

BiLSTM-Ridge Regression Meta Learning Model for Few-Shot Logs Classification

Hanji Shen* **, Jun Li*, Xiang Fu***

* Computer Network Information Center, Chinese Academy of Sciences, China

** University of Chinese Academy of Sciences, China

*** Blue Sky Frontier Science and Technology Innovation Center, Beijing, China
shenhanji@cnic.cn, jlee@cstnet.cn, fuxiangpku@163.com

Abstract—In the high threat attack classification detection task,, it is hard to find attack logs in the huge log set because the number of attack logs in the log set is particularly small, which is a major difficulty in the high threat attack classification detection task. This paper proposes a Meta learning classification model based on few-shot samples. Aiming at the uneven distribution of samples in the data set. Firstly, processing the source data set and enhance the data set to build a data set that can meet the C-way k-shot, then using fasttext to pre-train the data set samples, Finally, building a meta knowledge learner by using BiLSTM and support set meta training classifier by using ridge regression. The experimental results show that the proposed BiLSTM-Ridge regression model shows good results in the classification and detection of small sample attack logs on the real 335 attack log data sets of China Science and technology network. The accuracy of 5-way 3-shot can reach 66.47%, and the accuracy of 5-way 1-shot can reach 49.72%, which is improved compared with the typical small sample model.

Keywords—Few shot, Meta Learning, BiLSTM, Ridge Regression, High threat attack

I. INTRODUCTION

In recent years, with the significant improvement of computer computing power and the development of big data technology, deep learning models are more and more applied to natural language processing, speech recognition, image recognition and other fields, and have achieved blowout applications. In many large data sets, the recognition level of some deep learning model algorithms even exceeds the human resolution. However, the strong performance of these algorithms largely depends on the sufficiency of the data set, including the number of samples in the data set, whether the samples are well marked, and whether the distribution of samples covers the target domain. On the issue of high threat attack log classification detection targets in the cloud environment, the opportunity to obtain attack samples is very small. First, the number of high threat vulnerabilities is too small. Through the scarcity of log data samples of high threat attacks, it is difficult to establish a sufficient sample database to classify and detect high threat attack log data. In contrast, experienced network security engineers can complete rapid learning for a very small number of attack samples, and have the learning ability of rapid migration and generalization. In

the field of deep learning, the research that makes the model get rid of the dependence of a large number of data sets and have the ability of self-learning is called Meta learning.

At present, there are two main strategies for text classification with few shots: translation learning based method and Meta learning based method. By transferring the knowledge learned from the task to the new task or using the general domain pre training and fine tuning technology for small sample classification, pan et al. Proposed an improved hierarchical knowledge pool strategy based on pre training word embedding to transfer the knowledge from the source domain to the target domain [1], Gupta et al. [2] developed a binary classifier on the source domain, Classify new classes by adding a class identifier prefix to the input text. Typical algorithms of meta learning model include MAML [3] proposed by Finn et al., LSTM meta learning method [4] proposed by Ravi et al., and meta network model [5] proposed by munkhdalai et al; Typical algorithms for metric learning include the twin network model proposed by Koch et al. [6], the matching network model proposed by vinyas et al. [7], the prototype network model proposed by Snell et al. [8] and the relational network model proposed by Sung et al. [9]. BiLSTM can better understand context and can demonstrate its advantages in many different text mining tasks, such as sentiment analysis [10], sentence similarity [11], and named entity recognition [12]

This paper proposed a few-shot Meta learning model based on BiLSTM-Ridge regression. Aiming at the uneven distribution of data set samples, the source data set is pre-processed and enhanced to build a data set that can meet the C-way k-shot, and then the data set samples are pre trained with fasttext [13] to build a BiLSTM meta knowledge learner, Finally, ridge regression is used for support set learning training. Try to learn the knowledge generalization ability of network security engineers and support the classification and detection of attack logs under few samples.

II. BiLSTM-RIDGE META LEARNING MODEL

Our model is as Figure 1, mainly including source data set processing, meta knowledge learner, sample representation, meta learning ,meta training.

