BILEVEL PROGRAMMING ALGORITHMS FOR MACHINE LEARNING MODEL SELECTION

By

Gregory M. Moore

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: MATHEMATICS

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Examining Committee:

Kristin P. Bennett, Thesis Adviser

Sanmay Das, Member

Joseph G. Ecker, Member

John E. Mitchell, Member

Rensselaer Polytechnic Institute
Troy, New York

September 2010
(For Graduation December 2010)

ABSTRACT

This thesis represents an advance towards the goal of self-tuning supervised data mining as well as a significant innovation in scalable bilevel programming algorithms. Novel nonsmooth bilevel programming methods for training linear learning models with hyperparameters optimized via T-fold cross-validation (CV) are considered. Determining the optimal values of these hyperparameters is the goal of *model selection*. Three algorithms are developed and tested against previous algorithms: `ImpGrad`, `ImpBundle` and `PBP`. Also a new generalized version of ϵ-insensitive regression is developed. This new modeling task, called multiSVR, allows use of data sets with varying levels of quality. Two data sets are used to test this new 10-hyperparameter model.

Current model selection practice constructs models over a predefined grid of hyperparameter combinations and selects the best one, an inefficient heuristic algorithm. The proposed methods formulate the model selection CV problem as a bilevel problem. A bilevel problem is an optimization problem with constraints that are also an optimization program. The overall objective is referred to as the outer-level objective, and the inner optimization constraints are referred to as the inner-level problem. The bilevel framework simultaneously solves for the model hyperparameters (outer-level) and training weights (inner-level). The new methods address hyperparameter selection for linear support vector machines (SVM) and related machine learning problems.

The proposed approaches are the first to address the hyperparameter selection of SVM models expressed as unconstrained (potentially nonsmooth) optimization problems. Applying nonsmooth mathematical programming methods to nonsmooth linear SVM models has produced the fastest training algorithms to date. The bilevel programming approaches to hyperparameter optimization can produce two types of algorithms: explicit and implicit. Explicit algorithms simultaneously optimize the model weights and hyperparameters by replacing the inner-level problems by their optimality conditions. The proposed explicit algorithm `PBP` utilizes the optimality

conditions in nonsmooth form, contrasting with prior less scalable approaches that formulate the problem as a mathematical program with equilibrium constraints (MPEC). Implicit algorithms optimize only the hyperparameters by treating the model weights as implicit functions of the hyperparameters. The implicit algorithms work in a low dimensional space at the expense of having a very challenging nonsmooth nonconvex objective function. The proposed implicit algorithms work with the primal unconstrained SVM problem in contrast with prior implicit methods that use the smooth dual SVM problem.

The first proposed explicit approach, a penalized bilevel program (`PBP`) algorithm, treats the lower-level problems as unconstrained optimization problems that are replaced with their optimality conditions. The key innovation is that the inner-level problem can be replaced with a nonsmooth nonlinear constraints without introducing extra variables or constraints. Previous methods only used the smoothed primal or dual formulations which results in nonlinear optimality conditions that grow exponentially in the sample size. A novel bilevel programming algorithm to solve this class of problems is developed by penalizing the nonsmooth optimality condition and then solving the resulting problem with an approximation algorithm. Convergence analysis of the algorithm shows `PBP` converges to a solution satisfying the necessary optimality conditions of the penalized bilevel program.

The `ImpGrad` and `ImpBundle` algorithms are implicit gradient-based approaches. Similar to `PBP`, these algorithms treat the lower-level problems as unconstrained optimization problems, unlike the previous implicit methods which use the dual training formulation. This yields a linear optimality condition from which the gradient of the training problem can be computed. Similar to the previous implicit gradient-based algorithm, `ImpGrad` assumes that the bilevel program is locally smooth, and selects an arbitrary subgradient. `ImpBundle` expands on `ImpGrad` by not assuming smoothness. This improved algorithm uses directional derivatives in the search direction of interest to ensure that the subgradient chosen is not only valid, but also informative. This prevents searching in assent directions.

A new outer algorithm is developed to solve the `ImpBundle` algorithm. Rather than use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method of `ImpGrad` and

previous implicit gradient-based approaches, this method uses a novel nonsmooth nonconvex bundle method that allows for bounds constraints. This algorithm explicitly deals with the nonsmoothness and nonconvexity of the model selection problem.

Computational results compare `PBP`, `ImpGrad` and `ImpBundle` against previous bilevel algorithms. Results on cheminformatics problems are presented for the traditional ϵ-insensitive regression problem and a new multiSVR model with 10 hyperparameters. Model selection of the later problem is beyond the ability of most algorithms to solve efficiently. In fact a thorough grid search on this problem would require roughly 250 million training problems to be solved. The proposed implicit and algorithms are shown to scale well in the sample size and number of hyperparameters. Experimental results show than the proposed algorithms outperform grid search and prior smooth bilevel CV methods in terms of modeling performance and computational time. PBP produces the best overall generalization results but with larger computational times that the implicit algorithms. This increased speed permits modeling with an increased number of hyperparameters on massive datasets. The implicit algorithms are faster and applicable to a larger class of problems than PBP. But `ImpBundle` and particularly `ImpGrad` yielded slightly less robust generalization results on a few problems.