A Stochastic Gradient Descent Logistic Regression Software Program for Civil Engineering Data Classification Developed in .NET Framework

Phần mềm phân loại dữ liệu trong ngành xây dựng sử dụng thuật toán hồi quy logistic phát triển trên nền tảng .NET

Hoàng Nhật Đức[1] và **Nguyễn Huy Thành**[2]

[1]Viện Nghiên Cứu Phát Triển Công Nghệ Cao, Đại học Duy Tân, Đà Nẵng
Institute of Research and Development, Duy Tan University
*Email:*hoangnhatduc@dtu.edu.vn

[2]Công Ty Quản Lý Cầu Đường Đà Nẵng, Đà Nẵng
Da Nang Road and Bridge Management Company, Da Nang City
Email: huythanh307@gmail.com

Tóm tắt

Nghiên cứu này xây dựng mô hình phân loại dữ liệu dựa trên mô hình hồi quy logistic và thuật toán stochastic gradient descent. Mô hình có tên là SGD-LR được phát triển bằng ngôn ngữ C# trên nền tảng NET Framework 4.6.2. Khả năng phân tích dữ liệu của mô hình được minh chứng qua các dữ liệu thử nghiệm trong bài báo.

Từ khóa: Phân loại dữ liệu; Hồi quy logistic; Ngành xây dựng; Ngôn ngữ C#, .NET.

Abstract

This study constructs a Stochastic Gradient Descent Logistic Regression (SGD-LR) used for data classification. A software program based on the SGD-LR has been developed by the authors in Visual C# and .NET Framework 4.6.2. The program capability has been demonstrated by case studies in this work.

Key words: Data classification; Logistic regression, Civil Engineering; Visual C#; .NET Framework.

1. Introduction

Data classification is a widely encountered task in civil engineering [1,2]. This task typically involves the establishment of the mathematical relationship between a set of influencing factors and a modeled variable. The influencing factors are used as input information for the established mathematical model; the output of interest is the class label corresponding to the provided input information (see **Fig. 1**).

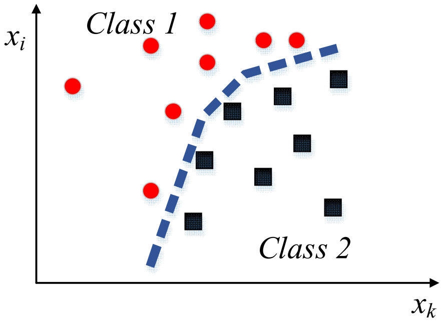


Fig. 1 The data classification process

The classification model can significantly help the efficiency of the decision making processes in the civil engineering field. Thus, data classification models are widely employed in civil engineering such as liquefaction prediction [3], groutability estimation [4-6], slope stability analysis [7], pavement distress recognition [8], etc. Among the classification models, the Logistic Regression (LR) is widely employed due to its fast model construction phase and good prediction performance [8,9]. Thus, this study aims at developing a software program based on the LR model and the Stochastic Gradient Descent (SGD) training algorithm. The SGD algorithm is used due to its effectiveness demonstrated in previous studies [8].

The software program, named as SGD-LR, has been developed in Visual C# .NET framework 4.6.2. A Graphical User Interface (GUI) has been constructed by the authors to ease the program implementation. The rest of the paper is organized as follows: the second section briefly mentions the formulation of the SGD-LR; two application cases of the software program are demonstrated in the third section; concluding remarks of this paper are stated in the final section.

2. Stochastic Gradient Descent Logistic Regression

The LR model can help to build a classification model that separates samples belonging to two possible categories namely the negative and postive classes. This machine learning model is straightforward to implement and its structure is also ease to interpret [10]. Successful applications of LR have been reported in various studies [11,12].