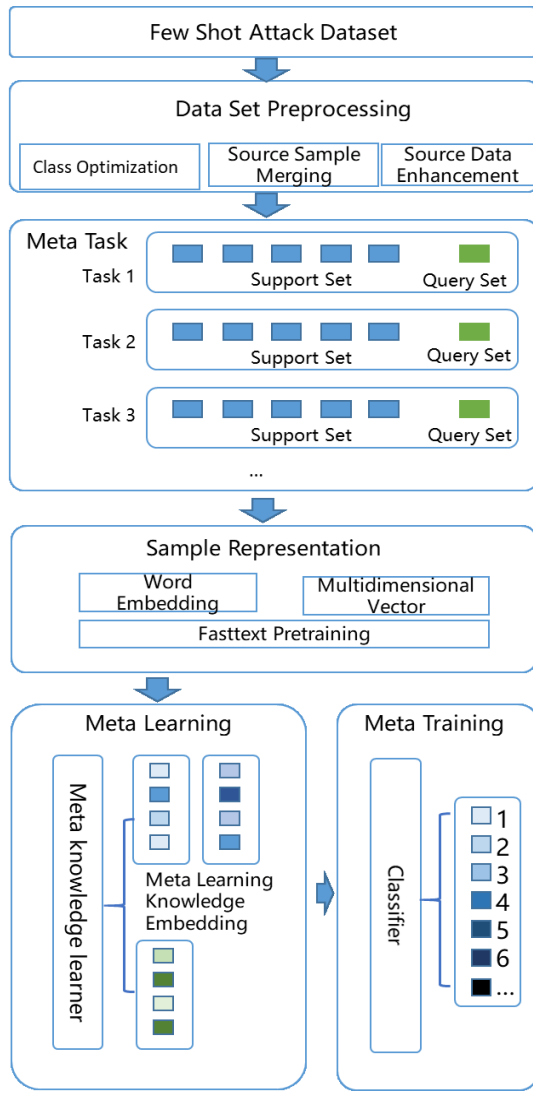


Figure 1. BILSTM-Ridge Meta Learning Model

A. Sample Representation

Sample representation aims to represent each word with a d -dimensional vector. We construct the d -dimensional vector with the word embedding, which is pre-trained by fasttext.

B. Meta Knowledge Learner

Meta knowledge learner aims to make the final classification result better. The module is mainly composed of a bidirectional LSTM (BiLSTM) and a full connection layer. We use BiLSTM to encode context embedding for each time step. The input of the module is a series of word vectors $Q: [q_1, \dots, q_m]$, where M represents the number of words in the sentence. The output is a matrix $v_{d \times m}^q$ that consists of context embedding.

$$\begin{aligned} \vec{v}_i^q &= \vec{LSTM}(\vec{v}_{i-1}^q, q_i) \quad i = 1, 2, \dots, m \\ \leftarrow v_i^q &= \leftarrow LSTM(\leftarrow v_{i+1}^q, q_i) \quad i = m, m-1, \dots, 1 \end{aligned}$$

$$v_i^q = \text{Concat}(\vec{v}_i^q, \leftarrow v_i^q) \quad i = 1, 2, \dots, m$$

$$v^q = [v_1^q, v_2^q, \dots, v_m^q]$$

Then, use the softmax function to obtain the output KQ. KQ is an n -dimensional vector representing the Meta knowledge of log samples, and N is the length of samples.

$$k^q = \text{Softmax}(\omega \times v^q + b)$$

C. Meta Training Classifier

We choose ridge regression to fit the support set: 1) if neural network is selected as the classifier, the number of samples in the support set is too small and the training is insufficient. 2) Ridge regression allows a closed form solution and reduces over fitting on small support sets through appropriate regularization.

Loss function: in each training set, we first fix the parameters of the Meta knowledge generator and update the parameters of the Meta training classifier through the support set. The loss function is as follows:

$$L(\theta) = \frac{1}{m} \sum_{i=1}^m (f_{\theta}(x^{(i)}) - y^{(i)})^2$$

Where m represents the number of samples in the support set, $f_{\theta}(x^{(i)})$ denotes the prediction of ridge regression, $y^{(i)}$ denotes the label of samples.

III. ALGORITHM DESIGN

The multi classification problem uses the small sample high threat log analysis method based on Meta learning, and constructs the small task set and test set of Meta learning based on the existing small sample data set. Before training, the data set is divided into two parts: the Meta training set M_{train} for generalized feature learning and the Meta test set M_{test} for testing the classification ability of the model. In order to effectively feedback the generalization ability of the model, M_{train} and M_{test} do not set the intersection.

A. Meta-training

N classes are extracted from the training data M_{train} . For each of these n classes, two subset examples are sampled as the support set M_{support} and the query set M_{query} . Next, input the support set M_{support} and the query set M_{query} into the model, and update the parameters by minimizing the loss in the query set M_{query} . The above process is called Meta training and will be repeated many times.

B. Meta-testing

After Meta training, the performance of the model will be evaluated through the same event based mechanism. In one test, n new classes will be extracted from M_{test} , which does not intersect M_{train} . Then, the support set and test set are extracted from N classes. Model parameters can be fine-tuned through small support sets.

C. Algorithm

Algorithm: BiLSTM-Ridge Regression Meta Learning Produce
Input: meta training set: $M_{train}\{M_{Support}, M_{query}\}$; meta testing set: $M_{test}\{M_{Support}, M_{test}\}$; number of meta training iterations: ep_{train} ; number of meta testing iterations: ep_{test} ; learner parameters: β ; classifier parameters: θ Output: trained parameters: θ ; average classification accuracy: acc_{right}
Meta-training: 1: Randomly initialize the parameters β, θ ; 2: for $ep=1$ to ep_{train} 3: Randomly select training tasks from M_{train} , Task set $T=\{T_1, T_2, T_3, \dots, T_N\}$, (Each task includes support set and query set); 4: for $T_n \in T$ 5: Fix θ by minimizing the loss of the classifier (Using $L(\theta) = \frac{1}{m} \sum_{i=1}^m (f_{\theta}(x^{(i)}) - y^{(i)})^2$); 6: end for 7: end for 8: Output θ ; Meta-testing: 9: for $ep=1$ to ep_{test} 10: Randomly select testing task T_j from M_{test} , T_j includes support set and query set; 11: Calculate the accuracy of testing task classification by using the trained meta learning model; 12: end for 13: Calculate the average classification accuracy of test tasks acc_{right} ; 14: Output acc_{right} ;

Calculate the average classification accuracy of test tasks, and evaluate the average classification accuracy of query sets in all test tasks. The formula is as follows:

$$acc_{right} = \frac{N_{right}}{Task_{num} \times k}$$

Where N_{right} denotes the correct quantity predicted in all query sets, $Task_{num}$ denotes the number of Meta testing tasks, k denotes the number of samples in each query set.

