Dean&Francis

ISSN 2959-6157

The Data Preprocessing Technique in Machine Learning

Jiale Deng[1, *]

[1]Pegasus California school, Qingdao, China

*Corresponding author: Djl17685516065@outlook.com

Abstract:

Data preprocessing has a vital impact on the performance of traditional machine learning. However, with the continuous development of deep learning technology and the powerful representation learning ability of neural networks, deep learning models can easily convert raw data into continuous feature representation and have made remarkable achievements in many downstream tasks. Recently, with the application of deep learning technology in more fields that require robustness and stability, its model bias caused by data quality problems are gradually exposed, which makes the data preprocessing technology regain the attention of researchers. This paper systematically expounds on the primary data preprocessing technology in machine learning and discusses the model bias caused by data deviation and the robustness in the face of attacks. In addition, this paper also illustrates the possibility of a data preprocessing foundation in solving the defects of deep learning foundation based on glove word vector and image classification task based on convolutional neural network. The research in this paper can provide a valuable reference for researchers in related fields.

Keywords: Data Preprocessing; Feature engineering; Machine learning; Deep learning.

1. Introduction

In machine learning, data preprocessing is a crucial step that directly affects the effect and accuracy of subsequent model training. The main goal of the conventional data preprocessing process is to clean, transform and standardize the original data, making it more suitable for machine learning algorithms. Data cleaning generally includes 1) missing value processing. Data missing values can be handled by deleting records containing missing values, filling missing values with mean/median/mode, or using interpolation methods. 2) Detection and treatment of outliers. Statistical methods (such as the 3σ principle), box charts (IQR) or model-based methods can be used to identify and deal with outliers in data. 3) Duplicate data deletion: Remove duplicate records in the data set to avoid affecting the model training results. Data conversion usually includes 1) feature extraction: extracting valuable features from the original data, such as extracting keywords through text analysis and extracting edges in image processing. 2) Feature construction: Create new features by combining existing features or applying mathematical transformations (such as logarithmic and polynomial transformations). 3) Normalization and standardization: Scaling the data to a standard range, such as min-max normalization and Z-score standardization, to improve the convergence speed and performance of the algorithm. In addition, for imbalanced data sets, resampling techniques (such as oversampling and undersampling) and synthetic minority oversampling techniques are usually needed to deal with the imbalance of categories in data sets.

With the development of deep learning and the arrival of the era of big data, representation-learning technology based on neural networks can quickly learn the feature distribution of data from large-scale data sets and transform the original data into continuous feature representations with rich semantic information. Typical technical representatives include word vector technology in natural language processing, visual Transformer in computer vision, and CLIP model integrating multimodal information. Due to the powerful ability of emerging representation learning methods in capturing data distribution law and the support of large-scale data sets, traditional data preprocessing has become an unnecessary link. However, with the rise of the significant language model, researchers gradually found that problems such as considerable model bias and illusion caused by data deviation became challenging to solve, which made researchers pay attention to the data preprocessing again [1-3].

In this paper, the importance of data preprocessing in the current training process of the deep learning model is expounded through the bias of word vectors in natural complementary processing and the anti-attack problem in computer vision. In the second section, this paper expounds on the bias in word vectors, and in the third section, it expounds on the problem of adversarial attack in computer vision. In the fourth section, this paper discusses the future research direction and challenges of data preprocessing technology in the era of deep learning and large language models.

2. Data Bias in Word Embedding

Word embedding is a technology in natural language processing that maps words or phrases in vocabulary into vector space [4]. This representation makes the similarity and contextual relationship between words can be quantified by the distance and direction of vectors, which significantly promotes the effect of natural language processing tasks. The training of word vectors is usually based on the context window, considering several words before and after the target word as its context. We can learn the relationship between words by training the model to predict the target word or context. With the continuous progress of deep learning technology, word vector technology is also developing. For example, pre-training language models (such as BERT, GPT, etc.) can learn more abundant language expressions by pre-training on a large-scale corpus, further improving the effect of natural language processing tasks. These models consider the contextual information of words and incorporate more advanced language features such as syntactic structure and semantic roles. However, in addition to learning the distribution law of features in language, the problem of language bias is also captured simultaneously. In this paper, the mainstream word vector Glove is selected, and a group of words that may be related to social prejudice are specially selected for similarity calculation, as shown in Table 1.

Table 1. Three Scheme comparing

	man	woman	CEO	manager	leader	chairman	nurse	cry
man	1	0.83	0.30	0.46	0.53	0.32	0.46	0.41
woman	789	1	0.20	0.24	0.40	0.19	0.61	0.41
CEO	213	654	1	0.63	0.43	0.77	0.18	0.03
manager				1	0.37	0.57	0.32	0.14
leader					1	0.57	0.19	0.21
chairman						1	0.12	0.05
nurse							1	0.25
cry								1

As seen from Table 1, male "man" is more similar to occupations with higher income and social status, such as "CEO", which aligns with social stereotypes. However, this situation needs to be considered when it appears in the word vector. Because there is no cleaning and correction of social prejudice in the data preprocessing stage, all prejudices that may exist in the original data will be encoded into the word embedding with a certain probability. This will seriously interfere with various downstream applications based on word embedding. Traditional data preprocessing emphasizes cleaning the noise in the data set. However, the above problems mean that it is a great challenge to design a new data preprocessing technology to clean the biased information in the original data set in natural language processing.

