Chemometrics and Intelligent Laboratory Systems 177 (2018) 74–82

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Automatic outlier sample detection based on regression analysis and repeated ensemble learning

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
Outlier samples
Ensemble learning
Predictive performance
QSPR
QSAR

ABSTRACT

The fields of chemoinformatics and chemometrics require regression models with high prediction performance. To construct predictive regression models by appropriately detecting outlier samples, a new outlier detection and regression method based on ensemble learning is proposed. Multiple regression models are constructed and y-values are estimated based on ensemble learning. Outlier samples are then detected by comprehensively considering all regression models. Furthermore, it is possible to detect outlier samples robustly and independently by repeated calculations. By analyzing a numerical simulation dataset, two quantitative structure-activity relationship datasets and two quantitative structure-property relationship datasets, it is confirmed that automatic outlier sample detection can be achieved, informative compounds can be selected, and the estimation performance of regression models is improved.

1. Introduction

In the fields of chemoinformatics and chemometrics, regression models between molecular descriptors quantifying chemical structures can be constructed with respect to activities, giving quantitative structure-activity relationships (QSARs) [1], and with respect to properties, giving quantitative structure-property relationships (QSPRs) [2]. Using regression models, it is possible to estimate the values of activities and properties in compounds for which these quantities have not been measured [3]. Because the activities and properties can only be estimated from information on chemical structures, virtual screening of chemical structures [4] and chemical structure generation with the desired activities and properties [5] are also possible.

Ideally, the constructed regression models should have high prediction performance. Using predictive regression models, it is possible to estimate the values of activities and properties with low errors. The workflow of constructing QSAR models with high prediction performance has been described in a previous report [6].

Any deterioration in the prediction performance of regression models may be due to underfitting, overfitting [7], the existence of unnecessary variables [8], or the existence of outlier samples [9]. Although these factors are all important, this study focuses on the existence and proper detection of outlier samples. The prediction performance of regression models is expected to be improved if the regression models are constructed after excluding the detected outlier samples. Furthermore, as the relationship between structure descriptor X and property or activity y is different in outliers from that in other samples, useful information may be acquired by detecting outlier samples.

Outlier sample detection is not only for improving the prediction performance of regression models. Outlier samples themselves have much information. Although there is a possibility of measurement errors of y, outlier samples are a trigger to elucidate new relationships between property/activity and chemical structures.

One of the most famous outlier sample detection methods is the three-sigma method (TS) [10]. For an X- or y-variable, samples with values that are more than three times the standard deviation from the average are detected as outlier samples. Although outliers that are far from the data distribution can be detected, average values and standard deviations are largely affected by the outliers themselves. If the outlier values are high, both the average and the standard deviation will be increased by the outliers. Therefore, the Hampel filter (HF) [11], which uses the median instead of the average and the median absolute deviation instead of the standard deviation, was developed. By using the median and the median absolute deviation, it is possible to detect outliers robustly.

Whereas TS and HF detect outlier samples using a single variable, multivariate outlier sample detection methods such as the convex hull [12,13], robust principal component analysis [14,15], k-nearest neighbor algorithm [16], support vector data description [17], and one-class support vector machine (OCSVM) [18] detect outlier samples by considering multiple variables simultaneously. OCSVM applies a

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

https://doi.org/10.1016/j.chemolab.2018.04.015
Received 9 September 2017; Received in revised form 28 February 2018; Accepted 17 April 2018
Available online 19 April 2018

support vector machine to the data domain estimation problem, and can detect samples existing in low-data-density domains from relatively few training data.

In regression analysis, it is important to quantitatively analyze the relationships between X and y to decrease errors between actual y-values and y-values calculated by regression models. Although the relationship between X and y must be consistent in a dataset to construct predictive regression models, outlier samples have the different relationship from that of normal samples. However, TS, HF, and multivariate methods such as OCSVM do not take the relationships between X and y into consideration in outlier sample detection. Because regression models express the relationship between X and y, they are considered reasonable for detecting outlier samples in which the relationship between X and y is different from that of the other samples. An outlier sample detection method that considers the relationship between X and y may detect outlier samples in which the errors between the actual y-values and the y-values estimated through cross-validation [19] are high. These samples exhibit a different relationship between X and y from that of the other samples, and can thus be classified as outliers.

