Expert Systems With Applications 51 (2016) 85–106

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Semi-supervised support vector regression based on self-training with label uncertainty: An application to virtual metrology in semiconductor manufacturing

Pilsung Kang[a], Dongil Kim[b,*], Sungzoon Cho[c]

[a] School of Industrial Management Engineering, Korea University, 02841 Seoul, South Korea
[b] Smart Manufacturing Technology Group, Korea Institute of Industrial Technology, 31056 Cheonan, South Korea
[c] Department of Industrial Engineering, Seoul National University, 08826 Seoul, South Korea

ARTICLE INFO

Keywords:
Semi-supervised learning
Support vector regression
Probabilistic local reconstruction
Data generation
Virtual metrology
Semiconductor manufacturing

ABSTRACT

Dataset size continues to increase and data are being collected from numerous applications. Because collecting labeled data is expensive and time consuming, the amount of unlabeled data is increasing. Semi-supervised learning (SSL) has been proposed to improve conventional supervised learning methods by training from both unlabeled and labeled data. In contrast to classification problems, the estimation of labels for unlabeled data presents added uncertainty for regression problems. In this paper, a semi-supervised support vector regression (SS-SVR) method based on self-training is proposed. The proposed method addresses the uncertainty of the estimated labels for unlabeled data. To measure labeling uncertainty, the label distribution of the unlabeled data is estimated with two probabilistic local reconstruction (PLR) models. Then, the training data are generated by oversampling from the unlabeled data and their estimated label distribution. The sampling rate is different based on uncertainty. Finally, expected margin-based pattern selection (EMPS) is employed to reduce training complexity. We verify the proposed method with 30 regression datasets and a real-world problem: virtual metrology (VM) in semiconductor manufacturing. The experiment results show that the proposed method improves the accuracy by 8% compared with conventional supervised SVR, and the training time for the proposed method is 20% shorter than that of the benchmark methods.

1. Introduction

Support vector regression (SVR), a regression version of support vector machines (SVM) (Vapnik, 1995), was proposed to solve nonlinear regression problems with a maximum margin algorithm (Smola & Schölkopf, 2002). SVR employs an ε-insensitive loss function (see Fig. 1); the training data whose margins are less than ε are not considered to be an error. Hence, an ε-sized insensitive tube (ε-insensitive tube or ε-tube) is constructed during SVR training. The data located on or outside the ε-tube are called support vectors, and the SVR regression function is formed as a linear combination of support vectors. SVR has the same advantages as those of SVM. SVR also maximizes the generalization performance by employing the structural risk minimization (SRM) principle with an ε-insensitive loss function, and it is capable of solving nonlinear problems with the kernel trick. With those advantages, SVR has been successfully applied to various areas: response modeling (Kim & Cho, 2012), virtual metrology (VM) (Kang, Kim, Lee, Doh, & Cho, 2011), finance prediction (Pai & Lin, 2005), time-series prediction (Thissen, van Brakel, de Weijer, Melssen, & Buydens, 2003), and environment application (Ortiz-García, Salcedo-Sanz, Pérez-Bellido, Portilla-Giqueras, & Prieto, 2010).

SVR was originally designed for application to the supervised learning problem. In supervised learning, the training dataset consists only of labeled data, $L = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{|L|}, y_{|L|})\}$, where $\mathbf{x}_l$, y_l, and $|L|$ represent the d-dimensional input variables, corresponding labels, and number of labeled data, respectively (see Eq. (1)). However, in some applications, because the labeled data are difficult, expensive, or time consuming to acquire, the amount of training data used is not sufficient for obtaining effective model performance. Conversely, the unlabeled data, $U = \{\mathbf{x}_{|L|+1}, \ldots, \mathbf{x}_{|L|+|U|}\}$, where $|U|$ is the number of unlabeled data, contains only the input variables (see Eq. (2)) and can be collected with less effort than the labeled data. Such unlabeled data are abundant in many applications, and hence the concept of training a model from those

* Corresponding author. Tel.: +82 31 8040 6774; fax: +82 31 8040 6170.
E-mail addresses: pilsung_kang@korea.ac.kr (P. Kang), dikim01@kitech.re.kr (D. Kim), zoon@snu.ac.kr (S. Cho).

http://dx.doi.org/10.1016/j.eswa.2015.12.027
0957-4174/

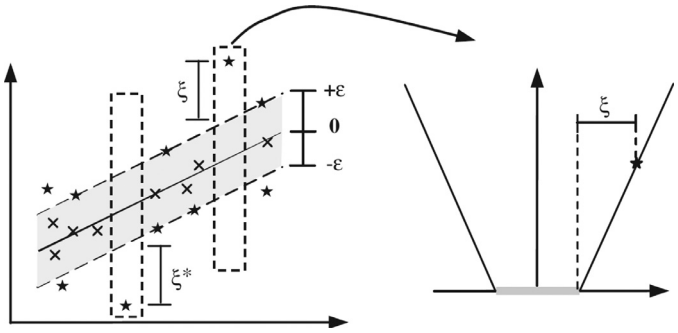

Fig. 1. ε -tube based on the margin of training data, and the ε-loss function of SVR (reprinted from Chen & Wang, 2007).

unlabeled and labeled data in order to improve model performance is proposed (Zhu, 2007).

$$L = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{|L|}, y_{|L|})\}, \quad \mathbf{x}_l \in \mathbb{R}^d, \quad y_l \in \mathbb{R}. \tag{1}$$

$$U = \{\mathbf{x}_{|L|+1}, \ldots, \mathbf{x}_{|L|+|U|}\}, \quad \mathbf{x}_u \in \mathbb{R}^d. \tag{2}$$

