Int. J. Math. And Appl., 6(1–E)(2018), 1239–1245

ISSN: 2347-1557

Available Online: http://ijmaa.in/

International Journal of Mathematics And its Applications

Credit Risk Analytic Using Logistic Regression and Decision Trees

Manoj Kumar Srivastava[1,*], Namita Srivastava[2] and Sandeep[3]

1 Department of Statistics, Institute of Social Sciences, Dr. B.R. Ambedkar University, Agra, Uttar Pradesh, India.

2 Department of Statistics, St. John's College, Agra, Uttar Pradesh, India.

3 NITI AAYOG, Government of India, New Delhi, India.

Abstract: This paper examines current practices, problems, and prospects of data and image classification. Data and Image classification is a complex process that may be affected by many factors. The emphasis is placed on the summarization of major advanced classification approaches and the techniques used for improving classification accuracy. In addition, some important issues affecting classification performance are discussed. This literature review suggests that designing a suitable data and image processing procedure is a prerequisite for a successful classification of remotely sensed data into a thematic map. Effective use of multiple features of remotely sensed data and the selection of a suitable classification method are especially significant for improving classification accuracy. Non parametric classifiers such as neural network, decision tree classifier, and knowledge based classification have increasingly become important approaches for multisource data classification. Integration of remote sensing, geographical information systems (GIS), and expert system emerges as a new research frontier. More research, however, is needed to identify and reduce uncertainties in the image-processing chain to improve classification accuracy.

Keywords: Credit Risk, Regression, Decision Trees.

1. Introduction

Remote sensing research focusing on image classification has long attracted the attention of the remote sensing community because classification results are the basis for many environmental and socioeconomic applications. Scientists and practitioners have made great efforts in developing advanced classification approaches and techniques for improving classification accuracy (Gong and Howarth 1992, Kontoes et al. 1993, Foody 1996, San Miguel Ayanz and Biging 1997, Aplin et al. 1999a, Stuckens et al. 2000, Franklin et al. 2002, Pal and Mather 2003, Gallego 2004). However, classifying remotely sensed data into a thematic map remains a challenge because many factors, such as the complexity of the landscape in a study area, selected remotely sensed data, and imageprocessing and classification approaches, may affect the success of a classification. Although much previous research and some books are specifically concerned with image classification (Tso and Mather 2001, Landgrebe 2003), a comprehensive uptodate review of classification approaches and techniques is not available. Continuous emergence of new classification algorithms and techniques in recent years necessitates such a review, which will be highly valuable for guiding or selecting a suitable classification procedure for a specific study. The foci of this paper are on providing a summarization of major advanced classification methods and techniques used for improving classification accuracy, and on discussing important issues affecting the success of image classifications. Common classification approaches, such as

[*] *E-mail: mkiss87@gmail.com*

ISODATA, Kmeans, minimum distance, and maximum likelihood, are not discussed here, since the readers can find them in many textbooks.

2. Remote-Sensing Classification Process

Remote sensing classification is a complex process and requires consideration of many factors. The major steps of image classification may include determination of a suitable classification system, selection of training samples, image preprocessing, feature extraction, selection of suitable classification approaches, postclassification processing, and accuracy assessment. The user's need, scale of the study area, economic condition, and analyst's skills are important factors influencing the selection of remotely sensed data, the design of the classification procedure, and the quality of the classification results. This section focuses on the description of the major steps that may be involved in image classification.

2.1. Selection of remotely sensed data

Remotely sensed data, including both airborne and spaceborne sensor data, vary in spatial, radiometric, spectral, and temporal resolutions. Understanding the strengths and weaknesses of different types of sensor data is essential for the selection of suitable remotely sensed data for image classification. Some previous literature has reviewed the characteristics of major types of remotesensing data (Barnsley 1999, Estes and Loveland 1999, Althausen 2002, Lefsky and Cohen 2003). For example, Barnsley (1999) and Lefsky and Cohen (2003) summarized the characteristics of different remotesensing data in spectral, radiometric, spatial, and temporal resolutions; polarization; and angularity. The selection of suitable sensor data is the first important step for a successful classification for a specific purpose (Phinn 1998, Jensen and Cowen 1999, Phinn et al. 2000, Lefsky and Cohen 2003). It requires considering such factors as user's need, the scale and characteristics of a study area, the availability of various image data and their characteristics, cost and time constraints, and the analyst's experience in using the selected image.Scale, image resolution, and the user's need are the most important factors affecting the selection of remotely sensed data. The user's need determines the nature of classification and the scale of the study area, thus affecting the selection of suitable spatial resolution of remotely sensed data. Previous research has explored the impacts of scale and resolution on remotesensing image classification (Quattrochi and Goodchild 1997). In general, a finescale classification system is needed for a classification at a local level, thus high spatial resolution data such as IKONOS and SPOT 5 HRG data are helpful. At a regional scale, medium spatial resolution data such as Landsat TM/ETM+, and Terra ASTER are the most frequently used data. At a continental or global scale, coarse spatial resolution Data such as AVHRR, MODIS, and SPOT Vegetation are preferable.

