Neurocomputing 235 (2017) 274–286

Contents lists available at ScienceDirect

Neurocomputing

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

Self-adaptive robust nonlinear regression for unknown noise via mixture of Gaussians

Haibo Wang[a], Yun Wang[b,*], Qinghua Hu[b]

[a] School of Economics and Management, Hubei University of Technology, Wuhan, Hubei, China
[b] School of Computer Science and Technology, Tianjin University, Tianjin, China

ARTICLE INFO

Communicated by Wei Chiang Hong

Keywords:
Self-adaptive nonlinear regression
Unknown noise
Mixture of Gaussians
Expectation maximization

ABSTRACT

For most regression problems, the optimal regression model can be obtained by minimizing a loss function, and the selection of loss functions has great effect on the performance of the derived regression model. Squared loss is widely used in regression. It is theoretically optimal for Gaussian noise. However, real data are usually polluted by complex and unknown noise, especially in the era of big data, the noise may not be fitted well by any single distribution. To address the above problem, two novel nonlinear regression models for single-task and multi-task problems are developed in this work, where the noise is fitted by Mixture of Gaussians. It was proved that any continuous distributions can be approximated by Mixture of Gaussians. To obtain the optimal parameters in the proposed models, an iterative algorithm based on Expectation Maximization is designed. The proposed models turn to be a self-adaptive robust nonlinear regression models. The experimental results on synthetic and real-world benchmark datasets show that the proposed models produce good performance compared with current regression algorithms and provide superior robustness.

1. Introduction

Regression, which is concerned to extract hidden rules from data, is an very old, but still a hot topic today [1]. The goal of regression is to predict the value of target variables given the value of a *D*-dimensional vector of independent variables [2]. Currently, regression analysis is widely used in various domains, such as gold returns [3], solar power output forecasting [4], face recognition [5] and so on. What attracts more attention is the performance of regression models in real complex conditions.

To develop regression algorithms, three important issues should be taken into consideration, namely model structures, objective functions and their corresponding optimization methods [1]. According to model structures, regression algorithms can generally be divided into two large categories: linear regression models and nonlinear regression models. As to objective functions, loss functions have great effect on the performance of regression models. The selection of loss functions is mostly dependent on the types of noises [6,7]. For example, squared loss is good for Gaussian noise, least absolute deviation loss for Laplace noise [8], and so on. After obtaining the objective functions, we should develop optimization methods to search the optimal solution under the optimization objective functions.

However, for some real-world applications, training sets are usually subject to unknown but complex noises. The underlying assumption of Gaussian distributed error term in traditional models will be not reliable in such case. There are two solutions to solve the regression problems [9]. The first solution is to diagnose the outliers, which can be seen as special noise with long tail [10], then the training samples processed by removing the detected outliers will be fed into the regression models [11]. The second solution is to construct a regression model which is robust to outliers directly [12].

For the first solutions, generally, outliers can be identified by using five basic plots (Graph of predicted residuals, Williams graph, Pregibon graph, McCulloh-and-Meeter graph, L-R graph) [13,14] and other additional methods mentioned in chemometrical textbooks [15]. Besides, there are many outlier detection techniques have been proposed recently, which can be divided four categories [16]: statistical [17], distance-based [18], density-based [19] and soft computing [20]. The final constructed regression model requires two-step procedure [13]. The final regression accuracy depends largely on the goodness of outliers detection results. And those outliers detections methods have the risks of identifying normal points as outliers. In this case, certain information in training samples will lose due to the reduction of useful normal samples, that will have great effects on regression performance, especially for the small size training samples. Moreover, detected outliers may be also contains certain information. For the above

* Corresponding author.
E-mail address: wangyun15@tju.edu.cn (Y. Wang).

http://dx.doi.org/10.1016/j.neucom.2017.01.024
Received 28 July 2016; Received in revised form 23 November 2016; Accepted 4 January 2017
Available online 16 January 2017
0925-2312/

considerations, in terms of regression modeling, we focus on the robust regression models.

In order to improve the robustness of the regression model, much effort has been made in the past few years. The common strategy to enhance the robustness of the regression model is to add weights to different samples. In [21], the authors pointed out that samples with large simulation residuals should be given small weights. And in [22], authors claimed that the relatively smaller weights should be given to the sample points with large distance to others. Four different types of weighting functions including Huber, Hampel, Logistic and Myriad are studied in [23], the results show that Logistic and Myriad weighting function are more robust than the other two functions in most cases [24]. However, it is a difficult task to determine the optimal weight to each sample. Some other researchers suggest using robust loss functions to reduce the effect of different noises. In [25], maximum correntropy criterion that comes from information theoretic learning is selected as a loss function, while a truncated least squares loss function is employed in [26]. This loss function is non-convex, which leads to a difficult optimization task. Besides, many other new regression models are proposed currently [27–30].