Semi-supervised learning (SSL) is one of the principal approaches for training a model from both labeled and unlabeled data. "SSL addresses this problem by using large amounts of unlabeled data, together with the labeled data, to build better classifiers" (Zhu, 2007). The research efforts for SSL can be categorized into five main directions: generative (Nigam & Ghani, 2000; Seeger, 2000), self-training (Mihalcea, 2004; Rosenberg, Hebert, & Schneideman, 2005), co-training (Blum & Mitchell, 1998; Mitchell, 1999), low density separation (Chapelle, Sindhwani, & Keerthi, 2008; Vapnik, 1998), and graph-based methods (Belkin, Niyogi, & Sindhwani, 2006; Sindhwani, Niyogi, & Belkin, 2005). Those methods have been applied widely to various real-world problems; however, most works for SSL were designed for classification problems. Because the label is a continuous real numbered variable for regression problems, most of these methods cannot be applied to regression problems directly (Cortes & Mohri, 2007; Pozdnoukhov & Bengio, 2006; Wang, Hua, Song, Dai, & Zhang, 2006).

Recently, SSL for regression has been proposed. SSL for regression is more complicated than SSL for classification. In order to utilize the unlabeled data for training, SSL for regression must estimate continuous-valued labels, explicitly or implicitly; only binary labels are required for SSL for classification. Co-training (Wang, Fu, , & Ma, 2011; Wang, Ma, & Wang, 2010; Zhou & Li, 2007), graph-based methods (Pozdnoukhov & Bengio, 2006), and kernel-based methods (Cortes & Mohri, 2007; Wang et al., 2006) have been proposed for SSL regression. However, these methods have limitations. The uncertainty of estimating labels for the unlabeled data is not considered. Because the labels for the unlabeled data should be estimated using mathematical models, a labeling uncertainty always exists. The estimated labels for the unlabeled data should be addressed differently based on their uncertainties. Moreover, because SVR is a margin-based method, only the estimated labels located on or outside the ε-tube influence the final SVR model. However, those methods, which employ a weighted average of nearest neighbors or a regression function trained by the labeled data, tend to estimate the labels of the unlabeled data inside the ε-tube. These do not improve the model accuracy. Finally, the time complexities of these methods are relatively high. Because co-training is a wrapper-based iterative approach, two base learners should be trained in each iteration. In addition, the graph-based methods need to calculate the entire kernel-based weight matrix for all data, including the unlabeled data.

In this paper, we propose a self-training based non-iterative semi-supervised support vector regression algorithm that estimates the label distribution of each unlabeled data point and oversamples based on the uncertainty of the labeling. The proposed method is designed to consider both accuracy and efficiency. The principal contribution of the proposed method can be summarized as follows:

(1) **The proposed method considers the uncertainty of the estimated labels of the unlabeled data.** In order to consider such labeling uncertainty, the proposed method estimates the label distribution (not a label value) of each unlabeled data point. Then, each unlabeled data point has an input value and corresponding Gaussian label distribution that consists of the mean and variance. The lower the variance of the estimated label distribution, the lower is the uncertainty of the unlabeled data point. Conversely, if the variance of the estimated label distribution is greater, the labeling uncertainty for the unlabeled data point is higher. Probabilistic local reconstruction (PLR) (Lee, Kang, & Cho, 2014), a local topology-based linear reconstruction method, is employed for estimating the label distribution. The proposed method employs two PLR models with different settings to capture and conjugate the local and global topology of the unlabeled data.

(2) **The proposed method generates data by oversampling from the unlabeled data and their estimated label distribution.** The proposed method randomly generates multiple training data from the unlabeled data and their estimated label distributions in order to increase the probability of the unlabeled data affecting the final SVR training. Moreover, the sampling rate is different based on the uncertainty of the estimation for each unlabeled data point. For those unlabeled data with low labeling uncertainty, only a few samples are generated from the estimated label distribution. Conversely, for those unlabeled data with high labeling uncertainty, more samples are generated in order to represent the entire estimated label distribution.

(3) **The training complexity of the proposed method is relatively low.** The proposed method employs a non-iterative algorithm. In addition, the proposed method does not need to construct a large-sized graph on the labeled and unlabeled data. Hence, the proposed method demands significantly less complexity than the iterative or graph-based methods. Moreover, to reduce the training data generated by oversampling, an additional training data selection method,expected margin-based pattern selection (EMPS) (Kim & Cho, 2012), is employed for training efficiency.

The performance of the algorithm is verified using sufficient benchmark regression datasets. Based on the experiments conducted on 30 regression datasets, it is verified that the proposed method improves accuracy by approximately 8% over the conventional supervised SVR. In addition, the training time of the proposed method, including the construction of the final SVR, is reduced to only 20% of the benchmark methods. We also demonstrate a real-world application of VM in semiconductor manufacturing. VM is a mathematical model that employs regression algorithms to estimate the quality of wafers in a manufacturing process, which has a few labeled data and abundant unlabeled data. The proposed method shows excellent performance for the semi-supervised VM problem.

The remainder of this paper is organized as follows. In Section 2, a review of the SSL literature is presented. In Section 3, the proposed method for SS-SVR is proposed. In Sections 4 and 5, the experiment results on benchmark datasets and a real-world semiconductor manufacturing problem are presented, respectively. Finally, Section 6 concludes this paper with a summary and limitations of the proposed method and future works are discussed.