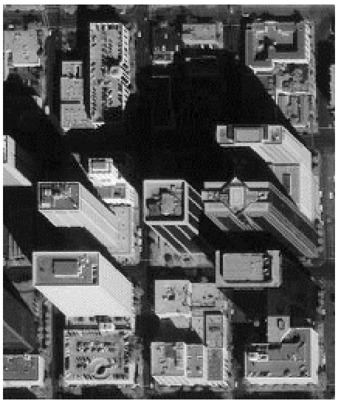

Figure 1. IKONOS DATASET

Another important factor influencing the selection of sensor data is the atmospheric condition. The frequent cloudy conditions in the moist tropical regions are often an obstacle for capturing high quality optical sensor data. Therefore, different kinds of radar data serve as an important supplementary data source. Since multiple sources of sensor data are now readily available, image analysts have more choices to select suitable remotely sensed data for a specific study. A combination of multisensor data with various image characteristics is usually beneficial to the research (Lefsky and Cohen 2003). In this situation, economic condition is often an important factor that affects the selection of remotely sensed data and the time and labour that can be devoted to the classification procedure, thus affecting the quality of the classification results.

2.2. Selection of a classification system and training samples

A suitable classification system and a sufficient number of training samples are prerequisites for a successful classification. Cingolani et al., (2004) identified three major problems when medium spatial resolution data are used for vegetation classifications: defining adequate hierarchical levels for mapping, defining discrete land-cover units discernible by selected remote-sensing data, and selecting representative training sites. In general, a classification system is designed based on the user's need, spatial resolution of selected remotely sensed data, compatibility with previous work, image-processing and classification algorithms available, and time constraints. Such a system should be informative, exhaustive, and separable (Jensen 1996, Landgrebe 2003). In many cases, a hierarchical classification system is adopted to take different conditions into account.

A sufficient number of training samples and their representativeness are critical for image classifications (Hubert-Moy et al., 2001, Chen and Stow 2002, Landgrebe 2003, Mather 2004). Training samples are usually collected from fieldwork, or from fine spatial resolution aerial photographs and satellite images. Different collection strategies, such as single pixel, seed, and polygon, may be used, but they would influence classification results, especially for classifications with fine spatial resolution image data (Chen and Stow 2002). When the landscape of a study area is complex and heterogeneous, selecting sufficient training samples becomes difficult. This problem would be complicated if medium or coarse spatial resolution data are used for classification, because a large volume of mixed pixels may occur. Therefore, selection of training samples must consider the spatial resolution of the remote-sensing data being used, availability of ground reference data, and the complexity of landscapes in the study area.

2.3. Data preprocessing

Image preprocessing may include the detection and restoration of bad lines, geometric rectification or image registration, radiometric calibration and atmospheric correction, and topographic correction. If different ancillary data are used, data conversion among different sources or formats and quality evaluation of these data are also necessary before they can be incorporated into a classification procedure. Accurate geometric rectification or image registration of remotely sensed data is a prerequisite for a combination of different source data in a classification process. Many textbooks and articles have described this topic in detail (Jensen 1996, Toutin 2004). Therefore, it is not discussed here.

If a single date image is used in classification, atmospheric correction may not be required (Song et al., 2001). When multitemporal or multisensor data are used, atmospheric calibration is mandatory. This is especially true when multisensor data, such as Landsat TM and SPOT or Landsat TM and radar data, are integrated for an image classification. A variety of methods, ranging from simple relative calibration and dark-object subtraction to calibration approaches based on complex models (e.g. 6S), have been developed for radiometric and atmospheric normalization and correction (Markham and Barker 1987, Gilabert et al., 1994, Chavez 1996, Stefan and Itten 1997, Vermote et al., 1997, Tokola et al., 1999, Heo and FitzHugh 2000, Song et al., 2001, Du et al., 2002, McGovern et al., 2002, Canty et al., 2004, Hadjimitsis et al., 2004).