For the purpose of data modeling, the output of a LR model (y) is often denoted as 1 for positive cases and 0 for negative cases [8]. An input feature is given in the form of $x_i = x_{i1}, x_{i2}, ..., x_{iD}$ where D is the number of the features used for classification. $\theta = \theta_0, \theta_1, \theta_2, ..., \theta_D$ represents the model parameters. The quantity $h_\theta(x_i)$ is employed to express the probability of the positive class output of raveling. $h_\theta(x_i)$ is calculated as follows [13]:

$$h_\theta(x_i) = h_\theta(x_{i1}, x_{i2}, ..., x_{iD}) = \frac{1}{1+\exp(-\eta_i)} = \frac{1}{1+\exp(-\theta^T x_i)} \qquad (1)$$

where $\eta_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + ... + \theta_D x_{iD} = \theta^T x_i$.

$g(\eta_i) = \dfrac{1}{1+\exp(-\eta_i)}$ is known as the logistic function; its derivative is given as follows [14]:

$$g'(\eta_i) = g(\eta_i) \times (1 - g(\eta_i)) \tag{2}$$

The probabilities of the positive (*y* = 1) and negative (*y*=0) classes are given as follows:

$$P(y_i = 1 | x_i, \theta) = h_\theta(x_i) \tag{3}$$

$$P(y_i = 0 | x_i, \theta) = 1 - h_\theta(x_i) \tag{4}$$

Hence, the output probability can be given as follows [14]:

$$P(y_i = 0 | x_i, \theta) = (h_\theta(x_i))^{y_i} (1 - h_\theta(x_i))^{1-y_i} \tag{5}$$

Thus, the likelihood of the parameters can be expressed as follows [14]:

$$L(\theta) = \prod_{i=1}^{M} (h_\theta(x_i))^{y_i} (1 - h_\theta(x_i))^{1-y_i} \tag{6}$$

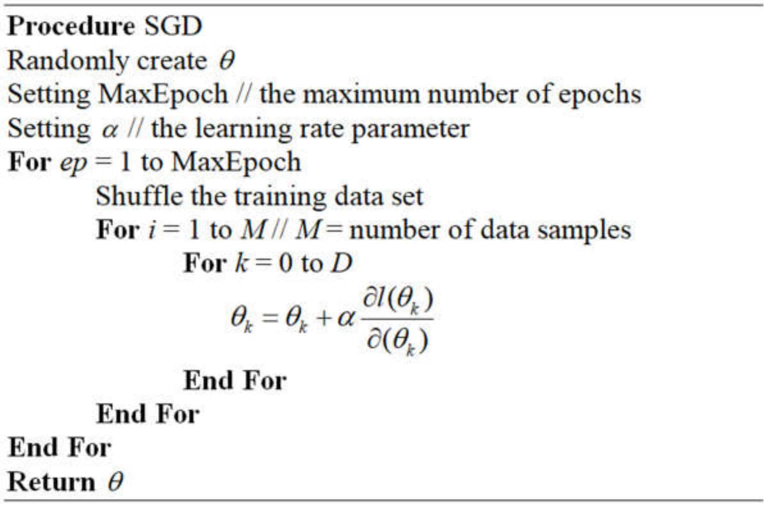
where *M* represents the number of data samples.

In order to determine the set of model parameters θ, one has to maximize the following log likelihood function:

$$l(\theta) = \log(L(\theta)) = \sum_{i=1}^{M} y_i \log(h_\theta(x_i) + (1 - y_i)(1 - \log(h_\theta(x_i)) \tag{7}$$

The SGD algorithm (see **Fig. 2**) is employed in this study to identify the set of model parameters θ. Using the SGD, the parameter θ is updated according to the following rule:

