CLASSIFIER SHOOTOUT: A QUANTITATIVE ASSESSMENT OF THREE POPULAR IMAGE CLASSIFICATION METHODOLOGIES

David M. Johnson, Geographer
National Agricultural Statistics Service
United States Department of Agriculture
3251 Old Lee Highway, suite 305
Fairfax, VA 22030
dave_johnson@nass.usda.gov

ABSTRACT

The effectiveness of three popular commercial supervised image classification methods is presented. They are the maximum likelihood (Leica Geosystems Imagine), classification tree (Rulequest See5.0), and object-oriented (Definiens Professional). Analyses of the methods are tested against Landsat TM and Resourcesat AWiFS data. The imagery utilized was acquired during mid-summer over the agriculturally intensive Upper Midwestern United States and Mississippi River Alluvial Plain. Ground truth data from the comprehensive June Agricultural Survey is used for training and validation of the classification outputs. Quantitative assessments of each classification, emphasizing the ability to discriminate agricultural cover types at regional scales, are compared along with subjective ratings. Practical considerations are also discussed for efficiently and cost effectively using the tools beyond a research environment. Results are guiding the National Agricultural Statistics Service's ability to improve and expand its annual Cropland Data Layer campaign.

INTRODUCTION

Since the early 1990s, the National Agricultural Statistics Service (NASS) has developed summer-based land cover classifications, with an emphasis on documenting row crop agriculture, over certain intensive growing regions of the United States (US). The main goal of the work has been to derive acreage statistic estimates to improve or validate those collected from NASS' traditional annual probability survey programs. Landsat-5 TM and Landsat-7 ETM+ imagery have traditionally been used as the data source, but within the last couple of years a switch to Resourcesat-1 56m resolution AWiFS data was made by the agency. Regardless, TM, ETM+ and AWiFS imagery are all well suited for monitoring agriculture.

NASS began the crop classification program before the advent of today's commercial image processing software and has relied on an internally developed program, PEDITOR, in order to manage the image and ground truth data and perform the classification. PEDITOR classifications are based on maximum likelihood (ML) analysis. The training data to derive this supervised style classification usually comes from the June Agricultural Survey (JAS) which is a probability-based area frame survey sponsored annually by NASS. Ground truth is enumerated during June at thousands of approximately one square mile sample sites focused within the intensive agricultural regions of the US. This data set allows NASS to provide planted acreage estimates for the major crop commodities to policy makers in a timely fashion. As a side benefit, it also allows for a natural mechanism to perform supervised style image classification to better map the distribution of crops.

Given the advance of "commercial-off-the-shelf" GIS software and, in particular, the methodologies they rely on to derive image classification, NASS has spent time evaluating the costs and benefits of the most promising ones. In particular, NASS has investigated boosted classification tree (BCT) products as derived from Rulequest See5.0 and object-oriented (OO) analysis via Definiens Professional (formally known as eCognition). Both have created a certain amount of "buzz" within the remote sensing community and from an operational standpoint have become viable. To many, including NASS, the ML method is considered the longtime standard and thus an appropriate benchmark in which to compare others.

The ultimate goal of the NASS spatial analysis group is to most accurately document agricultural cover types, essentially creating a cropland census, for a given growing season over large regions. To reach this goal, other practical factors besides just map accuracy as derived from image processing software are of concern. They include ease of use, cost, scalability, repeatability, and speed. In other words, the most accurate method may not necessarily be the best if it is difficult to use, expensive, unable to accommodate large datasets, gives different users different results, or time consuming. Thus, findings go beyond that of a just a typical error matrix of accuracy assessment results.

BACKGROUND

Many methods for image classification exist but probably the most common and widely used methodology is the ML. The basic premise is to assign each raw image pixel to a predetermined class deemed to be the most probable based on the mean measurement and variance of the spectral properties of ground truth data. ML analysis is parametric in nature and assumes the distribution of the training data is normal. Most commercial remote sensing software has the ability to derive a ML classification. NASS' in-house software, PEDITOR, performs ML classifications but because it is not widely distributed (although it is free and in the public domain), has a step learning curve, a limited graphical use interface (GUI), and data type limitations, this paper relies on the commercial version within Leica Geosystems ERDAS Imagine. The underlying routines should be the same for both though and it is believed the classification output would be very similar. The estimated cost for an ERDAS Imagine, a full blown raster based GIS package, is around a few thousand US dollars. Other comparable software that can also perform a ML classification is likely similarly priced.

