Optimizing unsupervised classifications of remotely sensed imagery with a data-assisted labeling approach

Ruili Lang, Guofan Shao*, Bryan C. Pijanowski, Richard L. Farnsworth

Department of Forestry and Natural Resources, 715 West State Street, West Lafayette, IN 47907-2061, Purdue University, USA

Received 27 May 2007; received in revised form 25 October 2007; accepted 25 October 2007

Abstract

The quality of remotely sensed land use and land cover (LULC) maps is affected by the accuracy of image data classifications. Various efforts have been made in advancing supervised or unsupervised classification methods to increase the repeatability and accuracy of LULC mapping. This study incorporates a data-assisted labeling approach (DALA) into the unsupervised classification of remotely sensed imagery. The DALA-unsupervised classification algorithm consists of three steps: (1) creation of \( N \) spectral-class maps using Iterative Self-Organizing Data Analysis Technique Algorithm (ISODATA); (2) development of LULC maps with assistance of reference data; and (3) accuracy assessments of all the LULC maps using independent reference data and selection of one LULC map with the highest accuracy. Classification experiments with a composite image of a Landsat Thematic Mapper (TM) image and an Enhanced Thematic Mapper Plus (ETM+) image suggest that DALA was effective in making unsupervised classification process more objective, automatic, and accurate. A comparison between the DALA-unsupervised classifications and some conventional classifications suggests that the DALA-unsupervised classification algorithm yielded better classification accuracies compared to these conventional approaches. Such a simple, effective approach has not been systematically examined before but has great potential for many applications in the geosciences.

\( \text{c} \) 2008 Elsevier Ltd. All rights reserved.

Keywords: LULC mapping; Automated classification; Landsat; Classification accuracy; Classification algorithm

1. Introduction

The launch of Landsat-1 in 1972 initiated a new era of providing satellite data in digital format to users. Efforts to develop algorithms for image classification have continued since the 1970s. Combined with the rapid evolution of Geographic Information Systems, researchers have quickly created computer-aided analysis tools to produce land use and land cover (LULC) maps (Bauer et al., 1994). These maps have been used to analyze the impacts of land use change on the environment (Cihlar, 2000), improve land use planning and natural resource management (Treitz and Rogan, 2004), and better understand ecological processes on Earth (Fassnacht et al., 2006).

The quality of LULC maps is determined mainly by the accuracy of an image classification approach selected by the analyst (Congalton, 1991; Congalton and Green, 1999). Soon after the availability of Landsat data, the United States Geological Survey

---

*Corresponding author.

E-mail address: shao@purdue.edu (G. Shao).
comparatively well separated (Lillesand et al., 2004). However, real-world LULC mapping rarely attains this level of accuracy. For example, overall accuracy of LULC mapping with Landsat Thematic Mapper (TM) data for the eastern United States was 81% for Anderson level I (i.e. water, urban, barren land, forest, agricultural land, wetland, and rangeland) and 60% for Anderson level II (Vogelmann et al., 2001). Given our inability to consistently achieve the minimum requirement of 85%, should we lower the minimum-accuracy requirement? For the purpose of landscape assessment, a lower-than-85% accuracy seems unacceptable (Shao et al., 2001; Lunetta and Lyon, 2004). Groom et al. (2006) suggest re-examining the classificatory and informational implications of image data used in landscape ecology.

Of the many classification approaches available, most researchers use either supervised or unsupervised classification. Supervised classification consists of two stages: training and classification. In the training stage, the analyst identifies representative training areas and develops a numerical description of the spectral attributes of each land cover type of interest in the scene. In the classification stage, each pixel in the image data set is categorized into the land cover class it most closely resembles (Lillesand et al., 2004). The accuracy of supervised classification largely depends on the quality of the training data (McCaffrey and Franklin, 1993). In general, training data are collected at homogenous locations (Jensen, 2005) and represent sub-types of information classes (i.e. LULC types). The locations and sample size of training data are difficult to be optimized depending on image data types and classifiers to be used.

