Recent HARVIST Results: Classifying Crops from Remote Sensing Data

Kiri Wagstaff and Dominic Mazzoni (kiri.wagstaff@jpl.nasa.gov)
Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109

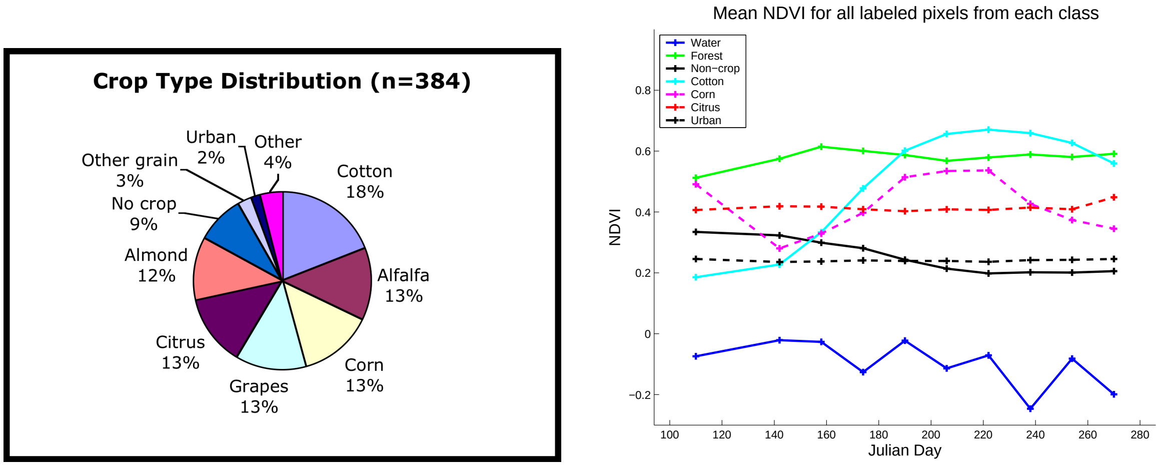
Abstract

In this paper, we report on recent results from the Heterogeneous Agricultural Research Via Interactive, Scalable Technology (HARVIST) project. HARVIST seeks to provide the tools and scalability required to enable practicioners to analyze large, diverse data sets that may come from different data sources. We have focused on agricultural applications, and our current results demonstrate the ability of the system to train a crop type classifier that operates on orbital remote sensing images. We find that this classifier can label crops with an accuracy of 82%, comparable to other published results.

1 Introduction

The Heterogeneous Agricultural Research Via Interactive, Scalable Technology (HARVIST) project commenced on October 1, 2004. Our goal is to integrate multiple Earth Science data sources into a single graphical user interface that allows for the investigation of connections between different variables (Wagstaff et al., 2005). In particular, we focus on relationships between weather and crop yield, but the system we are creating will be capable of integrating data for other studies as well. The data sources are heterogeneous in that they contain information at different spatial, spectral, and temporal resolutions. The HARVIST system will provide multiple machine learning and data analysis algorithms that can be applied to the data. Specifically, we have incorporated support vector machines (SVMs; classification) and clustering (discovery) methods into the system, and our next steps will include adding multivariate spatial modeling (regression and prediction) methods.

Our recent focus has been on building a crop type classifier that operates on remote sensing data collected by the Multi-angle Imaging Spectroradiometer (MISR) instrument (Diner et al., 1988). This classifier can be used to help construct land use/land cover (LULC) data bases for use by scientists or policy makers (Wardlow & Egbert, 2005). This paper reports on preliminary results we have obtained, using ground truth data we collected in August, 2005, for development and evaluation of the classifier. We find that a trained SVM can label cropland with 82% accuracy, comparable to other results on a similar problem, despite differences in the remote sensing instrument, target location, and machine learning method. Taken together, these advances illustrate the benefits of automated data analysis methods. We plan to extend and enhance these results in several ways, including the incorporation of soil type information and developing a regression method to estimate crop yields (e.g., expected bushels of corn per acre) from remote sensing data. We are also very interested in working with scientists on related problems for which HARVIST can be of use.



(a) Distribution of observed crop types; cotton is most prevalent.

(b) 2005 time series plot of the mean NDVI, computed from MISR observations, for crop and non-crop regions.

Figure 1: Ground truth data we collected near Bakersfield, CA, in August, 2005.