In cross-validation, however, if samples happen to have high y-errors, they are wrongly considered as outliers. On the contrary, when samples happen to have low y-errors, it is impossible to detect them as outliers. These issues are particularly evident when the number of training samples is low and the number of X-variables is high, in which case even linear regression models can be unstable, and when the y-errors happen to be high or are conversely easy to decrease. Deng et al. proposed an outlier sample detection method based on ensemble learning [20]. In sample bagging based on partial least-squares (PLS) [21], outliers having large average y-errors are defined as outliers of y and those having a large standard deviation in their y-errors are defined as outliers of X. However, the thresholds of the average and standard deviation were not shown to detect outliers, and it is unclear whether outlier samples can be determined automatically. Automatic outlier is impossible because there are no explicit rules in outlier detection. In addition, there remains a danger that normal samples are considered as outlier samples and outlier samples are not detected since outlier samples are included in a dataset in modeling; the average and the standard deviation are affected outliers; and regression models can be unstable.

This study proposes a robust and automatic outlier sample detection method based on ensemble learning and regression analysis. Regression models are repeatedly constructed by changing the training samples in ensemble learning, and samples having large errors between the center of the estimated values and the actual value are judged as outlier samples. Using the median rather than the mean as the center, and the median absolute deviation rather than the standard deviation, the influence of unstable local regression models on the detection of outlier samples is reduced. In addition, ensemble learning reduces the influence of y-errors that happen to be large or small in a local model construction and the noise of the estimated value.

After detecting outlier samples, ensemble learning is performed again with normal samples and the y-values are estimated for all samples. It is possible to prevent the estimated y-values from being influenced by outlier samples. Additionally, samples detected as outliers by chance can be changed to normal samples. This enables robust and automatic outlier sample detection and improved estimation performance to be achieved simultaneously.

To confirm the effectiveness of the proposed method, a set of numerical simulation data, two sets of QSPR data and two sets of QSAR data are used. Compared with other outlier sample detection methods, the proposed method classifies fewer samples as outliers, thus improving the overall estimation performance in regression analysis.

2. Method

First, TS, HF, OCSVM, and cross-validation error-based outlier sample detection (CVE) are introduced as general outlier sample detection methods. The proposed outlier sample detection method based on ensemble learning is then discussed in detail.

2.1. Three-sigma method (TS)

The TS method detects outliers whose absolute values exceed three times the standard deviation from the average for a single variable. When the data distribution of a variable follows the normal distribution, the probability that a value is within three times the standard deviation of the average is 99.73%. Values outside this interval can be regarded as outliers.

First, standardize a variable x as:

$$z^{(k)} = \frac{x^{(k)} - \mu}{\sigma}, \tag{1}$$

where $x^{(k)}$ is the value of the kth sample and $z^{(k)}$ is the value of the kth sample after normalization. μ is the average of x and is given as follows:

$$\mu = \frac{\sum_{k=1}^{n} x^{(k)}}{n}, \tag{2}$$

where n is the number of samples. σ is the standard deviation of x, given by:

$$\sigma = \sqrt{\frac{\sum_{k=1}^{n} (x^{(k)} - \mu)^2}{n-1}}. \tag{3}$$

Samples for which the absolute value of $z^{(k)}$ is greater than 3 are detected as outlier samples.

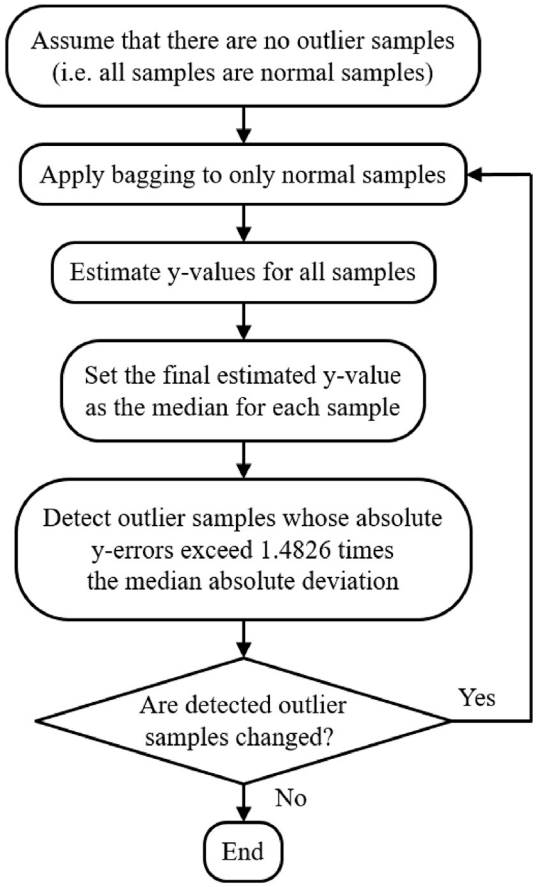

Fig. 1. Flow of the proposed ELO method.