3. Data Noise in the Adversarial Attack

Adversarial attack are an essential research topic in machine learning and computer security. It refers to the attacker modifying the input data by carefully designing minor disturbances, thus making the machine learning model produce the wrong output. This attack method poses a potential threat to many systems based on machine learning, such as automatic driving, image recognition, voice recognition, etc. There is no strong evidence to prove the cause of the Adversarial attack. However, a mainstream view is that the specific noise in the original image data has yet to be cleaned up in the data preprocessing stage.

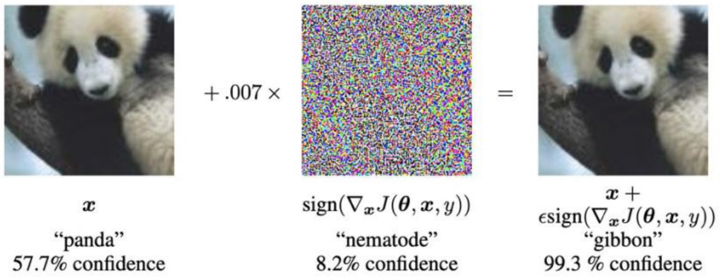
+ .007 ×
=
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Fig. 1 An example of adversarial example generated by GoogLeNet. Slight perturbation mislead the GoogLeNet's judgment [5].

From the example in Figure 1 in the application field of image classification, It is simple to understand how a neural network model may have positively identified a picture of a panda. Next, a slight perturbation is added to this photo, resulting in a generated photo that looks almost indistinguishable from the original from a human perspective. However, inputting the generated photo into the same neural network model results in a high-confidence (99.3%) misclassification judgment (gibbon).

The attack method used in the experiments in this paper is FGSM (Fast Gradient Sign Attack). FGSM, a straightforward and very effective algorithm for producing adversarial examples, is one of the simplest and most well-known image adversarial attack techniques, as shown in formula (1):

$$x' = x + \epsilon sign\left(\nabla_x L\left(\theta, x, t\right)\right) \qquad (1)$$

As shown in formula (2), by using one-step gradient descent. the formula (3) can be solved:

$$minimize L\left(\theta, x', t\right) \qquad (2)$$

$$\text{s.t.} \quad \| x' - x \|_\infty \leqslant \epsilon \quad \text{and} \quad x' \in \left[0,1\right]^m$$

The method searches the the original image's ϵ neighborhood for the perturbation signal value, enabling the adversarial example x' to be classified as the wrong label t. Compared with the method of Szegedy et al., the FGSM algorithm needs only one back-propagation process to generate adversarial examples.

During network training, FGSM learns the input image features and obtains the classification probability through the softmax or sigmoid layers. Then calculate the loss value with the obtained classification probability and the real label, return the loss value and calculate the gradient, that is, gradient backpropagation. To make the loss value greater than the loss value of the entire image when the changed image is fed into the classification network, it is simply necessary to add the calculated gradient direction to the input image. CIFAR-10 dataset is trained and tested using LeNet, ResNet-18, and VGG16 neural network models [6,7]. Moreover, use FGSM to attack, get the experimental results.

Table 2. Attack effect of FGSM attack method on CNNs

Epsilon	0	0.05	0.1	0.15	0.2	0.25	0.3
LeNet	0.53	0.2	0.06	0.02	0.01	0.0045	0.0028
ResNet18	0.77	0.02	0.02	0.02	0.02	0.03	0.04
VGG16	0.80	0.066	0.02	0.02	0.02	0.03	0.04

Using the LeNet model to train and test CIFAR-10, in the case of Epsilon=0, the accuracy is only 0.5336. It shows that the CIFAR-10 data set is much more complicated than the previously mentioned handwritten digit picture dataset. The ResNet-18 model has a correct rate of 0.7665 when Epsilon=0, which shows that the ResNet-18 model is better than LeNet. Of course, the capacity of the ResNet-18 model should be more significant. The accuracy of the VGG16 model is the highest among the three models, reaching 0.8002. When the FGSM attack is added, the accuracy of the three models shows a linear decline. It is not difficult to show that the FGSM attack is effective for the three models of LeNet, ResNet-18, and VGG16, as shown in Table 1. A more vivid example is shown in Figure 2.