Topographic correction is another important aspect if the study area is located in rugged or mountainous regions (Teillet et al., 1982, Civco 1989, Colby 1991, Meyer et al., 1993, Richter 1997, Gu and Gillespie 1998, Hale and Rock 2003). A detailed description of atmospheric and topographic correction is beyond the scope of this paper. Interested readers may check relevant references to identify a suitable approach for a specific study.

3. Feature extraction and selection

Selecting suitable variables is a critical step for successfully implementing an image classification. Many potential variables may be used in image classification, including spectral signatures, vegetation indices, transformed images, textural or contextual information, multitemporal images, multisensor images, and ancillary data. Due to different capabilities in land cover separability, the use of too many variables in a classification procedure may decrease classification accuracy (Hughes 1968, Price et al., 2002). It is important to select only the variables that are most useful for separating land-cover or vegetation classes, especially when hyperspectral or multisource data are employed. Many approaches, such as principal component analysis, minimum noise fraction transform, discriminant analysis, decision boundary feature extraction, non-parametric weighted feature extraction, wavelet transform, and spectral mixture analysis (Myint 2001, Okin et al., 2001, Rashed et al., 2001, Asner and Heidebrecht 2002, Lobell et al., 2002, Neville et al., 2003, Landgrebe 2003, Platt and Goetz 2004) may be used for feature extraction, in order to reduce the data redundancy inherent in remotely sensed data or to extract specific land cover information.

Optimal selection of spectral bands for classifications has been extensively discussed in previous literature (Mausel et al., 1990, Jensen 1996, Landgrebe 2003). Graphic analysis (e.g. bar graph spectral plots, co spectral mean vector plots, two dimensional feature space plot, and ellipse plots) and statistical methods (e.g. average divergence, transformed divergence, Bhattacharyya distance, Jeffreys Matusita distance) have been used to identify an optimal subset of bands (Jensen 1996). Penaloza and Welch (1996) explored the fuzzy logic expert system for feature selection. Peddle and Ferguson (2002) examined three approaches (exhaustive search by recursion, isolated independent search, and sequential dependent search) for optimizing the selection of multisource data, and found that these approaches were applicable to a variety of data analyses. In practice, a comparison of different combinations of selected variables is often implemented, and a good reference dataset is vital. In particular, a good representative dataset for each class is key for implementing a supervised classification. The divergence related algorithms are often used to evaluate the class separability and then to refine the training samples for each class.

3.1. Selection of a suitable classification method

Many factors, such as spatial resolution of the remotely sensed data, different sources of data, a classification system, and availability of classification software must be taken into account when selecting a classification method for use. Different classification methods have their own merits. The question of which classification approach is suitable for a specific study is not easy to answer. Different classification results may be obtained depending on the classifier(s) chosen. These classifiers are based on different mathematical functions.

(a). 2D polynomial functions, such as:

$$Q_{2D}(XY) = \sum_{i=0}^{m} \sum_{j=0}^{n} a_{ij} X^i Y j$$

(b). 3D polynomial functions, such as:

$$Q_{3D}(XYZ) = \sum_{i=0}^{m} \sum_{j=0}^{n} \sum_{k=0}^{p} a_{ijk} X^i Y j Z^k$$

(c). 3D RF such as:

$$R_{3D}(XYZ) = \frac{\sum\limits_{i=0}^{m}\sum\limits_{j=0}^{n}\sum\limits_{k=0}^{p} a_{ijk}X^iYjZ^k}{\sum\limits_{i=0}^{m}\sum\limits_{j=0}^{n}\sum\limits_{k=0}^{p} b_{ijk}X^iY^jZ^k}$$

where:X, Y, Z are the terrain or cartographic coordinates; i, j, k are integer increments; and m, n and p are integer values, generally comprised between 0 and 3, with $m + n(+p)$ being the order of the polynomial functions, generally three.

