process at any point, thereby obtaining the best possible result for a given amount of computation time. Alternatively, the results can be displayed as they are generated, providing the user with real-time feedback about the current accuracy of classification.

Computational savings are realized through the guided application of resources only to those items that are estimated to be misclassified. The coarse approximation may suffice for items that can be classified easily, and more computation can be devoted to ambiguous or difficult cases. Thus, the algorithm enables the user to exert direct, dynamic control over the balance between classification speed and accuracy. When constraints on computation time and other resources preclude a totally accurate classification of all the data, this algorithm provides the best possible approximation to the classification of each item, rather than fully classifying only a fraction of the data set and leaving the rest marked "unknown."

This work was done by Kiri Wagstaff and Michael Kocurek of Caltech for NASA's Jet Propulsion Laboratory. Further information is contained in a TSP (see page 1).

The software used in this innovation is available for commercial licensing. Please contact Karina Edmonds of the California Institute of Technology at (626) 395-2322. Refer to NPO-44089.

Active Learning With Irrelevant Examples

Classification algorithms can be trained to recognize and reject irrelevant data.

NASA's Jet Propulsion Laboratory, Pasadena, California

An improved active learning method has been devised for training data classifiers. One example of a data classifier is the algorithm used by the United States Postal Service since the 1960s to recognize scans of handwritten digits for processing zip codes. Active learning algorithms enable rapid training with minimal investment of time on the part of human experts to provide training examples consisting of correctly classified (labeled) input data. They function by identifying which examples would be most profitable for a human expert to label. The goal is to maximize classifier accuracy while minimizing the number of examples the expert must label.

Although there are several well-established methods for active learning, they may not operate well when irrelevant examples are present in the data set. That is, they may select an item for labeling that the expert simply cannot assign to any of the valid classes. In the context of classifying handwritten digits, the irrelevant items may include stray marks, smudges, and mis-scans. Querying the expert about these items results in wasted time or erroneous labels, if the expert is forced to assign the item to one of the valid classes.

In contrast, the new algorithm provides a specific mechanism for avoiding querying the irrelevant items. This algorithm has two components: an active learner (which could be a conventional active learning algorithm) and a relevance classifier. The combination of these components yields a method, denoted Relevance Bias, that enables the active learner to avoid querying irrelevant data so as to increase its learning rate and efficiency when irrelevant items are present.

The algorithm collects irrelevant data in a set of rejected examples, then trains the relevance classifier to distinguish between labeled (relevant) training examples and the rejected ones. The active learner combines its ranking of the items with the probability that they are relevant to yield a final decision about which item to present to the expert for labeling. Experiments on several data sets have demonstrated that the Relevance Bias approach significantly decreases the number of irrelevant items queried and also accelerates learning speed.

This work was done by Kiri Wagstaff of Caltech and Dominic Mazzoni of Google, Inc. for NASA's Jet Propulsion Laboratory. For more information, contact iaoffice@jpl.nasa.gov. NPO-44094

A Data Matrix Method for Improving the Quantification of Element Percentages of SEM/EDX Analysis

John F. Kennedy Space Center, Florida

A simple 2D $M \times N$ matrix involving sample preparation enables the microanalyst to peer below the noise floor of element percentages reported by the SEM/EDX (scanning electron microscopy/energy dispersive x-ray) analysis, thus yielding more meaningful data.

Using the example of a 2×3 sample set, there are $M = 2$ concentration levels of the original mix under test: 10 percent ilmenite (90 percent silica) and 20 percent ilmenite (80 percent silica). For each of these M samples, $N = 3$ separate SEM/EDX samples were drawn. In this test, ilmenite is the element of interest. By plotting the linear trend of the M sample's known concentration versus the average of the N samples, a much higher resolution of elemental analysis can be performed. The resulting trend also shows how the noise is affecting the data, and at what point (of smaller concentrations) is it impractical to try to extract any further useful data.

This work was done by John Lane of Kennedy Space Center. For further information, contact the Kennedy Innovative Partnerships Program Office at (321) 861-7158. KSC-13303