Dynamics of Continuous, Discrete and Impulsive Systems
Series **B:** *Applications & Algorithms* **17 (2010)** 741-749
 http://www.watam.org

NONSMOOTH OPTIMIZATION ALGORITHM FOR SEMI-SUPERVISED DATA CLASSIFICATION

Burak Ordin

Department of Mathematics, Faculty of Science,
Ege University, Bornova, 35100, Izmir, Turkey
email: burak.ordin@ege.edu.tr

Abstract. In this paper, we develop a new algorithm for solving semi-supervised data classification problems. Given a training set of labeled data and a working set of unlabeled data, semi-supervised vector machine constructs a support vector machine using both the training and working sets. We formulate this problem as a nonsmooth optimization problem and then apply the quasisecant method for its solution. We evaluate the new algorithm applying it to some test data sets and report the results of numerical experiments.

Keywords. Nonsmooth optimization, nonconvex optimization, semi-supervised data classification, subdifferential.

AMS (MOS) subject classification: 65K05, 90C25.

1 Introduction

The supervised data classification is an important task in data mining. It has many applications in engineering, medicine, business etc. The purpose of supervised data classification is to establish rules for the classification of some observations assuming that the classes of data are known [1].

Support vector machines algorithms are known to be powerful data classification algorithms [14, 15]. They have been applied to solve large scale data classification problems such as the text categorization [11].

In the semi-supervised classification, only partial information is available about the data labels. In particular, referring to the training set as the set of the labeled objects and the testing set as the set of the unlabeled objects, the basic idea is to construct the classifier on the basis of the information coming from both of them. Namely semi-supervised classification is a compromise between supervised and unsupervised classification, whose objective is to take advantage from both of them [1, 2, 12, 13, 16].

Bennett and Demiriz [7] formulate the semi-supervised support vector machine as a mixed integer programming problem. Their formulation requires the introduction of a binary variable for each unlabeled data point in the training set. This makes the problem difficult to solve for large unlabeled data.