2 Preliminary Crop Type Classification Results

In August, 2005, we conducted a field study in the area surrounding Bakersfield, California. We surveyed 384 crop fields and, for each one, recorded the latitude, longitude, and crop type being grown. We also took digital pictures of the majority of the fields for later reference. The top five crops (in terms of number of fields observed) were cotton, corn, citrus, grapes, and almonds; see Figure 1(a). We also collected observations of uncultivated fields and urban areas. It was not possible to collect a regularly spaced sample of fields, due to fences and warnings against trespassing on private land. Therefore, our sample is biased in that we were restricted to fields that were near roads or highways. We were nevertheless able to collect a diverse sample of crops that has provided the necessary basis for developing a crop type classifier.

Once we had collected the crop type labels, we used the latitude and longitude data to match each labeled field with the corresponding location in MISR's remote sensing observations (275 meters per pixel). We created a time series data set that includes observations at four spectral bands, for each location, every 16 days from January to September, 2005. For this study, we only used observations from MISR's nadir-pointing camera, but in the future we intend to incorporate observations from the other angles as well. Gobron et al. (2002), among others, have shown that the use of MISR's multiple angles significantly increases our ability to discover the structure and photosynthetic activity of surface vegetation.

Figure 1(b) shows the mean NDVI (Normalized Difference Vegetation Index) for several of the crop and non-crop regions over the growing season (late March to late September). The crop growth profiles for cotton and corn have a distinctly different shape than that of the forest or non-crop (uncultivated) regions. Citrus is a tree crop, and it is unlikely that we can detect the ripening of the fruit from orbital observations. Therefore, its time profile remains flat, similar to the forest observations. However, its NDVI values are sufficiently distinct to permit us to distinguish citrus trees from forest areas.

	Water	Forest	Non-crop	Almond	Cotton	Alfalfa	Grapes	Corn	Citrus	Urban
Water	**35**	0	0	0	0	0	0	0	0	0
Forest	0	**165**	0	1	0	2	0	0	7	3
Non-crop	0	0	**123**	2	0	0	25	1	1	0
Almond	0	0	3	**94**	0	0	5	2	2	16
Cotton	0	0	0	0	**246**	0	7	0	1	0
Alfalfa	0	0	0	12	3	**46**	0	8	3	2
Grapes	0	0	2	0	12	2	**59**	0	12	1
Corn	0	0	2	26	5	8	3	**38**	2	2
Citrus	0	0	3	14	0	0	6	0	**63**	10
Urban	0	0	3	0	0	0	0	0	0	**147**

Table 1: Confusion matrix for MISR pixels from Bakersfield, CA, when training an SVM on the northern half of each class and testing it on the southern half. Each entry indicates the number of test pixels of the given row type that were classified by the SVM as the column type. Bolded entries show the number of correctly classified items.

Next, we trained an SVM (Cortes & Vapnik, 1995) to classify each pixel into one of ten classes: almond, cotton, alfalfa, grapes, corn, citrus, water, forest, non-crop, and urban. This SVM used a Gaussian kernel ($\sigma = 1.0, C = 10$). The inputs consisted of the time series observations (11 time points × 4 spectral bands = 44 features). The 384 fields that we had recorded resulted in 2480 labeled pixels. We split the data set randomly into two parts for training and testing. By including only observations from the growing season, we were able to exclude most of the cloudy days, which might otherwise skew the classification. The overall classification accuracy obtained was 91%. However, this data set contains strong spatial dependencies. That is, adjacent pixels are likely to be of the same type. Therefore, it can be misleading to randomly partition the data set into training and test sets. To better judge the generalization ability of our crop type classifier, we split the data set spatially: we used the northern half of each crop's pixels for training and the southern half of each crop's pixels for testing.

Using this approach, we obtained an overall accuracy of 82%. We further analyzed the accuracy rates for the different classes in the test set through the use of the confusion matrix shown in Table 1. Values in the (bolded) diagonal entries indicate the number of pixels correctly classified into each class, and the sum of these values, divided by the total number of labeled pixels (1243 in the test set), yields the overall accuracy figure. However, the confusion matrix also allows us to determine which classes are most difficult to distinguish. For example, 26 pixels that were labeled as "corn" were incorrectly classified as "almond". In fact, the "corn" class overall was the most difficult one to label correctly: of the 86 pixels from this class, the classifier only correctly identified 38 (44%). This performance could be improved by increasing the number of labels, and therefore the amount of training data, for this class. In contrast, none of the other classes were ever confused with pixels from the "water" class. In addition, the classifier was 100% correct in its labeling of "forest" pixels. This analysis suggests direct ways in which we might improve our classifier, such as acquiring more labeled cornfields, as well as classes that already have good performance and do not require a further investment of effort to collect ground truth.