Unsupervised classification techniques do not utilize training data as the basis for classification. The basic premise of unsupervised classification is that spectral values within a given land cover type should be close together in the feature space, whereas data values in different classes should be comparatively well separated (Lillesand et al., 2004). Various algorithms exist to create spectral classes. One of the most commonly used algorithms is the Iterative Self-Organizing Data Analysis Technique Algorithm (ISODATA), which uses the Euclidean distance in the feature space to assign every pixel to a cluster through a number of iterations (Jensen, 2005; ERDAS, 1999). Spectral classes identified by unsupervised classifications are the natural, inherent groupings of spectral values within a scene of remote sensing data (Campbell, 2007). An unsupervised classification consists of clustering and labeling spectral classes. When ISODATA is performed, it is common for an analyst to select the number of spectral classes (NSC), a convergence threshold (CT), and number of iterations for the algorithm, which introduces considerable subjectivity into the classification process.

In the labeling stage, it is the analyst who assigns spectral classes to information classes. The labeling of spectral classes can be time consuming, labor intensive, subjective, and even error prone (Lillesand et al., 2004; Aronoff, 2005). This is mainly because there is often no one-to-one correspondence between spectral and information classes (Lillesand et al., 2004; Campbell, 2007). Sometimes analysts have no confidence in labeling some spectral classes into information classes (Jensen, 2005). Subjectivity is thus inevitable in the second step of unsupervised classification. This is particularly true when there is high NSC, viz., 50 or more. Therefore, it is impractical to compare many possible classification results out of different parameter combinations in the clustering process. As a result, the potential of unsupervised classification cannot be fully realized.

Serious limitations exist for both supervised and unsupervised classification approaches. As a result of these limitations, accuracy suffers as does the ability of researchers to repeat the analysis and achieve similar results. For example, in an analysis of two LULC products from the Data and Information System of International Geosphere–Biosphere Programme and University of Maryland, both of which were derived from the same 1992–1993 Advanced Very High Resolution Radiometer data set, Hansen and Reed (2000) found a per-pixel agreement of only 74%. Such disparities may be due to the subjectivity that exists in both approaches (Cardille et al., 1996; Wayman et al., 2001). Wayman et al. (2001) proposed a hybrid classification technique called Iterative Guided Spectral Class Rejection (IGSCR), which uses a guided “cluster busting” approach (Jensen, 2005) to create a single signature file through multiple iterations with unsupervised classifications and then classify the whole image data with a supervised classification method. Like other hybrid classification methods, IGSCR represents an important improvement in image data classification. However,
little attention has been paid to systematic studies about maximizing classification accuracy with different parameter combinations associated with unsupervised classifications. There are still subjectivities in the existing classification techniques.

This paper introduces an automated, data-driven labeling approach which can turn unsupervised classification from a relatively experience-dependent, subjective approach to an experience-independent, objective approach. We show how the repeatability, accuracy, and efficiency of image data classification can be improved by using this method compared to conventional approaches.

2. The design of an automated labeling approach

An ideal unsupervised classification algorithm is an automatic and objective process that generates high-accuracy maps. The data-assisted labeling approach (DALA) proposed in this paper is a significant improvement toward addressing the limitations of both classification methods. It involves three general steps (Fig. 1):

- Creation of \( N \) spectral-class maps using ISO- DATA with multiple combinations of reasonable NSC and CT values.
- Development of \( N \) LULC maps from the spectral-class maps with assistance of Reference Data I.
- Accuracy assessments of each of the \( N \) LULC maps using Reference Data II and selection of the most accurate LULC map.

The first step is the ordinary clustering process in unsupervised classification. \( N \) depends on how many NSC and CT options are used. For manual labeling processes following clustering, a lower NSC normally means lower classification. When NSC is high, a labeling process will be labor intensive and time consuming. A high NSC also creates spectral classes between typical classes and subjective errors can be easily made for labeling the untypical classes in practice. This is a limiting factor for the unsupervised classification. However, a high NSC is not a limitation for DALA.

The second step is a labeling process assisted with reference data, called Reference Data I. Such reference data are collected at an information-class level rather than a sub-type level that is normally applied to training data in supervised classification. By overlaying Reference Data I with a spectral-class map, a table with intersected information classes and spectral classes is obtained (such as Table 1). An analyst can systematically and automatically assign spectral classes to information classes by certain rules, such as a majority rule, under which a spectral class belongs to the information class that has the highest pixel number of Reference Data I corresponding to the same spectral class. Using a Landsat TM image as a case study (see below), the first spectral group contains six reference pixels for forest (Table 1), 189 pixels for agriculture, three pixels for urban, and nine pixels for water. In this case, the agriculture class is assigned to the first spectral class. Similar operations are carried out on the remaining classes. Such an automated-labeling procedure results in a LULC map from each of \( N \) spectral-class maps.