Classification tree (often referred to more generally as decision tree) analysis has gained favor of late. Its basic premise is to divide the pixels based on their spectral properties into hierarchical subgroups. The divisions take place in a binomial fashion and at each step aims to minimize the heterogeneity among the split data. A primary advantage is the procedure is not parametric in nature and thus no prior assumption about the distribution of the data is needed. Because of this, it allows one the ability to utilize non-continuous (e.g. ancillary classifications) and non-normally distributed data (e.g. elevation) into deriving classifications. Various methods on how to exactly derive the decision trees have been created but shown to have similar outcomes. Thus, different commercial products defined as classification tree software exist (none of which are embedded within a GIS). The discussion here utilizes Rulequest See5.0 for a couple of major reasons. First off, it incorporates an advanced technique known as "boosting" which for most cases improves the classification outcome. Secondly, an interface has been written for it by the United States Geological Survey it to facilitate the easy transfer the analysis data to and from ERDAS Imagine. A single license of See5.0 costs many hundreds of US$ at the time of writing. However, See5.0 is only a text based analysis tool and not a GIS in itself and thus reliant on image processing software (i.e. ERDAS Imagine).

Object-oriented analysis is relatively new and incorporates contextual image information into the classification in addition to spectral information. The leading commercial software for this approach is Definiens Professional. There is a certain amount of excitement within the remote sensing community for the methodology because of the promise that spatial information can increase classification accuracy over pure pixel-based approaches. The OO approach is basically a two step process. The first is "segmenting" of the raw imagery into a full lattice of vector polygons which attempt to maximize the within polygon homogeneity based on the pixel reflectance values (i.e. divide the image into areas of like pixels). The second step is the classification of those polygons into known cover types through rules either based on spectral characteristics of the pixels within each polygon (e.g. mean, standard deviation) or on the geometric properties of the polygons themselves (e.g. size, orientation, relation to surrounding polygons). The rules can be derived either by training samples or through user defined knowledge-based decision rules. At the time or writing, Definiens' software license costs many thousands of US$. It is a GIS per se but lacks many of the normal raster based tools such as image mosaicking, reprojecting, clipping, resampling etc. thus in most work environments it would need to be used in conjunction with true image processing software.

METHODS AND RESULTS SUMMARY

NASS has experimented with all the of the methodologies in variety of US agriculture regions including the citrus region of Florida, the Mississippi River Alluvial Plain, and the upper Midwest in areas like North Dakota and Wisconsin. Data sources to test have always been Landsat (TM and ETM+), and Resourcesat-1 (AWiFS and LISS-III). All testing have been performed using independent training and validation sets of ground truth data and error matrices constructed. Most recent tests have been run on an Intel Dual Core Windows XP based PC with 2GB RAM.

The first test, a shootout between OO versus ML, was for a pilot project to detect citrus groves within Indian River County in the state of Florida utilizing Landsat TM data. The goal was to produce a land cover product as to detect where citrus groves had either gone out of or into production. Because NASS already maintains a biennial inventory of the citrus groves, a mechanism to have a full census of ground truth for perfect accuracy assessment was already in place. A best classification effort was made with the ML method and then with OO. Within the OO paradigm both the membership function and the nearest neighbor approaches were attempted. The nearest neighbor method was found to be preferred and more accurate. However, the ML still outperformed either OO classification effort modestly. A hybrid approached using the base classification from the ML, overlaying the lattice of OO segments, and performing a majority filter within each polygon gave the best overall results by a couple of percentage points.

