Semi-supervised Learning for Regression with Co-training by Committee

Mohamed Farouk Abdel Hady*, Friedhelm Schwenker, and Günther Palm

Institute of Neural Information Processing
University of Ulm
D-89069 Ulm, Germany
{mohamed.abdel-hady,friedhelm.schwenker,guenther.palm}@uni-ulm.de

Abstract. Semi-supervised learning is a paradigm that exploits the unlabeled data in addition to the labeled data to improve the generalization error of a supervised learning algorithm. Although in real-world applications regression is as important as classification, most of the research in semi-supervised learning concentrates on classification. In particular, although Co-Training is a popular semi-supervised learning algorithm, there is not much work to develop new Co-Training style algorithms for semi-supervised regression. In this paper, a semi-supervised regression framework, denoted by *CoBCReg* is proposed, in which an ensemble of diverse regressors is used for semi-supervised learning that requires neither redundant independent views nor different base learning algorithms. Experimental results show that *CoBCReg* can effectively exploit unlabeled data to improve the regression estimates.

1 Introduction

Many real-world data mining applications have a large amount of unlabeled data but labeling data is often difficult, expensive, or time consuming, as it requires the effort of human experts for annotation. *Semi-supervised learning* (SSL) refers to methods that exploits the unlabeled data in addition to the labeled data to improve the generalization error of a supervised learning algorithm. Readers interested in recent advances of SSL are directed to [1].

Co-Training is a popular SSL paradigm introduced by Blum and Mitchell [2] where two classifiers are trained iteratively on two sufficient and independent views. That is, two sets of features that are conditionally independent given the class and each of which is sufficient for learning. At the initial iteration, two classifiers are trained using the available labeled training examples. Then at each further iteration, each classifier labels and selects some unlabeled examples

* This paper is based on work done within the Transregional Collaborative Research Centre SFB/TRR 62 *Companion-Technology for Cognitive Technical Systems* funded by the German Research Foundation (DFG). The first author was supported by a scholarship of the German Academic Exchange Service (DAAD) and a travel grant from the European Neural Network Society (ENNS).

C. Alippi et al. (Eds.): ICANN 2009, Part I, LNCS 5768, pp. 121–130, 2009.

to augment the training set of the other. The aim is that one classifier can improve the accuracy of the other by providing it with informative examples. Although, multi-view *Co-Training* is applicable on certain domains, its multi-view requirement is impractical in many real-world applications. Goldman and Zhou [3] presented a single-view SSL method, called *Statistical Co-learning*. Two different supervised learning algorithms are used to partition the input space into a set of equivalence classes and k-fold cross validation is applied:(1) to select the most confident examples to label at each iteration and (2) to combine the two hypotheses producing the final decision. Zhou and Li [4] present a new *Co-Training* style SSL method called *Tri-Training*. An initial ensemble of three classifiers is trained by data sets generated via bootstrap sampling from the original labeled training set [5]. These classifiers are then refined during the *Tri-Training* process, and the final hypothesis is produced via majority voting.

Although the success of the above SSL approaches for classification, there is not much work on SSL for regression. Zhou et al. [6] proposed a *Co-Training* style semi-supervised regression algorithm called *CoReg*. This algorithm employs two diverse k-Nearest Neighbor (kNN) regressors that were instantiated using two different values of the Minkowski distance order. The *labeling confidence* is estimated such that the most confidently labeled example is the one which keeps the regressor most consistent with the existing labeled training set.

Our main contributions are: (1) A new single-view committee-based semi-supervised regression algorithm, called *CoBCReg* that extends the standard *Co-Training* algorithm. It is based on an ensemble of *RBF network* regressors constructed by *Bagging* [5]. (2) A new *Gaussian basis function* that is based on Minkowski distance instead of Euclidean distance. For the effectiveness of *CoBCReg*, there must be some diversity among the committee members and *CoBCReg* should maintain this diversity during the SSL process. This is achieved not only by training regressors using different training subsets but also through using different distance measures and different random initialization of the regressors parameters. The applicability of the proposed algorithm is broader than standard *Co-Training* algorithm because it does not require multiple redundant and independent views.

2 Co-training by Committee for Regression (*CoBCReg*)

There are two potential problems that can prevent any *Co-Training* style algorithm from exploiting the unlabeled data to improve the performance and these problems are the motivations for this study. Firstly the outputs of unlabeled examples are incorrectly estimated by a regressor that leads to adding noisy examples to the training set of the other regressor. Secondly there is no guarantee that the newly-predicted examples selected by a regressor as *most confident examples* will be *informative examples* for the other regressor. In order to mitigate the former problem, a committee of predictors is used in *CoBCReg* to predict the unlabeled examples instead of a single predictor. For the latter problem, each regressor selects the most informative examples for itself.