Accuracy assessment, step 3, uses Reference Data II to estimate the accuracy of each of the LULC maps created in step 2. Reference Data II must be independent of Reference Data I. The standard matrix-table approach can be used to assess each LULC map (Congalton and Green, 1999). The best map can be selected by referring to one or multiple accuracy indices, such as overall accuracy and Kappa statistic.
To automatically implement DALA, we wrote a computer program in Microsoft Visual Basic (VB) to implement Step 2 above. Macro languages can be used to link the three steps within commonly used remote sensing software such as ERDAS IMAGINE (www.leica.com).

3. Comparative classification experiments

A scene of Landsat TM data of path 22 and row 32, acquired on September 4, 1998, and a scene of Enhanced Thematic Mapper Plus (ETM+) data of the same location, covering northern Indiana, acquired on June 16, 2001, were stacked and fused into a composite multi-spectral image containing spectral features from two separate growing seasons. The resulting image contained 15 m pixels and 13 bands (excluding the thermal infrared band). To reduce processing time, we selected four 8 by 8 km² sample sites within the north-central Indiana composite image where LULC data were available for reference data (Fig. 2). The goal of the image data classifications was to discriminate four USGS-Level-I LULC types (Anderson et al., 1976): urban, agriculture, forest, and water.

The LULC data for the four 8 by 8 km² sample sites, delineated from 1 m-resolution ortho-photographs acquired in 1998 by the USGS, was used as LULC reference data. After vector-to-raster conversion, two sets of sample points (15 m pixels) were randomly selected from the LULC reference data and labeled Reference Data I and Reference Data II. The former was used for labeling in Step 2; the latter, for accuracy assessments in Step 3. Each set of reference data contained almost 8000 random pixels. There was a time difference of nearly 3 years between the reference aerial photographs and 2001 Landsat data. Changes in LULC during the 3 years were visually detected. To avoid the effects of LULC change on image data classifications, we removed the changed areas from both reference data and Landsat data.

Image data classifications were performed with ERDAS IMAGINE 8.7 and the VB program. Consistent with our three-step DALA-unsupervised classification algorithm, 24 LULC maps were created and compared (see below). The 24 LULC maps represent combinations of eight NSC geometric-progression values (4, 8, 16, 32, 64, 128, 256, and 512) and three CT values (95%, 97%, and 99%) with the number of iterations set at 100 in ISODATA to ensure convergence. Each classification was assessed using the error-matrix approach (Congalton and Green, 1999). Because the balance between producer’s accuracy (PA) and user’s accuracy (UA) is important to estimate the area for each LULC type (Shao et al., 2003), an index named $D_{\text{max}}$, defined as maximum absolute difference between PA and UA ($D_{\text{max}} = |PA - UA|$) for all the land use types within a matrix from each LULC map, was also used to evaluate each classification. The balance between PA and UA becomes better as $D_{\text{max}}$ decreases. The Z test (Congalton and Green, 1999) was used to quantitatively compare the LULC maps.

The same data set was used in an independent classification project in a remote sensing course, where nine graduate students classified the image data with conventional supervised and unsupervised classifications. To encourage the students to strive for the best classification rate, their grade was partially determined by the classification accuracy. To ensure all the maps share the same standard of classification accuracy, all the students used the Reference Data II to assess their classifications. Each student classified the image data into LULC maps using ISODATA unsupervised, minimum-distance (MD) supervised, and maximum-likelihood
ML) supervised classifications. Thus, the nine student analysts generated a total of 27 LULC maps. The classifications from the class project were used to evaluate the effectiveness of DALA.

4. Classification results

The overall accuracies of the 24 thematic maps generated with the DALA classification experiments ranged from 67.1% to 86.4% for NSC = 4–512 and CT = 95–99%. The map with 86.4% classification accuracy was obtained with NSC = 512 and CT = 99%. Apparently, NSC and CT had different effects on classification accuracy (Fig. 3a). In other words, classification accuracy was more affected with NSC (Table 2) than with CT (Table 3). However, when CT was higher, the increase in NSC became more effective on increasing classification accuracy (Table 2). For example, classification accuracy stopped increasing when NSC reached 64 for CT = 95% or 97% but continued to rise when NSC was increased from 64 to 128 for CT = 99%. In all the cases, continued increase in NSC beyond 128 was not significantly effective for improving classification accuracy. CT seemed to have more effect on classification accuracy only when NSC was relatively low (Table 3). When NSC was 16 or higher, CT was no longer effective for improving classification accuracy.

