Heterogeneous Agricultural Research Via Interactive, Scalable Technology (HARVIST): Infusion for the USDA

ESTO Final Report: January 1, 2007 – September 30, 2007
Submitted by Kiri L. Wagstaff, kiri.wagstaff@jpl.nasa.gov, 818-393-6393
Co-investigators: Dennis Corwin (USDA)
Contributors: Lukas Mandrake, Alex Roper, and Lucas Scharenbroich

Project Overview

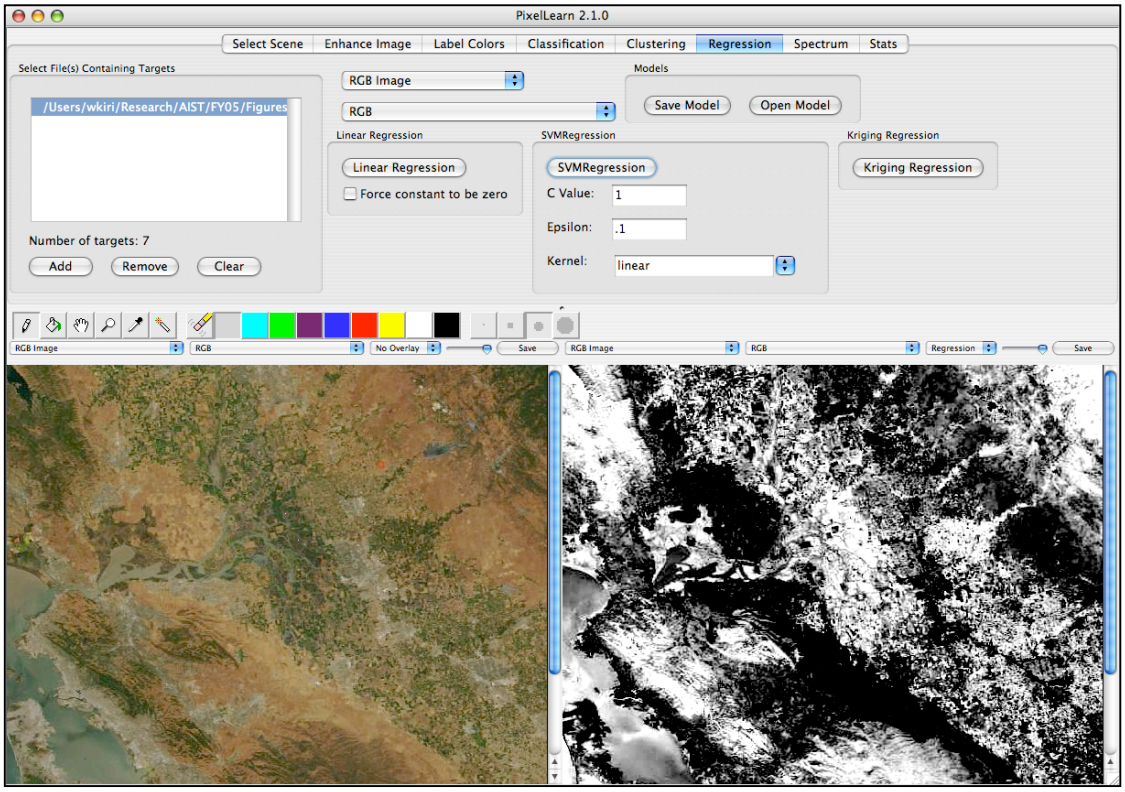
The HARVIST project was originally funded by the NASA ESTO program from October 2004 to October 2006. During the course of this project, we extended an existing graphical classification interface, PixelLearn, to support data sets relevant to earth science studies and added the ability to perform a variety of clustering techniques. We also added a technical advance, progressive rendering, that greatly increased the interactivity of the application. Finally, we conducted experiments with regression methods such as SVMs and Bayesian approaches. We applied these algorithms to problems such as crop type classification and crop yield prediction. As a result of this effort, the technology involved in PixelLearn moved from TRL 4 (individual components) to TRL 6 (integrated system).

