Chapter 12
An Optimized Semi-Supervised Learning Approach for High Dimensional Datasets

Nesma Settouti
Tlemcen University, Algeria

Mohammed El Amine Bechar
Tlemcen University, Algeria

Mostafa El Habib Daho
Tlemcen University, Algeria

Mohammed Amine Chikh
Tlemcen University, Algeria

ABSTRACT

The semi-supervised learning is one of the most interesting fields for research developments in the machine learning domain beyond the scope of supervised learning from data. Medical diagnostic process works mostly in supervised mode, but in reality, we are in the presence of a large amount of unlabeled samples and a small set of labeled examples characterized by thousands of features. This problem is known under the term "the curse of dimensionality". In this study, we propose, as solution, a new approach in semi-supervised learning that we would call Optim Co-forest. The Optim Co-forest algorithm combines the re-sampling data approach (Bagging Breiman, 1996) with two selection strategies. The first one involves selecting random subset of parameters to construct the ensemble of classifiers following the principle of Co-forest (Li & Zhou, 2007). The second strategy is an extension of the importance measure of Random Forest (RF; Breiman, 2001). Experiments on high dimensional datasets confirm the power of the adopted selection strategies in the scalability of our method.

INTRODUCTION

One of the strongest problems afflicting current machine learning techniques is dataset dimensionality. Nowadays, with the advance of technologies, in many applications of real world problems, we deal with data from a few dozen to many thousands of dimensions. The analysis of higher dimensional datasets is difficult, not only because they are large in terms of the number of observations, but also because of the large number of variables (features) that can be generated with the modern automatic acquisition methods. In fact, most applications allow to obtain many features and samples at low cost. However,

DOI: 10.4018/978-1-5225-2607-0.ch012

the relevant features are often more difficult to be obtain than the others. This is particularly true in the prediction problems.

In these application fields, the learning task is confronted with another important detail where, new samples are easily generated; nevertheless, labeling data can be costly and time consuming. For example, with the fast development of the Internet, it is easy to get billions of Web pages from Web servers. However, the classification of web pages into classes is a long and difficult task. Also in the field of speech recognition, registration gives a huge amount of audio data whose cost is negligible. However, labeling them requires someone to listen and understand later. Similar situations apply to remote sensing, face recognition, medical imaging, image search by content (Zhou and Goldman, 2004) and intrusion detection in computer networks (Roli, 2005).

The availability of unlabeled data and the difficulty of obtaining labels, make the semi-supervised learning methods gained great importance. The question that arises is whether the knowledge of points with labels is sufficient to construct a decision function that can correctly predict the labels of unlabeled points. Different approaches propose to deduct unlabeled points of additional information and include them in the learning problem.

Different kinds of approaches have been developed to achieve the semi-supervised learning task. There are mainly three paradigms (Chapelle, O. et al., 2006; Cornuéjols and Miclet, 2010) that address the problem of combination of labeled and unlabeled to improve the performances. Therefore, we include in brief these categories:

- **Semi-Supervised Learning (SSL):** Refers to methods that attempt to exploit unlabeled data for supervised learning where unlabeled examples are different from test examples; or exploiting labeled data for unsupervised learning.
- **The Transductive Learning:** Assemble methods that attempt also to exploit the unlabeled examples, but assuming unlabeled examples are exactly the test examples.
- **The Active Learning:** Refers to methods that select unlabeled examples that are the most important, and an oracle can be proposed for the labeling of these instances; the objective is to minimize the labeling data (Freund, Y. et al., 1997). Sometimes it is called selective sampling or sample selection.

In this paper, we focus on improving the performance of supervised classification using unlabeled data (SSL). In this context the two main contributions of this work are the treatment of the following questions: “How to judge the relevance of a model using unlabeled data? ” And ”How to improve the performance of the model ?”.

Many semi-supervised learning (SSL) algorithms have been proposed, among which the ”*Co-forest*” algorithms are widely used. We present in this work an optimized *Co-forest* algorithm. It uses a relevant random subspace method to form an initial ensemble of classifiers, where each classifier is trained with different relevant subspace of the original feature space. Unlike the prior work of (Li and Zhou, 2007) on *Co-forest*, our method uses a feature importance measure in semi-supervised learning by the ensemble of classifiers. Each classifier’s prediction on new unlabeled data with relevant features are combined and then used to enlarge the training set of others. The classifiers ensemble are refined through the enlarged training set. Experiments on high and small data sets show the ability and effectiveness of *Optim Co-forest* to select and measure importance to improve the performance of the ensemble of classifiers learned with a small amount of labeled samples by exploiting unlabeled samples. A comparative result