 2014) and Landsat BRDF inversion approaches do not work reliably because of the narrow Landsat field of view which precludes sampling of the intrinsic reflectance anisotropy of most land surfaces (Roy, Wulder, et al., 2014; Shuai, Masek, Gao, Schaaf, & He, 2014).

The current state of the practice for large area multi-temporal land cover classification is to derive metrics from the time series and then classify the metrics bands with a supervised (i.e., training data dependent) non-parametric classification approach. The choice of metrics, usually the maximum and quartile values of spectral indices and spectral bands over the time series, has been justified empirically in terms of attempting to capture seasonal land cover class spectral variations in a way that is robust to missing data (Broich et al., 2011; DeFries, Hansen, & Townshend, 1995; Friedl et al., 2010; Hansen et al., 2011, 2014). The nonlinearity in the metrics is assumed implicitly to be handled by the statistical classifier, which may be reasonable when non-parametric classifiers are used. However, while providing a practical and useful form of data reduction, the metrics-based classification approaches may not be optimal as compared with linear and nonlinear DR techniques, because they may use less of the temporal information available in the satellite time series.

In this paper, a recent nonlinear DR method developed for application to hyperspectral data is refined for application to multispectral satellite time series that have missing observations. The methodology can be applied to any satellite time series but is demonstrated in this study using Landsat data. Landsat data have an established and rapidly evolving heritage for multi-temporal classification of land cover, change, and disturbance (Hansen & Loveland, 2012; Roy, Wulder, et al., 2014), but often have missing data due to cloud obscuration (Kovalskyy & Roy, 2013) and, in historical data, due to satellite reception and acquisition issues (Loveland & Dwyer, 2012). The Landsat data and study areas are first described, followed by a description and justification for selection of the Laplacian Eigenmaps (LE) nonlinear DR method (Belkin & Niyogi, 2002) and how it was refined to handle missing data. Experiments are conducted in three study areas in Texas, Kansas, and South Dakota that encompass complex agricultural landscapes with differing amounts of missing Landsat data. Supervised random forest classification of the output of the refined DR method applied to 52 weeks of Landsat time series are compared with supervised random forest classification of the conventional metrics derived from the Landsat data. The classification accuracies of the LE and the metrics-based approaches generated using varying proportions of training data, defined by sampling United States Department of Agriculture (USDA) Cropland Data Layer (CDL) data, are examined, and implications and recommendations for future research are discussed.

2. Data

2.1. Landsat data

Web Enabled Landsat Data (WELD) Version 1.5 products were obtained from the USGS National Center for Earth Resources Observation and Science (EROS) (http://e4ftl01.cr.usgs.gov/WELD/). The products are defined in the Albers Equal Area conic projection in separate geolocated tiles of 5000 × 5000 30 m pixels and because they are temporally aligned, they are straightforward to use for time series classification applications. For example, they were used to generate 30 m conterminous United States (CONUS) annual land cover (Hansen et al., 2011) and 5-year land cover change (Hansen et al., 2014) classifications. Weekly, monthly, seasonal and annual WELD products are generated by application of a temporal compositing scheme to select a single best pixel observation for each reporting period (Roy et al., 2010). The annual product is generated by compositing 52 weeks of Landsat data, and the seasonal, monthly and weekly products are generated by compositing the Landsat data acquired in each season, month, or week, respectively. The version 1.5 WELD products were generated from every available Landsat 7 ETM + Level 1 T processed image in the USGS Landsat archive with cloud cover ≤80%. Standard Level 1 T processing includes radiometric correction, systematic geometric correction, precision correction using ground control chips, and the use of a digital elevation model to correct parallax error due to local topographic relief, with a CONUS geolocation error less than 30 m (Lee, Storey, Choate, & Hayes, 2004).

In this study WELD products for climate year 2010 were used. The annual (December 2009 to November 2010), Summer (June to August 2010), Autumn (September to November 2010), seven monthly (April through October 2010), and 52 weekly products (1st December 2009 to November 30th 2010) were used. Their specific usage is described in Section 5.

All the WELD products store for each 30 m pixel location the six reflective top of atmosphere Landsat 7 Enhanced Thematic Mapper Plus (ETM +) bands, the two top of atmosphere thermal bands, bit packed band saturation information, Normalized Difference Vegetation Index (derived as the near-infrared minus the red reflectance divided by their sum), two cloud masks, the day of the year that the pixel value was sensed on, and the number of Landsat observations considered in the product period (week, month, season, or annual) (Roy et al., 2010). In this study, the Landsat ETM + reflective wavelength bands 2 (green, 0.53–0.61 μm), 3 (red: 0.63–0.69 μm), 4 (near-infrared: 0.78–0.90 μm), 5 (middle-infrared: 1.55–1.75 μm), 7 (middle-infrared: 2.09–2.35 μm) and the NDVI were used. The shortest wavelength Landsat ETM + band 1 (blue: 0.45–0.52 μm) was not used because it is overly sensitive to atmospheric scattering (Ju, Roy, Vermote, Masek, & Kovalskyy, 2012; Roy, Qin, et al., 2014). All WELD pixel values flagged as cloudy in both the cloud masks were removed as they were highly likely to be cloud contaminated (Roy et al., 2010).

2.2. Cropland Data Layer

The United States Department of Agriculture (USDA) National Agricultural Statistics Service (NASS) Cropland Data Layer (CDL) for 2010 was obtained from the CDL web site (http://nassgeodata.gmu.edu/CropScape/). The CDL data were used as a source of supervised classification training and test data. The CDL is generated annually using moderate resolution satellite imagery and extensive agricultural ground