Testing of coincident, in date and area, TM and AWiFS was performed over the Mississippi River Alluvial Plain. BCT versus ML maps were constructed. Classification trees were shown to be slightly better than the ML. Both classifications were found to be improvable by a few percentage points by running a minimum mapping unit "clump;-and-eliminate" style filter after the fact to clean up spurious pixels. It has been found through sensitivity testing that a minimum mapping value of 20 acres (corresponding to 26 AWiFS pixels or 90 TM) is the best. This is also consistent with average field sizes within the US which tend to not be smaller than 20 acres.

More comprehensive cropland classification derived for a large section of North Dakota using all 3 methods on a single date Resourcesat-1 AWiFS and LISS-III data were constructed. BCT was found to be the best. LISS-III (a 23 m spatial resolution product) was also attempted and BCT was still found to be the best. However, the OO method was not testable because it simply could not handle the data volume. Results were similar (with BCT the best) when run with the same style data over Wisconsin.

DISCUSSION AND CONCLUSION

The BCT analysis has proven to NASS as being as equally accurate, and almost always better, than either the ML or OO. To rank them would place BCT first, ML second and OO third in terms of resulting map accuracy. All results are within a few percentage points of each other and some times not considered significant.

BCT is the most "black box" approach which has its pros and cons. Primarily, there is a feeling of lack of control but of the flip side it does not seem to matter since the classification outperforms the others. Some tuning of See5.0 input parameters, such as pruning levels, have shown to have little impact. However, not using See5.0's boosting option always resulted in lower accuracies by a few percentage points. The BCT analysis seems most reliant on have a representatively sample of ground truth but which is typically not a problem for NASS but could be for others.

OO results were the most varied but they never outperformed the BCT or ML methods. This is probably due to several reasons. First, by nature agricultural fields within the US do not tend to vary in size or shape based on what is planted in them. In other words, a corn field only varies from a soybean field based on its spectral properties. Thus, the OO approach does not gain in performance based on contextual information. This notion was reinforced within the analysis tool within Definiens which always showed spectral reflectance to be the primary variable to classify the polygons. Second, both Landsat and Resourcesat pixels are relatively coarse. OO analysis probably would show more impact if the objects of interest had a higher pixel to object ratio (and thus more ability for spatial properties to show up). Although it was not tested directly, the OO method seems it would perform best when the classified feature of interested all contain unique scale attributes. Furthermore, OO would probably have the most to gain with a lot of user interaction of the input parameters and classification schemes. This can be good for the expert user but is less attractive to the novice. And finally, even though output was typically not as good in terms of map accuracy with the OO, subjectively it is easy not to be impressed by the look of the results which lack the speckling of traditional pixel based classifiers. Some might choose the OO approach simply for that reason.

The very best results were derived from a hybrid approach with initial classification using BCT and then post "polishing" them either though a clump-and-eliminate filter (within ERDAS) or a majority filter utilizing the derived OO segments and polygon boundaries. The clump-and-eliminate filter provided a bit more boost to the output accuracy, but again subjectively the OO often looks more pleasing. There is a suggestion with higher resolution data the hybrid OO would win out.

For good or bad, Definiens Professional has a lot of input parameters and there is feeling that with optimization they can lead to the best classification. For example, at the initial step of segmenting the image one is faced with the selection of "scale" and "color" parameters. It is not obvious what is best. Within the Florida example, sensitivity testing was performed to try to optimize the best values. Surprisingly, it was found they have very little impact (unless changed radically) on the output accuracy. The best approach seems to be to pick the default values for the initial multi-spectral resolution segmentation and then as a second step perform spectral difference segmentation with a small value (say 3) to refine the polygons. For the best accuracy is seems the prudent to stick will smaller versus larger polygons if unsure.

On a related note, Definiens Professional was also tested as a potential solution to the application of counting citrus trees in support of the Biennial Commercial Citrus Inventory Program administered by the NASS Florida Field Office. The basic premise was to utilize higher resolution imagery (approximately 1 m sq) and design an OO method to automate the tree counting. Segmentation was performed to roughly divide tree crowns and shadows from the surrounding. Then a decision rule was performed to select all of the polygons that contained lesser spectral means than those surrounding (thus picking all of the shadows). Finally those shadows were all added up (it was not as successful using the tree tops). The results tended to be biased conservatively (i.e. didn't count as many tree shadows as really exist) but with an adjustment factored in it could came up with reasonable tree counts. This type of more GIS-based analysis seems to be were the OO approach can really be beneficial. And again, it reinforces the idea that the OO approach has the most to gain with higher resolution imagery.