Another method to obtain a robust linear regression model is to model the noise comprehensively by mixture distributions. Mixture of Gaussian (MoG) [31], which can approximate any continues noise distributions [32] and is successfully applied in many domains and achieves great success, is used to fit the noise in regression problem. In [33], an autoregressive model was proposed with the noise fitted by MoG. Recently, other mixture distributions such as Mixture of t distributions [34] and scale mixtures of skew-normal distributions [35] were also employed to fit the unknown noise in LR model.

Besides, mixture distributions were also applied in nonlinear regression models. In [36,37], nonlinear regression models based on scale mixtures of skew-normal distributions were proposed with all parameters estimated by Bayesian inference and EM algorithm, respectively. Also, heteroscedastic nonlinear regression models based on scale mixtures of skew-normal distributions were proposed with parameters estimated by EM algorithm in [38]. However, the limitation of the above nonlinear regression models is that the nonlinear functions or nonlinear models are known in advance although the mixture distributions can fit the noise in nonlinear regression models. In literatures, there exist many nonlinear regression models such as SVM, LSSVM and ELM etc. to approximate the unknown nonlinear relationship between inputs and outputs. According to the theory of Bayesian inference, square loss function is optimal when the noise is Gaussian distributed [1]. However, in reality, the noise in real-world is complex or unknown, a single distribution to describe the real noise is improper.

In order to solve the problem of mismatch between the loss function and the real unknown or complex noise distribution, and motivated by successful applications of MoG in linear and nonlinear regression problems under the condition that the linear and nonlinear relationships are known, in this paper, a novel nonlinear regression model is proposed to approximate the unknown nonlinear relationships between inputs and outputs with the feature of noise comprehensively described by MoG. In our paper, in order to solve the single-task and multi-tasks regression problems in reality with unknown or complex noise, we propose two robust nonlinear regression models: single-task nonlinear regression model (SNLR-MoG) and multi-tasks nonlinear regression model (MNLR-MoG), which are all under the MoG noise distribution assumption and are all optimized within EM frameworks.

The contributions of this paper are summarized as follows:

(1) A nonlinear regression technique based on MoG is proposed to build nonlinear regression model with unknown noise.
(2) Expectation Maximization is introduced to solve the proposed nonlinear regression model.
(3) Extensive experiments are conducted, and the regression results are compared with seven popular regression models and show that the proposed model has great advantages under complex or unknown noise conditions.

The rest of this paper is organized as follows. Section 2 describes some related works about LR model under different type of noise. The objective functions of proposed two nonlinear regression models are described in section 3, and the corresponding training processes of two models are introduced in Section 4. Experiments on synthetic datasets and real-world benchmark datasets are carried out and the corresponding results are shown in Section 5. Section 6 concludes the paper.

2. Related work

In this section, some related works about linear regression models under different types of noise are presented. The linear regression model is expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1}$$

where $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$ means the vector of dependent variable, and $\mathbf{X} = [x_1, x_2, \ldots, x_n]^T$ is the matrix of independent variable, $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_k]^T$ is the vector of regression coefficients, $\boldsymbol{e} = [e_1, e_2, \ldots, e_n]^T$ means the vector of model noise. Assuming that the noise of LR model is a Gaussian with zero mean and unknown variance, namely

$$p(e_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right) \tag{2}$$

In this case, the likelihood of $\boldsymbol{e}$ can be written as

$$p(\boldsymbol{e}) = \prod_{i=1}^{n} p(e_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n} e_i^2}{2\sigma^2}\right) \tag{3}$$

By changing from e_i to y_i, the corresponding density is expressed as

$$p(\mathbf{y}|\boldsymbol{\beta}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n} (y_i - x_i\boldsymbol{\beta})^2}{2\sigma^2}\right) \tag{4}$$

Given the likelihood function above, then the log-likelihood can be easily computed as follows:

$$L(\mathbf{y}|\boldsymbol{\beta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} (y_i - x_i\boldsymbol{\beta})^2}{2\sigma^2} \tag{5}$$

The estimated values of parameter $\boldsymbol{\beta}$ can be obtained by maximizing the log-likelihood function $L(\mathbf{y}|\boldsymbol{\beta})$, namely

$$\hat{\boldsymbol{\beta}} = \arg\max_{\beta} L(\mathbf{y}|\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - x_i\boldsymbol{\beta})^2 \tag{6}$$

In this case, maximizing log-likelihood is equivalent to minimizing the sum of squared residuals, and then we have

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{7}$$

Concluded from the above interference, when given the noise distribution $p(e_i)$ in linear regression model, the regression coefficient can be obtained by maximizing the log-likelihood function $L(\mathbf{y}|\boldsymbol{\beta})$. Generally, the assumption that the noise is Gaussian distributed is sometimes improper in real-world applications. Therefore, in linear regression model, the original assumption that the noise obeys Gaussian distribution is replaced by the assumption that the noise obeys different types of distributions, such as Laplace distribution, Beta distribution and Huber distribution etc. And hence, the different improved linear regression models with different single noise distributions are proposed. Table 1 shows the different noise distributions and their corresponding optimal objective functions.

In practice, the noise is complex and unknown if the data are