In late 2005, we established a connection with soil scientists from the USDA Soil Salinity Laboratory and invited them to JPL to see a demonstration of PixelLearn. Our principal contact has been Dr. Dennis Corwin. We also asked them to provide an overview of the studies they were conducting so that we might see how our software could be of benefit to their investigations. They were in the planning stages for a soil salinity study in the Red River Valley that was to begin in June, 2006. We saw that there was an opportunity to assist in this study by providing PixelLearn's data analysis capabilities, namely the ability to read in very large data sets, the ability to display multiple overlays, and the ability to perform exploratory analysis such as clustering and classification of the data. We requested and received $75K from the ATI program to take the PixelLearn application one step further to TRL 7 by delivering it to the USDA scientists and tailoring it to their needs. Our primary goals in this effort have been to support the data sets used by the USDA scientists, add regression capabilities into PixelLearn, ultimately to provide a tool that they will find useful in their ongoing soil salinity studies. Overall, this project has been very successful, with multiple deliveries of PixelLearn to the USDA and the potential for additional follow-on work in the future.

Accomplishments

1. ***Data Type Support.*** We have extended the PixelLearn system to support additional data formats needed by the USDA. Specifically, PixelLearn can now load and analyze MODIS type 9 and type 15 data, and it automatically generates an NDVI (Normalized Difference Vegetation Index) view of the data when Red and IR bands are available. We have also added support for loading data stored in HDF files when running on a Windows platform.

2. ***Regression Algorithms.*** PixelLearn previously provided classification and clustering algorithms for analyzing input data sets. We added two regression algorithms (linear regression and Support Vector Regression) that permit PixelLearn to build models that can



predict real-valued quantities, such as temperature or rainfall, from remote sensing data. PixelLearn also permits the user to load and save trained regression models, an important capability when working with a variety of data sets or sharing models with collaborators. These new capabilities are shown in the screenshot above. There are 7 pixels associated with real-valued targets, these pixels are marked in the left panel but too small to see individually. In this example, higher (brighter) values correspond to vegetated areas. The trained SVM Regression model assigns real-valued outputs to the rest of the scene, as shown on the right. Consistent with the general PixelLearn philosophy, the user can experiment with different parameter settings and instantly observe the impact of those settings in the visual output. The USDA soil scientists plan to use this capability for estimating soil salinity values based on a collection of discrete field samples.

3. ***PixelLearn Infusion Success.*** We provided three releases of PixelLearn, of increasing capability, to the USDA. The final release was provided on August 24, 2007 and also included the first PixelLearn User's Guide. The PixelLearn TRL has advanced from 6 to 7. Perhaps most tellingly, Dr. Corwin of the USDA has chosen to request additional funds in a follow-on NRCS (National Resources Conservation Service) proposal to support additional development of PixelLearn for his work. He wrote, "*I feel that PixelLearn can be a useful tool for NRCS and I want to continue work with you and your staff in modifying PixelLearn, which means getting you support funds from NRCS.*" He plans to include a request for $50K to support PixelLearn, out of a total $250K budget for his project.

4. ***Publications.*** The abstracts of these papers are appended below.
    a. We submitted a journal paper in May describing work previously accomplished under the HARVIST project: "Progressive Refinement for Support Vector Machines," by Kiri L. Wagstaff, Michael Kocurek, and Dominic Mazzoni, to the *Data Mining and Knowledge Discovery* journal.
    b. We had a paper accepted describing a method for identifying which MODIS pixels correspond to which crop (their "salience" with respect to that crop) and presented it at the 2007 ICML Workshop on Constrained Optimization and Structured Output Spaces in June, 2007. The paper is "Salience Assignment for Multiple-Instance Regression," by Kiri L. Wagstaff and Terran Lane.
    c. We have submitted an abstract in September describing our advances in crop yield prediction from remote sensing data to the 2007 Fall Meeting of the American Geophysical Union: "County-Level Crop Yield Prediction Using Remote Sensing Data," by Kiri L. Wagstaff, Alex Roper, and Terran Lane.