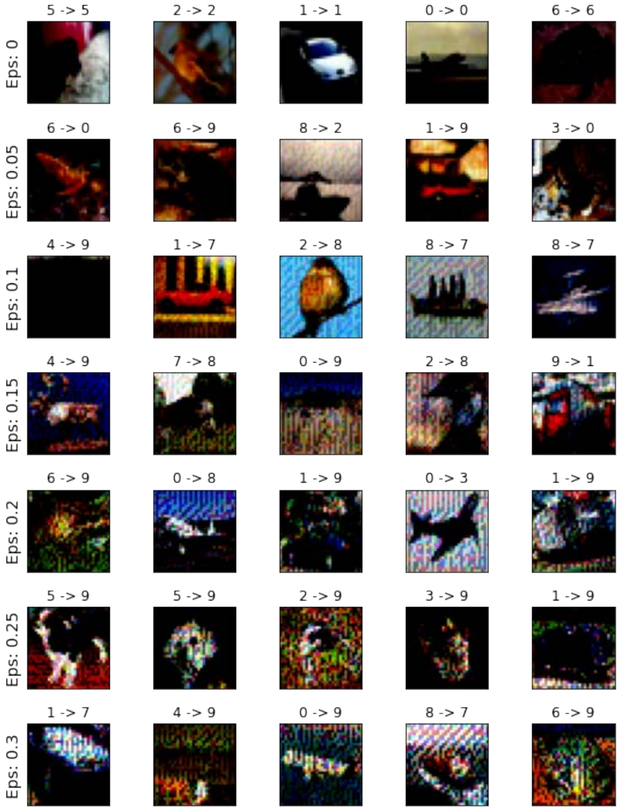

Fig. 2 Image samples under different attack intensities

As can be seen from Figure 2, different degrees of noise appear in the original image. In computer vision, this kind of similar noise is often removed in the data preprocessing stage, thus ensuring that the model is not affected by low-quality images. An intuitive understanding is that this denoising in the data preprocessing stage leads to the model's vulnerability when it faces noise attacks. Therefore, the existing work enhances the robustness of the model by adding noise in the training stage. However, the above problems have not been completely solved. In the data preprocessing stage, the contradictory training skills of reducing noise and adding noise mean that researchers in related fields need to rethink data characteristics and how to deal with noise in the data preprocessing stage.

In addition to the above problems, with the advent of big data, data preprocessing technology is also facing other challenges. For example, with data volume increase, the time and resources required for data preprocessing will also increase significantly. Especially for large-scale data sets, the preprocessing process may be very time-consuming and occupy many computing resources. With the diversification of data sources, the processing of heterogeneous data will also become an important development direction. In addition, data security has become a key issue in big data and cloud computing. Future data preprocessing technology needs to pay more attention to data security and privacy protection to ensure data security and compliance in the preprocessing process.

4. Conclusion

With the continuous development of deep learning technology and the powerful representation learning ability of neural networks, deep learning models can easily convert raw data into continuous feature representation and have made remarkable achievements in many downstream tasks. In recent years, with the application of deep learning technology in more fields that require robustness and stability, the model deviation caused by data quality problems is gradually exposed, which makes the data preprocessing technology get researchers' attention again. This paper systematically expounds the primary data preprocessing techniques in machine learning, and the model deviation caused by data deviation and the robustness against attacks are discussed. In addition, this paper also illustrates the possibility of a data preprocessing foundation in solving the defects of deep learning foundation based on glove word vector and image classification task based on convolutional neural network. Finally, this paper discusses other challenges and development directions of data preprocessing technology. The research in this paper can provide a valuable reference for researchers in related fields.

References

[1] Chunlong Xia, Xinliang Wang, Feng Lv, et al. Vit-comer: Vision transformer with convolutional multi-scale feature interaction for dense predictions//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 5493-5502.

[2] S. Joshua Johnson, M. Ramakrishna Murty, I. Navakanth. A

ISSN 2959-6157

detailed review on word embedding techniques with emphasis on word2vec. Multimedia Tools and Applications, 2024, 83(13): 37979-38007.
[3] Lijie Fan, Dilip Krishnan, Phillip Isola, et al. Improving clip training with language rewrites. Advances in Neural Information Processing Systems, 2024, 36.
[4] Andrea Vallebueno, Cassandra Handan-Nader, Christopher D. Manning, et al. Statistical Uncertainty in Word Embeddings: GloVe-V. arXiv preprint arXiv:2406.12165, 2024.
[5] Linyu Tang, Lei Zhang. Robust Overfitting Does Matter: Test-Time Adversarial Purification With FGSM[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 24347-24356.
[6] Farhan Ullah, Irfan Ullah, Rehan Ullah Khan, et al. Conventional to deep ensemble methods for hyperspectral image classification: A comprehensive survey. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024.
[7] Ashish Bajaj, Dinesh Kumar Vishwakarma. A state-of-the-art review on adversarial machine learning in image classification. Multimedia Tools and Applications, 2024, 83(3): 9351-9416.