Wardlow and Egbert (2005) explored the use of time series obtained by the Moderate Resolution Imaging Spectroradiometer (MODIS) to classify land cover types across the state of Kansas. The MODIS bands they used have a spatial resolution of 250 meters per pixel, similar to that of MISR. They used a hierarchical approach based on decision trees to identify corn, sorghum, soybeans, alfalfa, winter wheat, and fallow fields. They report an accuracy of 84%, slightly higher than ours. However, random sampling was used to create the training and test data sets, which (as we discovered) can inflate accuracy results, so it is difficult to compare these numbers directly.

3 Collaboration Lessons Learned

The most useful lesson learned in this project, with respect to collaborations, is that it is most effective to explicitly seek out experts and invite them to JPL to obtain critical feedback and information about new problems. We met with Marcel Schaap, Dennis Corwin, and Scott Lesch of the U.S. Salinity Laboratory to discuss how HARVIST could be of use for their soil studies. There is potential for collaboration on a soil salinity study starting in June 2006. We also aim to involve agricultural scientists in our current work on crop yield prediction.

4 Conclusions and Next Steps

In this paper, we have presented the HARVIST project, which seeks to provide the tools and scalability required to enable practicioners to analyze large, diverse data sets. Our recent results indicate promising behavior for the task of classifying remote sensing pixels according to the crop present, if there is one. We have identified avenues for further improvement, such as collecting additional ground truth for low-accuracy classes (e.g., corn) and incorporating other data sources, such as soil types.

Given these preliminary results, we plan to further evaluate our crop type classifier by applying it to a larger MISR data set that covers the state of Kansas (100 counties). We will determine how well our ground truth labels from California, and an SVM classifier trained on that information, can generalize to new observations from a different part of the country.

In addition, another major goal of this project is to demonstrate its utility by generating crop *yield* predictions, based on the time series remote sensing observations. We will compare our predictions to actual yield values published by the USDA at a per-county level. We will compare the quality of predictions obtained with and without our crop type classifier's decisions as an additional input. That is, we will specialize our predictions for a given crop based only on remote sensing observations from terrain in which that crop is currently grown, as determined by our classifier.

Acknowledgements

We would like to thank Stephan Sain for his contributions to this project. Mike Kocurek and Tim Stough helped collect the ground-truth crop classifications. We would also like to thank the MISR team, without whom this work would not have been possible. This work was supported by a two-year grant from the NASA Earth-Sun Systems Technology Office and was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References

Cortes, C., & Vapnik, V. (1995). Support-vector network. *Machine Learning*, *20*, 273–297.

Diner, D. J., Beckert, J. C., Reilly, T. H., Bruegge, C. J., Conel, J. E., Kahn, R., Martonchik, J. V., Ackerman, T. P., Gordon, H. R., Muller, J.-P., Myneni, R., Sellers, R. J., Pinty, B., & Verstraete, M. M. (1988). Multiangle Imaging Spectroradiometer (MISR) instrument description and experiment overview. *IEEE Transactions on Geoscience and Remote Sensing*, *36*, 1072–1087.

Gobron, N., Pinty, B., Verstraete, M. M., Widlowski, J.-L., & Diner, D. J. (2002). Uniqueness of multiangular measurements—Part II: Joint retrieval of vegetation structure and photosynthetic activity from MISR. *IEEE Transactions on Geoscience and Remote Sensing*, *40*, 1574–1592.

Wagstaff, K. L., Mazzoni, D., & Sain, S. (2005). HARVIST: A system for agricultural and weather studies using advanced statistical models. *Proceedings of the Earth-Sun System Technology Conference.*

Wardlow, B. D., & Egbert, S. L. (2005). State-level crop mapping in the U.S. Central Great Plains agroecosystem using MODIS 250-meter NDVI data. *Pecora 16: Global Priorities in Land Remote Sensing.*