$$\theta_k = \theta_k + \alpha(y_i - h_\theta(x_i))x_{i,k} \tag{8}$$

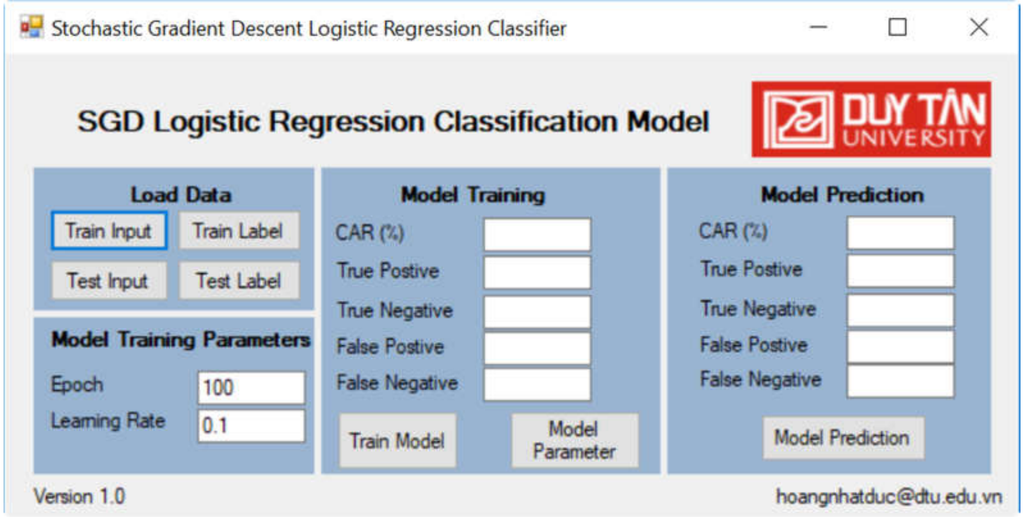


```
Procedure SGD
Randomly create $\theta$
Setting MaxEpoch // the maximum number of epochs
Setting $\alpha$ // the learning rate parameter
For $ep$ = 1 to MaxEpoch
    Shuffle the training data set
    For $i$ = 1 to $M$ // $M$ = number of data samples
        For $k$ = 0 to $D$
            $\theta_k = \theta_k + \alpha \frac{\partial l(\theta_k)}{\partial(\theta_k)}$
        End For
    End For
End For
Return $\theta$
```

Fig. 2 The SGD algorithm for training the LR model

3. The SGD-LR Software Program Applications

To automatically implement the LR model employing the SGD algorithm, a software program has been developed in .NET framework 4.6.2. The Graphical user interface (GUI) of the software program is shown in **Fig. 3**. The current program supports a single run of model training and testing phase. The training input file and its corresponding label file (TrainInput.csv and TrainLabel.csv) are stored in .csv file. Similarly, the testing input file and its corresponding label file should be named as TestInput.csv and TestLabel.csv. The user needs to provide the number of training epoch and the learning rate shown in the panel of Model Training Parameters.


Epoch 100

Learning Rate 0.1

Fig. 3 The SGD-LR software program

Notably, the input features of the data set should been normalized by the Z-score equation given as follows:

$$X_{ZN} = \frac{X_o - m_X}{s_X} \quad (9)$$

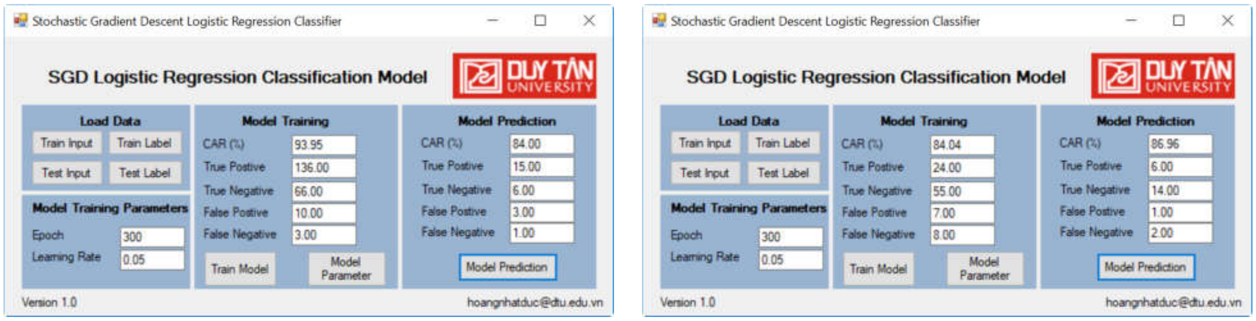
where X_o and X_{ZN} denote an original and a normalized input feature, respectively. m_X and s_X denote the mean and the standard deviation of the original input feature, respectively.

