Dr. THOMAS DIETTERICH

Project Title: Content-Based Image Retrieval of Herbarium Samples

Activities:
We have photographed plant samples from the OSU Herbarium and collected and scanned samples of the corresponding species from trees in Oregon. We have implemented and experimented with algorithms for matching shapes. First, we studied dynamic time warping (edit distance) algorithms for comparing the shapes of two isolated leaves. Then we considered matching isolated leaves against images in which leaves overlap or are incomplete. We imported gap-penalty dynamic programming algorithms from computational biology and implemented and tuned them. We also conducted experiments with using edit distances as approximate kernels in support vector machines. We also studied ways of combining isolated leaf data with data from Herbarium samples to obtain higher classification accuracy.

Findings:
Dynamic programming algorithms give excellent performance for shape matching. Using a 9-nearest neighbor classification algorithm, we can correctly classify isolated or overlapping leaves for four species of plants with error rates of less than five percent. We found that edit distance, with appropriate transformation and tuning, work well as kernels. We developed algorithms and methodology for exploiting auxiliary data sources (in this case, Herbarium samples) to improve classification accuracy in sparse-data situations.

Contributions to the discipline:
We have shown how to extend DP warping to account for occlusion in image matching. We have developed a new approach to solving learning problems where the training and test data have different properties. Our approach relies on using a small amount of training data that is representative of the test data and combining it with the much larger amount of non-representative training data. This problem arises in many applications, and we believe our solution will work in general.

Contributions beyond the discipline:
We have begun to show how to perform content-based image retrieval from plant image databases. This has potential applications in checking for correctness, consistency, and duplication within and between plant image databases. It also suggests a way that members of the public could look up the genus and species of unknown plants. Similar techniques should enable automated census of beetles and bees, with application to studies of forest and agricultural ecology and environmental monitoring.

Project Title: Machine Learning for Sequential and Spatial Data

Activities:
We focused our work on a new method developed by researchers at CMU and Whizbang Labs known as the Conditional Random Field, which was published in the summer of 2001. We developed a gradient descent method for training Conditional Random Fields. We implemented this method, first in Matlab and then in C. Because it is slow, we then created an optimized, parallelized implementation that runs via MPI on our cluster of linux boxes. Unfortunately, this was still too slow, so we developed a new method that extends Friedman’s gradient tree boosting idea to the sequential data setting. This gave us a speedup of approximately 10,000 times.

We have implemented and tested a general recurrent sliding window classifier framework for the WEKA machine learning package, which we will begin distributing later this year.
Findings:
We found that the CRFs are far too slow to provide the basis for an off-the-shelf algorithm for sequential supervised learning. Each iteration of gradient descent was requiring more than 50 hours of CPU time even with 16-fold parallel evaluation of the gradient. Gradient boosting looks like it will be fast enough to make CRFs a practical methods for sequential learning.

Contributions to the discipline:
We have shown how to extend Friedman's gradient tree boosting idea to the sequential data problem. This has highlighted several new research issues.

Project Title: Sub-pixel classification of remote sensed images

Activities:
We developed, implemented, and tested a new regression-tree algorithm for sub-pixel land-cover abundance estimation. We also implemented improved methods for linear unmixing (a popular algorithm in remote sensing) and neural networks.

Findings:
Experiments on four simulated and real remote sensing data sets showed that the bagged regression tree algorithm is superior to the neural network and linear unmixing methods.

Contributions to the discipline:
We have introduced the machine learning community to a new kind of classification problem and demonstrated that it cannot be solved by reducing it to standard supervised classification. We have developed an elegant yet effective algorithm, based on regression trees, for solving the problem.

Contributions beyond the discipline:
We hope to convince remote sensing scientists to employ our improved algorithm in their work on computing planet-wide continuous fields for land cover abundances.

Project Title: Bias-Variance Analysis of Support Vector Machine Classifiers

Activities:
We conducted a large experimental study to measure the bias and variance of SVMs while varying the parameters of the SVM (the C parameter) and parameters of the gaussian and polynomial kernels. We measured bias and variance both in synthetic problems and on some benchmark data sets from the Irvine repository of machine learning databases.

Findings:
We found that SVMs exhibit significant bias when the kernel parameter sigma is too small. This was surprising, because one would expect that small sigma would lead to low bias and high variance. However, the SVM ends up identifying relatively few support vectors and these fit the training data well but give essentially zero responses on the test data (and hence, high bias). We also found that SVMs exhibit significant variance for small samples, despite the tuning of the C parameter. We demonstrated that SVMs can be tuned so that when combined with bagging, they give better results than a single well-tuned SVM.

Contributions to the discipline:
We have provided further insight into the behavior of SVMs. We have also shown that they are quite robust to mistuning of the parameters, which is a very reassuring result.
Project Title: Supervised Reinforcement Learning

Activities:
Developed, tested, and evaluated three new algorithms for reinforcement learning: (a) a method based on regression-tree approximations of the value function, (b) a method based on support-vector methodology that constructs a scoring function from which a good policy can be computed, and (c) a method based on performing gradient ascent search to improve the performance of a parameterized policy.

Findings:
The regression tree method works very well for deterministic and nearly-deterministic problems such as arise in combinatorial search and deterministic control. The support-vector method also works very well and is able to generalize well to new problems. The model-based policy gradient method is the best of these three approaches. We proved that model-based policy gradient computed from a small subspace of the entire state space converges (under certain technical conditions) to a policy that is approximately a local optimum in the policy space.

Contributions (within the discipline):
There are two factors limiting the widespread application of reinforcement learning. First, the only algorithms whose performance is well-understood do not scale up to large problems. Second, the only algorithms which scale to large problems are unreliable and often behave poorly. Our research makes incremental contributions toward algorithms that will scale to large problems while also having proofs of convergence and stability.

Project Title: Automatic Extraction of Label Data from Herbarium Samples

Activities:
Benchmarked commercial OCR packages for reading the typewritten and printed labels of Herbarium samples. Obtained initial results reading labels for samples of 22 fern species.

Findings:
The Omnipage 8 system gave the best results so far. Attempts to preprocess the image to improve performance generally hurt performance of Omnipage 8 and did not help it.