The $D_{\text{max}}$ values of the 24 maps varied tremendously across the combinations of NSC and CT values (Fig. 3b). When NSC reached 128, $D_{\text{max}}$ values became consistently lower. Based on the relationships of NSC with both overall classification accuracy and $D_{\text{max}}$ values (Fig. 3), the processes of improving the LULC maps could be divided into three stages: low-accuracy stage when NSC = 4; unstable-accuracy stage when NSC = 8-64; and high-accuracy stage when NSC = 128–512. Maps created at the low-accuracy stage were impractical and unacceptable because analysts rarely used such a low NSC even in conventional classification methods. Maps generated at the unstable-accuracy stage were commonly made with conventional classification methods but had high variations in classification accuracy. Classification accuracy temporarily stopped increasing and even became lower when NSC was changed from 8 to 16 (Fig. 3a, Table 2), a common phenomenon observed with conventional unsupervised classifications. Maps created at the high-accuracy stage could hardly be obtained with conventional classification methods because NSC was too high for manual labeling processes. Based on the $Z$-test, most of the maps at the unstable-accuracy stage were different in classification accuracy while the maps at the high-accuracy stage were virtually the same in classification accuracy (Table 2). The high-accuracy stage also means a stable or equilibrium stage.
The overall accuracy of the 27 LULC maps generated by the student analysts ranged from 72.2% to 85.1%. Average overall accuracy equaled 80.6 ± 5.24% (between 72.2 and 84.7%) for maps created with unsupervised classification, 82.6 ± 2.33% (between 78.8 and 84.5%) with ML supervised classification, and 82.6 ± 2.80% (between 77.1 and 85.1%) with MD supervised classification. The $Z$ value between the best LULC map from the student analysts and the map created with NSC = 512 and CT = 99% from the DALA experiment equaled 3.26. At the 99% confidence level, the difference in classification accuracy between the two maps was significant, indicating that the classification accuracy with DALA-unsupervised classification algorithm could be greater than any maps.
5. Discussion and conclusions

The DALA-unsupervised classification algorithm proposed and evaluated in this paper addressed many of the critical obstacles that confront analysts who use digital remotely sensed data to develop LULC maps. First, subjectivity introduced by the analyst in his/her selection of one pair of NSC and CT values is minimized by using multiple NSC and CT values and generating many different spectral-class maps. The classification experiment suggests that the DALA-unsupervised classification algorithm can easily overcome the limitations caused by a high number of NSC, and the increase in NSC to 128 can be an effective measure to significantly improve classification accuracy.

The substitution of reference data for the analyst’s experience or ground knowledge in assigning spectral classes to information classes in Step 2 is the greatest strength of the DALA-unsupervised classification algorithm. Unlike the traditional unsupervised classification approaches, where subjectivity increases, repeatability decreases, and labor and costs increase as NSC increases, the DALA-unsupervised classification makes it possible for analysts to categorize dozens to hundreds of spectral classes into a handful of meaningful information classes without subjectivity and difficulty. The automated classification procedure allows numerous classifications with different parameter combinations, which was not possible with conventional labeling approaches in the past. By assessing all the classifications with an independent reference data set, it is possible to maximize classification accuracy with DALA-unsupervised classification algorithm. The comparative classifications in this study demonstrated such a possibility.

The troublesome labeling process involved in conventional unsupervised classification is similar to the training set refinement step in supervised classification (Lillesand et al., 2004). One may argue that the use of reference data really represents the training step in supervised classification, which reintroduces the limitations we supposedly addressed with DALA. This is not likely the case, however. The training data for supervised classification are normally collected at a sub-type level whereas the reference data for DALA are collected at a type (information class) level. The latter is much easier to determine than the former. An independent LULC map is an excellent resource for reference data. A combination of independently derived LULC maps allows one to establish consistency across several years for millions of reference points or pixels. In other words, we should not expend additional time and incur additional costs to create new reference data for the sake of creating data, especially if it is not of better quality than the existing data from earlier LULC maps.