A common practice is to randomly divide the collected dataset into two sets of training (90%) and testing data (10%). The training set is used to construct the SGD-LR model; the testing set is used for evaluating the model generalization capability. To evaluate a LR model performance, the Classification Accuracy Rate (CAR) is often employed:

$$\text{CAR} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (10)$$

where TP, TN, FP, and FN denote the true positive, true negative, false positive, and false negative values, respectively.

To demonstrate the operation of the newly developed software program, two case studies related to groutability estimation are employed. Data in first case study is collected from the previous work of [6]. In the latter case study, data presented in the previous work of [4,15] is employed. In both case studies, a set of input features is employed to predict the state of groutability (either groutable which is the positive class or ungroutable which is the negative class). The prediction results of the two case studies are reported in **Fig**. **4**, **5**, and **6**. Good CAR values are observed for both case studies (84.00% and 86.96%).



(a)

Model Training Parameters

Epoch	300
Learning Rate	0.05

	Model Training	Model Prediction
CAR (%)	93.95	84.00
True Postive	136.00	15.00
True Negative	66.00	6.00
False Postive	10.00	3.00
False Negative	3.00	1.00

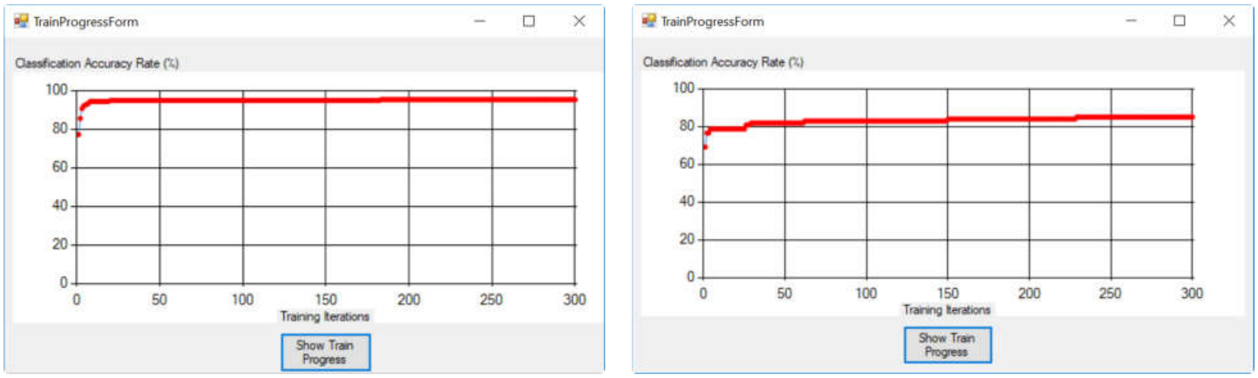
(b)

Model Training Parameters

Epoch	300
Learning Rate	0.05

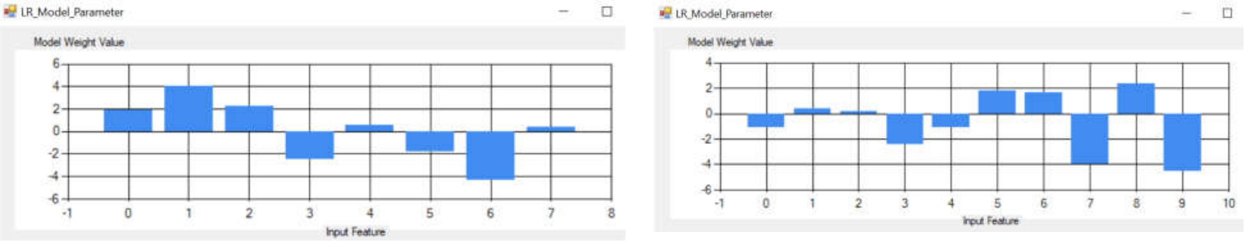
	Model Training	Model Prediction
CAR (%)	84.04	86.96
True Postive	24.00	6.00
True Negative	55.00	14.00
False Postive	7.00	1.00
False Negative	8.00	2.00

Fig. 4 Prediction results: (a) Case study 1 and (b) Case study 2



(a) (b)

Fig. 5 The model training progress: (a) Case study 1 and (b) Case study 2



(a) (b)

Fig. 6 The model parameters: (a) Case study 1 and (b) Case study 2

4. Conclusion

Data classification is an important task in civil engineering. This study develops a software program based on the LR algorithm and the SGD training algorithm. The applicability of the software program has been illustrated by two case studies using collected data sets of the groutability estimation. Good predictive results show that the SGD-LR software program can be a useful tool to assist decision makers in civil engineering.