5. ***Student involvement.*** Alex Roper, a rising junior at the California Institute of Technology, participated in this project through a 10-week Summer Undergraduate Research Fellowship (SURF) internship. He contributed significantly to the new PixelLearn capabilities and conducted several experiments that provided improved crop yield prediction results, as described in the AGU abstract.

Paper Abstracts

Progressive Refinement for Support Vector Machines
by Kiri L. Wagstaff, Michael Kocurek, and Dominic Mazzoni
submitted to the *Data Mining and Knowledge Discovery* journal

Support vector machines (SVMs) have good accuracy and generalization properties, but they tend to be slow to classify new examples. In contrast to previous work that aims to reduce the time required to fully classify all examples, we present a method that provides the best-possible classification given a specific amount of computational time. We construct two SVMs: a "full" SVM that is optimized for high accuracy and a reduced-set SVM that provides extremely fast, but less accurate, classifications. We apply the reduced SVM to the data set, estimate the posterior probability that each classification is correct, and then use the full SVM to reclassify items in order of their likelihood of misclassification. Our experimental results show that this method rapidly achieves high accuracy, by selectively devoting resources (reclassification) only where needed. It also provides the first such progressive SVM solution that can be applied to multiclass problems.

Salience Assignment for Multiple Instance Regression
by Kiri L. Wagstaff and Terran Lane
ICML 2007 Workshop on Constrained Optimization and Structured Output Spaces

We present a Multiple-Instance Learning (MIL) algorithm for determining the salience of each item in each bag with respect to the bag's real-valued label. We use an alternating-projections constrained optimization approach to simultaneously learn a regression model and estimate all salience values. We evaluate this algorithm on a significant real-world problem, crop yield modeling, and demonstrate that it provides more extensive, intuitive, and stable salience models than Primary-Instance Regression, which selects a single relevant item from each bag.

County-Level Crop Yield Prediction Using Remote Sensing Data
by Kiri L. Wagstaff, Alex Roper, and Terran Lane;
submitted to the 2007 Fall Meeting of the American Geophysical Union.

Early estimates of crop yield, particularly at a fine scale, can inform precision agriculture efforts. The USDA National Agricultural Statistics Service (NASS) currently provides estimates of yield on a monthly basis for each state. These estimates are based on phone interviews with farmers and in-situ examination of randomly selected plots. We seek to provide predictions at a much higher spatial resolution, on a more frequent basis, using remote sensing observations. We use publicly available data from the MODIS (Moderate Resolution Imaging Spectroradiometer) instruments on the Aqua and Terra spacecraft. These observations have a spatial resolution of 250 m and consist of two spectral bands (red and infra-red) with a repeat period of 8 days.

As part of the HARVIST (Heterogeneous Agricultural Research Via Interactive, Scalable Technology) project, we have created statistical crop yield models using historical MODIS data combined with the per-county yield reported by the USDA at the end of the growing season. In our approach, we analyze 100 randomly selected historical pixels from each county to generate a yield prediction for the county as a whole. We construct a time series for each pixel that consists of its NDVI (Normalized Difference Vegetation Index) value observed during each 8-day time period to date. We then cluster all pixels together to identify groups of distinct elements different crops, bodies of water, urban areas, desert, etc.) and create a regression model for each one. For each crop of interest, the model that best predicts that crop's historical yield is selected. These models can then be applied to data from subsequent years to generate predictions for the future.

We applied this approach to data from California and Kansas for corn and wheat. We found that, in general, the yield prediction error decreased as the harvest time approached. In California, distinctly different models were selected to predict corn and wheat, permitting specialization for each crop type. The best models from 2001 predicted yield for 2002 with a 10% (corn) and 23% (wheat) relative error three months before harvest. In Kansas, the 2001 models for corn and wheat were not well distinguished, providing good predictions for wheat (19% error three months before harvest) but poor predictions for corn (55% error three months before harvest). In post-analysis, we found that the 2001 pixel NDVI time series for Kansas are much more homogeneous than those for California, making it difficult to select crop-specific models. We are currently working on incorporating historical data from additional years, which will provide more diversity and potentially better predictions. We are also in the process of applying this technique to additional crops.