A comparison of the DALA-unsupervised classifications with the conventional supervised or unsupervised classifications not only substantiated the hypothesized benefits, especially increased accuracy of the DALA-unsupervised classification algorithm, but the comparison also provided insights on the relative importance of NSC and CT values in the development of LULC maps. The DALA-unsupervised approach, for example, can fully utilize the advantages of ISODATA and spectral discrimination of natural clusters. As NSC values increased, classification accuracy increased for all three stages. The unstable-accuracy stage reflects the dynamics of classification accuracy. A temporary stability or decrease in classification accuracy at the unstable-accuracy stage is a misleading signal to an analyst, who may halt using higher NSC in conventional unsupervised classifications. When the DALA-unsupervised classifications are performed, an automated labeling process makes it possible to use high NSC and a high-accuracy can be obtained. The data-assisted process ensures objectivity of the image data classification, thus substantiating the advantages of the DALA method over conventional unsupervised and supervised classification methods.

Second, CT exhibits a minor role in the DALA-unsupervised classification algorithm compared to NSC. The effect of a CT value is noticeable for low values of NSC, viz. less than the 16 obtained in this study. However, higher CT can help achieve more
stable relationships between classification accuracy and NSC. Therefore, high CT is necessary for avoiding the temporarily-stable phenomenon. Unfortunately, either greater CT or greater NSC values create greater computational demands. If the image data set is large, increasing NSC is more effective than increasing CT for increasing classification accuracy given computation capabilities.

Third, using higher NSC values requires more reference data points. Ideally, the two reference data sets need to cover the intersections of spectral classes and information clusters. Though this requirement and our program’s demand for two reference data create the impression of a large data set, however, our case study suggests data requirements are manageable. The two reference data sets, almost 16,000 pixels, represent only 0.4% of the entire image. This small percentage of the map image devoted to reference data allowed us to utilize as many as 512 spectral classes and assign spectral classes to four LULC types while improving overall classification accuracy.

Fourth, creating more LULC types generally requires a larger number of NSC and, thus, requires larger reference data sets as well. The minimum sample size should be set to ensure there is at least one sample point for each spectral class. If there is an empty match or an equal match of information classes from Reference Data I corresponding to a spectral class, the size of Reference Data I needs to be increased. Our experience suggests that if a pixel is the sample unit and pixels are selected with spatial overlays of reference polygons, the sample size can be easily increased while the proportion of sample-data area to the entire image area can still remain small. There is supposed to be an optimum NSC relative to a given number of LULC types because low NSC results in low classification accuracy and high NSC has a high demand on reference data. In this classification experiment, the optimum NSC was 128, the beginning of the high-accuracy stage. However, the optimum NSC could vary with data types and classification schemes, and has to be determined by using the DALA-unsupervised classification algorithm on a case-by-case basis.

When a higher number of LULC types are involved and/or a larger geographic area is targeted in classification, the same three-stage DALA-unsupervised classification algorithm can still be used. Because the classification procedure is automated, increasing LULC and/or expanding land area would not necessarily increase classification complexity or time. However, any increase in LULC and/or land area would require a higher NSC and, in turn, would require a larger sample size for the two reference data sets. In this case, the same principles of sampling technique in accuracy assessment (Congalton and Green, 1999) are applicable for collecting the two reference data sets.

In summary, conventional unsupervised and supervised classification approaches have advantages and disadvantages (Campbell, 2007). The DALA-unsupervised classification algorithm described and tested in this paper represents an important step in strengthening the advantages of unsupervised classification while addressing the limitations of this method. Specifically, the DALA-unsupervised reduces the analyst’s role significantly while expanding the analysis to include multiple NSC and CT values and capitalizing on previous work from other projects. The end result suggests that the development of more accurate LULC maps can be realized by reducing the subjectivity introduced into a classification project by the analyst. The automated labeling process can increase repeatability and lower time commitments and costs of a classification project. More experiments need to be conducted to validate or invalidate the relative impact of NSC values on classification accuracy as well as substantiate our belief that the reference data sets do not increase significantly as the area in the experiment increases. It is also likely that with each new experiment, the DALA-unsupervised classification algorithm will become even more standardized, thus further contributing to its widespread use.

Acknowledgments

The authors thank nine graduate students for their active involvement in the comparative classification experiment. They are Amelie Davis, Jonah Duckles, Martin Gauthier, Julie Crick, James Monty, Brian Napoletano, Avantika Regmi, Angel Torres, and Yong Yuan. The two anonymous reviewers provided detailed comments and constructive suggestions in revising the manuscript. This work was supported in part by the Department of Forestry and Natural Resources, Purdue University.

References