3.2. Post classification processing

Traditional per pixel classifiers may lead to ‘salt and pepper’ effects in classification maps. A majority filter is often applied to reduce the noises. Most image classification is based on remotely sensed spectral responses. Due to the complexity of biophysical environments, spectral confusion is common among land cover classes. Thus, ancillary data are often used to modify the classification image based on established expert rules. For example, forest distribution in mountainous areas is related to elevation, slope, and aspects. Data describing terrain characteristics can therefore be used to modify classification results based on the knowledge of specific vegetation classes and topographic factors. In urban areas, housing or population density is related to urban land use distribution patterns, and such data can be used to correct some classification confusions between commercial and high-intensity residential areas or between recreational grass and crops. Although commercial and high-intensity residential areas have similar spectral signatures, their population densities are considerably different. Similarly, recreational grass is often found in residential areas, but pasture and crops are largely located away from residential areas, with sparse houses and a low population density. Thus, expert knowledge can be developed based on the relationships between housing or population densities and urban land use classes to help separate recreational grass from pasture and crops. Previous research has indicated that post classification processing is an important step in improving the quality of classifications (Harris and Ventura 1995, Murai and Omatu 1997, Stefanov et al., 2001, Lu and Weng 2004).

3.3. Evaluation of classification performance

Evaluation of classification results is an important process in the classification procedure. Different approaches may be employed, ranging from a qualitative evaluation based on expert knowledge to a quantitative accuracy assessment based on sampling strategies. To evaluate the performance of a classification method, Cihlar et al., (1998) proposed six criteria: accuracy, reproducibility, robustness, ability to fully use the information content of the data, uniform applicability, and objectiveness. In reality, no classification algorithm can satisfy all these requirements nor be applicable to all studies, due to different environmental settings and datasets used. De Fries and Chan (2000) suggested the use of multiple criteria to evaluate the suitability of algorithms. These criteria include classification accuracy, computational resources, stability of the algorithm, and robustness to noise in the training data. Classification accuracy assessment is, however, the most common approach for an evaluation of classification performance, which is detailed in §3.

4. Classification Accuracy Assessment

Before implementing a classification accuracy assessment, one needs to know the sources of errors (Congalton and Green 1993, Powell et al., 2004). In addition to errors from the classification itself, other sources of errors, such as position errors resulting from the registration, interpretation errors, and poor quality of training or test samples, all affect classification accuracy. In the process of accuracy assessment, it is commonly assumed that the difference between an image classification result and the reference data is due to the classification error. However, in order to provide a reliable report on classification accuracy, non image classification errors should also be examined, especially when reference data are not obtained from a field survey. A classification accuracy assessment generally includes three basic components: sampling design, response design, and estimation and analysis procedures (Stehman and Czaplewski 1998). Selection of a suitable sampling strategy is a critical step (Congalton 1991). The major components of a sampling strategy include sampling unit (pixels or polygons), sampling design, and sample size (Muller *et al.* 1998). Possible sampling designs include random, stratified random, systematic, double, and cluster sampling. A detailed description of sampling techniques can be found in previous literature such as Stehman and Czaplewski (1998) and Congalton and Green (1999).

The error matrix approach is the one most widely used in accuracy assessment (Foody 2002b). In order to properly generate an error matrix, one must consider the following factors: (1) reference data collection, (2) classification scheme, (3) sampling scheme, (4) spatial autocorrelation, and (5) sample size and sample unit (Congalton and Plourde 2002). After generation of an error matrix, other important accuracy assessment elements, such as overall accuracy, omission error, commission error, and kappa coefficient, can be derived. Previous literature has defined the meanings and provided computation methods for these elements (Congalton and Mead 1983, Hudson and Ramm 1987, Congalton 1991, Janssen and van der Wel 1994, Kalkhan et al. 1997, Stehman 1996, 1997, Congalton and Green 1999, Smits et al. 1999, Congalton and Plourde 2002, Foody 2002b, 2004a). Meanwhile, many authors, such as Congalton (1991), Janssen and van der Wel (1994), Smits et al. (1999), and Foody (2002b), have conducted reviews on classification accuracy assessment. They have assessed the status of accuracy assessment of image classification, and discussed relevant issues. Congalton and Green (1999) systematically reviewed the concept of basic accuracy assessment and some advanced topics involved in fuzzy logic and multilayer assessments, and explained principles and practical considerations in designing and conducting accuracy assessment of remote sensing data. The Kappa coefficient is a measure of overall statistical agreement of an error matrix, which takes non diagonal elements into account. Kappa analysis is recognized as a powerful method for analysing a single error matrix and for comparing the differences between various error matrices (Congalton 1991, Smits et al. 1999, Foody 2004a). Modified kappa coefficient and tau coefficient have been developed as improved measures of classification accuracy (Foody 1992, Ma and Redmond 1995). Moreover, accuracy assessment based on a normalized error matrix has been conducted, which is regarded as a better presentation than the conventional error matrix (Congalton 1991, Hardin and Shumway 1997, Stehman 2004).

