Chemometrics and Intelligent Laboratory Systems 182 (2018) 47–56

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Illustration of merits of semi-supervised learning in regression analysis

Hiromasa Kaneko

Department of Applied Chemistry, School of Science and Technology, Meiji University, 1-1-1 Higashi-Mita, Tama-ku, Kawasaki, Kanagawa 214-8571, Japan

ARTICLE INFO

Keywords:
Regression
Semi-supervised learning
Stability
Predictive performance
Applicability domains
QSPR
QSAR

ABSTRACT

Semi-supervised learning (SSL) is a method for learning the relationship between X and y, and the essential structure of the corresponding dataset, using both labeled and unlabeled data. In this paper, an approach to use a combination of labeled and unlabeled samples to reduce the dimension, then perform regression analysis using the labeled samples in a low-dimensional space is focused in SSL methods. While various SSL methods for regression have been developed, there has been insufficient discussion as to why SSL is effective in regression analysis. Therefore, in this study, the merits of SSL in regression analysis are discussed in terms of the stability or the robustness and applicability domains of regression models and prior distribution of X-variables. The superiorities of SSL methods over fully supervised methods in regression are demonstrated using data from numerical simulations, quantitative structure–activity relationships and quantitative structure–property relationships.

1. Introduction

In the fields of data analysis, chemoinformatics and chemometrics, chemical structures are transformed into molecular descriptors and handled quantitatively and qualitatively. The relationships between molecular descriptors and their properties/activities (y) are analyzed in quantitative structure-activity relationships (QSARs) [1] and quantitative structure-property relationships (QSPRs) [2]. When y is continuous values, it is regression, and when y is class values, it is classification. By constructing statistical models that connect information on the chemical structures of compounds with their properties and activities, it is possible to estimate the property- and activity-values of new chemicals, without experiments, by inputting their structures into the models. This has applications in virtual screening [3] and chemical structure generation [4].

Supervised learning is a method to learn the relationships between explanatory variables (X) and objective variables (y). In a chemical context, quantitative information on chemical structures corresponds to X, and the properties and activities of the corresponding compounds or samples are the y-values. Supervised learning can be classified into two forms: classification, in which y consists of categories, and regression, in which y consists of continuous values. Linear classification methods include linear discriminant analysis [5] and support vector machine (SVM) [6], while nonlinear classification methods include the k-nearest neighbor algorithm [7], decision tree (DT) [8], random forest (RF) [9] and nonlinear SVM. Likewise, regression analysis includes both linear methods, such as principal component regression (PCR) [10], partial least squares (PLS) [11] and support vector regression (SVR) [6], and nonlinear methods, such as DT, RF and nonlinear SVR.

There also exist numerous compounds and virtual chemical structures without y-values, i.e., for which the properties and activities have not been measured. For these structures, only information on the X-variables is available. Given a dataset of chemical structures (samples) without y-values, the essential structure hidden behind the dataset must be learned by a process called unsupervised learning. This process is classified into two forms: dimensionality reduction (or data visualization), and clustering. Linear methods for dimensionality reduction include principal component analysis (PCA) [10] and factor analysis (FA) [12], and nonlinear ones include kernel PCA (KPCA) [13], self-organizing map (SOM) [14], generative topographic mapping (GTM) [15], locally linear embedding (LLE) [16] and t-distributed stochastic neighbor embedding (tSNE) [17]. Clustering methods include hierarchical clustering [18] and k-means clustering [19]. Currently, the number of unlabeled data, i.e., samples without y-values, is enormous compared with the number of labeled data, i.e., samples with y-values. Thus, unsupervised learning plays a key role in data analysis, alongside supervised learning.

Semi-supervised learning (SSL) [20] is a method for learning the relationship between X and y, and the essential structure of the corresponding dataset, using both labeled and unlabeled data. In SSL, trial predictive models for classification and regression are constructed using not only labeled but also unlabeled data. SSL can be especially useful when the number of labeled data is small and the number of unlabeled data is large.

Various SSL methods have been developed for classification, such as self-training [21], co-training [22], expectation-maximization

E-mail address: hkaneko@meiji.ac.jp.

https://doi.org/10.1016/j.chemolab.2018.08.015
Received 9 December 2017; Received in revised form 27 August 2018; Accepted 30 August 2018
Available online 4 September 2018

(EM)-based SSL [23], transductive learning [24], semi-supervised SVM [25], label propagation [26] and label spreading [27]. Although there is still room for discussion on SSL in the case of classification, the basic mechanism of SSL in classification is to identify unlabeled samples that are close to labeled samples in a given class, re-classify those unlabeled samples as belonging to the class in question, and then perform classification again. This produces a set of discriminant functions that describe the data distribution of the unlabeled samples.

However, SSL methods have also been developed for regression. One such approach is to use a combination of labeled and unlabeled samples to reduce the dimension, then perform regression analysis using the labeled samples in a low-dimensional space. Dimension reduction methods used thus far include PCA [28], probabilistic PCA [29] and Gaussian fields [30]. Nie et al. summarized a framework of unsupervised dimension reduction methods for SSL [31]. For generative linear regression models, Chakrabortty and Cai discussed the efficiency of regressors in SSL based on joint probability distribution of X, which is related to robustness of models and ancillarity [32,33]. However, while various methods have been proposed, there has been insufficient discussion of the effectiveness of SSL in both linear and nonlinear regression analyses, to the best of the author's knowledge, and it is not easy to label unlabeled data as shown in "2. Method" section.