While the BCT analysis performed better than the other methods for the map output it is also worth considering the feasibility of using to in an operational setting where factors such as trainability, cost, and speed are important. Table 1 shows a rundown of the, albeit somewhat subjective, findings.

Table 1

	Maximum Likelihood Analysis	Boosted Classification Tree Analysis	Object-oriented Analysis
Map Accuracy	Good	Best	Fair
Ease of use	Good	Good	Good
Cost	Best	Good	Poor
Scalability	Fair	Best	Poor
Consistency	Good	Best	Fair
Speed	Good	Best	Poor

Ease of use of the software is probably the most subjective so difficult to judge. For example, even if software is complicated an expert user may not think so. The software cost is easily the most for the OO, likely over double or triple the other methods, while the BCT is only marginally greater than the ML. If real gains were being made then it might be worth the high cost but it is not clear that it is for NASS applications. The scalability is also a large downfall of OO. With increased data collection volumes, NASS has recently been running analysis over large regions which may include 10 to 20 Landsat-like scenes across space and time. OO simply cannot handle the sheer data throughput. BCT handles it with aplomb. Speed is correlated with the data handling to some extent and the BCT analysis can be finished sometimes before the initial segmentation of OO polygons is even finished. All things considered, BCT analysis wins out overall.

While the OO seems to have negatives it is not to suggest it is not worthwhile for certain applications. It likely sees bigger gains with higher resolution datasets and ones that don't contain a huge amount of pixels. Computational power is still on the increase so those issues may become lessened in the future. Also, BCT cannot be performed without training data so OO would be compelling for scenarios where rule-based decision needs to be derived ahead of time. Finally, the hybrid approach of BCT and OO can lead to the most accurate and visually appealing classifications.

REFERENCES

Boryan, C., and M. Craig, 2005. Multiresolution Landsat TM and AWiFS sensor assessment for crop area estimation in Nebraska, *Proceedings from Pecora 16*, Sioux Falls, South Dakota. American Society for

Photogrammetry and Remote Sensing, Bethesda, Maryland, unpaginated CD-ROM.
Congalton, R.G., and K. Green, 1999. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, CRC Press, Boca Raton, Florida, 137 p.
Craig, M.E., 2001. A resource sharing approach to crop identification and estimation, *ASPRS 2001 Proceedings of the 2001 Annual Conference*, St. Louis, Missouri. American Society for Photogrammetry and Remote Sensing, Bethesda, Maryland, unpaginated CD-ROM.
Definiens, 2003. eCognition User Guide, Version 4.0, URL: *http://www.definiens-imaging.com*.
Definiens, 2006. Professional 5 Reference Book, URL: *http://www.definiens-imaging.com*.
Definiens, 2006. Professional 5 User Guide, URL: *http://www.definiens-imaging.com*.
Defries, R.S. and J.C.W. Chan, 2000. Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data, *Remote Sensing of Environment, 74*:503-515.
Friedl, M.A., and C.E. Brodley, 1997. Decision tree classification of land cover from remotely sensed data, *Remote Sensing of Environment*, 61:399–409.
Lawrence, R.L., A. Bunn, S. Powell, and M. Zambon, 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis, *Remote Sensing of Environment*, 90:331–336.
Lawrence, R.L., and A. Wright, 2001. Rule-based classification systems using classification and regression trees (CART) analysis, *Photogrammetric Engineering* & *Remote Sensing*, 67(10):1137–1142.
Quinlan, J.R., 1993. *C4.5 Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, California.
Zambon, M., R. Lawrence, A. Bunn, and S. Powell, 2006. Effect of alternative splitting rules on image processing using classification tree analysis, *Photogrammetric Engineering and Remote Sensing, 72*(1): 25-30.