The error matrix approach is only suitable for 'hard' classification, assuming that the map categories are mutually exclusive and exhaustive and that each location belongs to a single category. This assumption is often violated, especially for classifications with coarse spatial resolution imagery. 'Soft' classifications have been performed to minimize the mixed pixel problem using a fuzzy logic. The traditional error matrix approach is not appropriate for evaluating these soft classification results. Accordingly, many new measures, such as conditional entropy and mutual information (Finn 1993, Maselli et al. 1994), fuzzy set approaches (Gopal and Woodcock 1994, Binaghi et al. 1999, Woodcock and Gopal 2000), symmetric index of information closeness (Foody 1996), Renyi generalized entropy function (Ricotta and Avena 2002), and parametric generalization of Morisita's index (Ricotta 2004) have been developed. However, one critical issue in assessing fuzzy classifications is the difficulty of collecting reference data. More research is thus needed to find a suitable approach for evaluating fuzzy classification results.

5. Conclusion

The error matrix approach is the most common accuracy assessment approach for categorical classes and classification of data. Uncertainty and confidence analysis of classification results has gained some attention (McIver and Friedl 2001, Liu *et al.* 2004), and spatially explicit data on mapping confidence are regarded as an important aspect in effectively employing classification results for decision making (McIver and Friedl 2001, Liu et al. 2004). Image classification has made great progress over the past decades in the following three areas: (1) development and use of advanced classification algorithms, such as subpixel, per-field, and knowledge-based classification algorithms; (2) use of multiple remote-sensing features, including spectral, spatial, multitemporal, and multisensor information; and (3) incorporation of ancillary data into classification procedures, including such data as topography, soil, road, and census data. Accuracy assessment is an integral part in an image classification procedure. Accuracy assessment based on error matrix is the most commonly employed approach for evaluating per-pixel classification, while fuzzy approaches are gaining attention for assessing fuzzy classification results. Uncertainty and error propagation in the image-processing chain is an important factor influencing classification accuracy. Identifying the weakest links in the chain and then reducing the uncertainties are critical for improvement of classification accuracy. The study of uncertainty will be an important topic in the future research of image classification.

References

[1] T. Toutin, *Geometric processing of remote sensing images: models, algorithms and methods*, International Journal of Remote Sensing, 25(10)(2004), 1893-1924.

[2] M. F. Goodchild and J. Proctor, *Scale in a digital geographic world*, Geographical and Environmental Modelling, 1(1)(1997), 5-23.

[3] J. W. Wiesel, *Image rectification and registration. International Archives of Photogrammetry and Remote Sensing, Rio de Janeiro*, Brazil (Brazil: ISPRS), 25(A3b)(1984), 1120-1129.

[4] D. Lu and Q. Weng, *A survey of image classification methods and techniques for improving classification performance*, International Journal of Remote Sensing, 28(5)(2007), 823-870.

[5] G. P. Asner and K. B. Heidebrecht, *Spectral unmixing of vegetation, soil and dry carbon cover in arid regions: comparing multispectral and hyperspectral observations*, International Journal of Remote Sensing, 23(2002), 3939-3958.

[6] I. Bloch, *Information combination operators for data fusion: a comparative review with classification*, IEEE Transactions on Systems, Man, and Cybernetics, 26(1996), 52-67.

[7] C. Conese and F. Maselli, *Evaluation of contextual, per-pixel and mixed classification procedures applied to a subtropical landscape*, Remote Sensing Reviews, 9(1994), 175-186

[8] G. M. Foody, *Hard and soft classifications by a neural network with a nonexhaustively defined set of classes*, International Journal of Remote Sensing, 23(2002), 3853-3864.

[9] E. Binaghi, P. Madella, M. G. Montesano and A. Rampini, *Fuzzy contextual classification of multisource remote sensing images*, IEEE Transactions on Geoscience and Remote Sensing, 35(1997), 326-339.

[10] C. Ricotta and G. C. Avena, *The influence of fuzzy set theory on the areal extent of thematic map classes*, International Journal of Remote Sensing, 20(1999), 201-205.

[11] L. Zadeh, *Fuzzy Sets*, Information and Control, 8(1965), 338-353.