Therefore, this study will discuss the merits of SSL in regression analysis. The discussion will focus on a method combining low-dimensional transform (in unsupervised learning) and regression (in supervised learning). Both unlabeled and labeled samples are used to reduce the dimension of a dataset, and regression analysis is then conducted using the labeled samples in a reduced-dimensional space. This method is called SLR, from **S**SL based on **l**ow-dimensional transform (in unsupervised learning) and **r**egression (in supervised learning). For generative linear regression models in SSL, please see reference [32]. The novelty of this paper is to expand SLR, which was only a combination of PCA and PLS, to any (nonlinear) dimensionality reduction such as PCA, FA, KPCA, SOM, GTM, LLE and tSNE and any (nonlinear) regression analysis methods such as PCR, PLS, SVR, DT and RF, and to clarify the merits of SLR and the reasons of predictive accuracy improvement using SLR, which were unclear.

This paper will discuss SLR from the viewpoints of the stability or the robustness of regression models, applicability domains (ADs) of regression models and prior distribution of X-variables. The merits of SLR will be discussed, such as its ability to stabilize regression models, enlarge the ADs of regression models and accurately obtain prior distributions of X-variables. To enlarge the ADs means that regression models estimate y-values for larger data domains in X-space. Moreover, the superiorities of SLR methods over supervised methods will be demonstrated using data from numerical simulations, QSAR and QSPR.

2. Method

2.1. Semi-supervised learning (SSL) in regression analysis

In this paper, SLR, which is an **S**SL method based on **l**ow-dimensional transform (in unsupervised learning) and **r**egression (in supervised learning) is targeted. Fig. 1 shows the basic concept of SLR. First, the dimensions of a dataset including both labeled and unlabeled samples are reduced using a dimensionality reduction method. The X-variables, the total number of which is m, are transformed to Z-variables, the total number of which is k ($m > k$). Then, regression analysis is performed between the Z-variables and y-variable using only the labeled samples, and a regression model is constructed. By inputting the values of Z into the regression model, the y-values can be estimated for the unlabeled samples.

Any dimensionality reduction method, such as PCA, FA, KPCA, SOM, GTM, LLE and tSNE, can be used. Likewise, any regression analysis method, such as PCR, PLS, (nonlinear) SVR, DT and RF, can be employed.

2.2. Discussion on merits of SSL in regression analysis

Unfortunately, it is difficult to obtain structural information on predictive models from unlabeled data both in regression and in classification. Fig. 2 shows the regression analysis of a dataset containing both labeled and unlabeled data. Given the labeled data, many candidate regression models are possible, such as those numbered 1, 2 and 3. However, no structural information on these regression models in X and y space can be obtained from the unlabeled data. In regression analysis, therefore, SSL does not offer the same benefit – i.e., that unlabeled data are transformed to labeled data based on the essential structure of the labeled and unlabeled data – as it does in classification.

Therefore, this study focuses on the merits of SSL in terms of the stability or the robustness of regression models, ADs of regression models and prior distribution of X-variables.

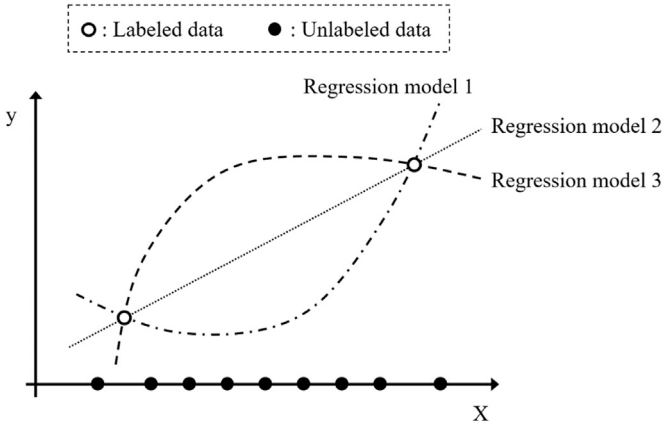

Fig. 2. Labeled data and unlabeled data in regression analysis. Given the labeled data (two empty circles), many candidate regression models are possible, such as those numbered 1, 2 and 3. However, no structural information on these regression models in X and y space is obtained from the unlabeled data (dots).

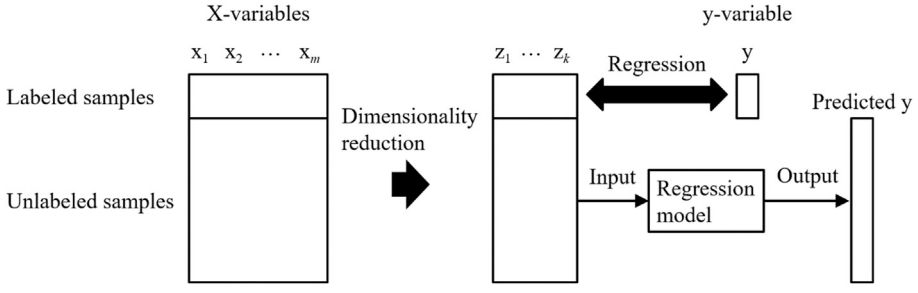

Fig. 1. Basic concept of semi-supervised learning based on a combination of unsupervised and supervised learning in regression analysis (SLR).