Supplementary materials

The software program can be downloaded via:
https://github.com/NhatDucHoang/SGD_LR_Software

References

1. Huang H, Burton HV (2019) Classification of in-plane failure modes for reinforced concrete frames with infills using machine learning. Journal of Building Engineering 25:100767. doi:https://doi.org/10.1016/j.jobe.2019.100767

2. Chang M, Maguire M, Sun Y (2019) Stochastic Modeling of Bridge Deterioration Using Classification Tree and Logistic Regression. Journal of Infrastructure Systems 25 (1):04018041. doi:10.1061/(ASCE)IS.1943-555X.0000466

3. Hoang N-D, Bui DT (2018) Predicting earthquake-induced soil liquefaction based on a hybridization of kernel Fisher discriminant analysis and a least squares support vector machine: a multi-dataset study. Bulletin of Engineering Geology and the Environment 77 (1):191-204. doi:10.1007/s10064-016-0924-0

4. Tekin E, Akbas SO (2017) Predicting groutability of granular soils using adaptive neuro-fuzzy inference system. Neural Computing and Applications. doi:10.1007/s00521-017-3140-3

5. Hoang N-D, Tien Bui D, Liao K-W (2016) Groutability estimation of grouting processes with cement grouts using Differential Flower Pollination Optimized Support Vector Machine. Applied Soft Computing 45:173-186. doi:https://doi.org/10.1016/j.asoc.2016.04.031

6. Liao K-W, Fan J-C, Huang C-L (2011) An artificial neural network for groutability prediction of permeation grouting with microfine cement grouts_DataFile. Computers and Geotechnics 38 (8):978-986. doi:https://doi.org/10.1016/j.compgeo.2011.07.008

7. Zhou LY, Shan FP, Shimizu K, Imoto T, Lateh H, Peng KS (2017) A comparative study of slope failure prediction using logistic regression, support vector machine and least square support vector machine models. AIP Conference Proceedings 1870 (1):060012. doi:10.1063/1.4995939

8. Hoang N-D (2019) Automatic detection of asphalt pavement raveling using image texture based feature extraction and stochastic gradient descent logistic regression. Automation in Construction 105:102843. doi:https://doi.org/10.1016/j.autcon.2019.102843

9. Kim K, Kim J, Kwak T-Y, Chung C-K (2018) Logistic regression model for sinkhole susceptibility due to damaged sewer pipes. Natural Hazards 93 (2):765-785. doi:10.1007/s11069-018-3323-y

10. Piegorsch WW (2015) Statistical Data Analytics: Foundations for Data Mining, Informatics, and Knowledge Discovery. John Wiley & Sons, Ltd, ISBN 978-1-118-61965-0,

11. Saha TK, Pal S (2019) Exploring physical wetland vulnerability of Atreyee river basin in India and Bangladesh using logistic regression and fuzzy logic approaches. Ecological Indicators 98:251-265. doi:https://doi.org/10.1016/j.ecolind.2018.11.009

12. Kim H, Hong T, Kim J (2019) Automatic ventilation control algorithm considering the indoor environmental quality factors and occupant ventilation behavior using a logistic regression model. Building and Environment. doi:https://doi.org/10.1016/j.buildenv.2019.02.032

13. Agresti A (2019) An introduction to categorical data analysis. John Wiley & Sons, Inc, Hoboken, NJ 07030, USA, ISBN 9781119405283,

14. Ng A (2018) Lecture notes. CS229 Machine Learning. Stanford University, http://cs229.stanford.edu/notes/cs229-notes1.pdf (Last Access 12/13/2018)

15. Cheng M-Y, Hoang N-D (2014) Groutability prediction of microfine cement based soil improvement using evolutionary LS-SVM inference model. Journal of Civil Engineering and Management 20 (6):839-848. doi:10.3846/13923730.2013.802717