Office of Science
U.S. DEPARTMENT OF ENERGY

Advanced Scientific Computing Research
Applied Mathematics
FY 2007 Accomplishment

"Training Support Vector Machines for Data Classification"

Dianne P. O'Leary, Andre L. Tits, Jin Hyuk Jung
University of Maryland
College Park, Maryland

Summary

Data classification is a key problem in science and engineering. For example, given the results of laboratory screenings on a set of patients for whom a diagnosis is known due to more invasive tests, we might want to develop a way to diagnose future patients without the results of the invasive tests. Similarly, we might want to evaluate a nuclear weapon stockpile using tests as simple as possible. One method for performing such data classification is a support vector machine, a mathematically determined separator between two groups of individuals. We have developed a faster way to perform the computation of the separator, thus making it practical to use more samples and define a more precise separator.

A support vector machine (SVM) is defined by a hyperplane whose dimension is determined by the number of test results available. This number is typically quite small, but the number of individuals tested can be quite large. We have a dilemma: more individuals can increase the quality of the hyperplane but also the cost to determine it.

Mathematically, the problem of determining the hyperplane is a quadratic programming problem with many more constraints than variables. Most of these constraints (individuals) are not relevant: the hyperplane would be the same whether we knew the test results for them or not. We have implemented a constraint reduction method for these problems, based on adaptive identification of individuals that do determine the hyperplane. This yields substantial savings in time for later iterations of the solution algorithm.

In Figure 1 we show timings comparing three variants of our algorithm with the standard IPM on four test problems obtained from Joshua Griffin. The variants differ in how they decide which data points to include; all of these variants, however, show substantial improvement over the standard algorithm. Figure 2 illustrates how the savings is achieved, using a toy problem with a separating ellipse for ease of visualization. At the beginning, all of the data points are used but as the algorithm proceeds, fewer and fewer of the points are used in the computation, until finally the important points are identified and the algorithm can complete.

In collaboration with Juan Mesa (Lawrence Berkeley National Laboratory) on a project for the 2007 Mathematical Sciences Research Institute Undergraduate Program, we are applying the algorithm to the problem of detecting supernovae in astronomical images.

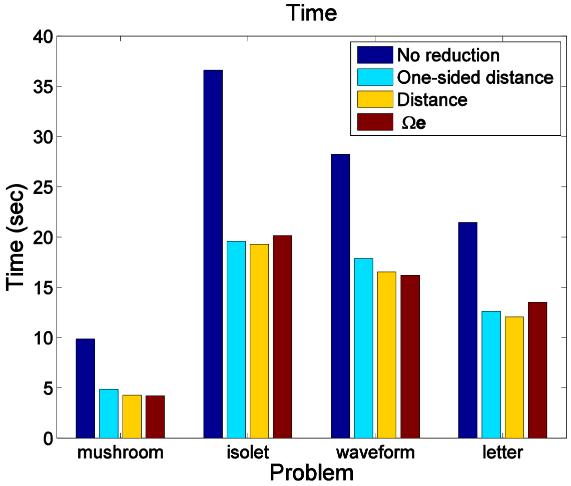

Figure 1. Time for a standard IPM (No reduction) and for three variants of the adaptive reduction algorithm on four test problems: mushroom, isolet, waveform, and letter.

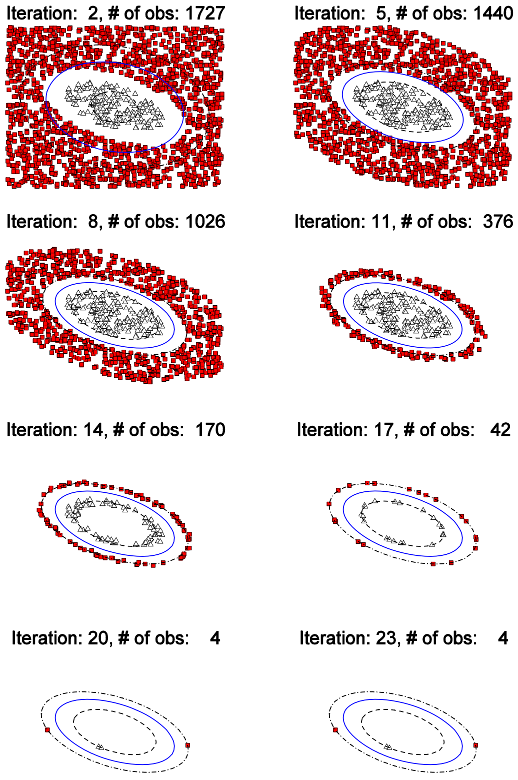

Figure 2 Snapshots of iterations of the adaptive IPM for finding a classifier for a toy problem in 2-dimensional input space. The data points that contribute to the computation are indicated by red boxes and gray triangles.

For further information on this subject contact:

Dr. Dianne P. O'Leary
University of Maryland
oleary@cs.umd.edu
Phone: 301-405-2678

Or
Dr. Anil Deane
Applied Mathematics Research Program
Office of Advanced Scientific Computing
Phone: 301-903-1465
deane@ascr.doe.gov