IV. EXPERIMENT AND ANALYSIS

The experiment environment adopts Huawei server Fusion Server G5500 (CentOS 7.6 system, Intel Golden 6230 Processor, 384 GB memory, Tesla V100 GPU Card), python development language and scikit-learn model library.

Taking the attack log of a site of China Science and technology network from October 20 to November 19, 2021 as the data set, it is found that there are 335 high threat logs. The sample distribution is as following table 1:

Table 1. Source Data Set Distribution

Attack Trying Mode	Count
Code injection	115
XML injection	74
Command injection	60

Exploit vulnerability	33
File restrictions	14
Local file contains	12
Directory information disclosure	10
SQL injection	7
File injection	4
Web Shell	2
Malformed request	1
Java attack	1
Pseudo protocol	1
XSS attack	1

Attack log 1:

```
POST /Autodiscover/Autodiscover.xml HTTP/1.1
Host: 159.XX.XX.XX
User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/78.0.3904.108 Safari/537.36
Content-Length: 314
Content-Type: application/xml

<!DOCTYPE xxe [
<!ELEMENT name ANY >
<!ENTITY xxe SYSTEM "file:///etc/passwd">]>
<Autodiscover
xmlns="http://schemas.microsoft.com/exchange/autodiscover/out
look/responseschema/2006a">
<Request>
<EmailAddress>aaaaa</EmailAddress>
<AcceptableResponseSchema>&xxe;</AcceptableResponseSche
ma>
</Request>
</Autodiscover>
```

Attack log 2:

```
GET /../../../../../../../../etc/passwd HTTP/1.1
Host: 159.XX.XX.XX
User-Agent: Mozilla/5.0 (X11; Linux x86_64)
AppleWebKit/537.36 (KHTML, like Gecko) Ubuntu
Chromium/80.0.3987.87 Chrome/80.0.3987.87 Safari/537.36
```

The attack data set can be regarded as a typical small sample. The sample can be regarded as a binary classification and multi classification problem in the whole log. Binary classification means to distinguish high threat logs and general logs from massive log samples; the multi classification problem is to further identify the attack label for the high threat log, which is similar to target recognition in image recognition, object recognition and object classification. Since most of high threat attacking behaviours cannot be recorded by protective equipment, human beings can learn from the learning process of known high threat log samples, and further draw inferences from one instance to identify whether the unmarked log is a high threat log. The semi supervised learning method can be used to train the model through these few labelled logs, and further migrate the training model to a large number of unlabelled samples for detection. This paper focuses on the multi classification of attack logs.

The data contains 335 pieces of data, belonging to 14 categories respectively, and the sample distribution of each

category shows a serious imbalance. Further analyse and process the metadata set, merge the samples with the same characteristics, and retain more diverse logs in each category, so as to obtain a better generalization data set. After processing, the data sets are distributed as Table 2:

Table 2. Data Set after Processing

Attack Trying Mode	Count
Code injection	43
XML injection	1
Command injection	4
Exploit vulnerability	33
File restrictions	14
Local file contains	5
Directory information disclosure	2
SQL injection	7
File injection	4
Web Shell	1
Malformed request	1
Java attack	1
Pseudo protocol	1
XSS attack	1

If we want to perform a log classification task of c-ways and k-shot (the data contains n categories and each category has k logs). For example, 5-ways and 2-shot classification means that there are only 2 logs under each classification. We can build many c-ways and k-shot tasks based on the data set, which will be used as the task source of Meta learning. The constructed tasks are divided into training task and test task. Each task contains its own training data and test data, which are called support set and query set respectively in Meta learning.

From the data set, there is only one sample under some classes. If we construct 5-way and 3-shot tasks, it is necessary to enhance the data set. The data set after data enhancement as Table 3:

Table 3. Data Set after Enhancement

Attack Trying Mode	Count
Code injection	43
XML injection	3
Command injection	4
Exploit vulnerability	33
File restrictions	14
Local file contains	5
Directory information disclosure	3
SQL injection	7
File injection	4
Web Shell	3
Malformed request	3
Java attack	3
Pseudo protocol	3

XSS attack	3
------------	---

Select 7 categories as the training set and 7 categories as the test set, and take the calculated mean of 10 experimental results as the final result. The results are compared with LSTM Meta learning model, model independent MAML Meta learning model, relational network and other mainstream methods. The results are as Table 4:

Table 4. Classification accuracy on 5-way 1-shot and 5-way 3-shot

Model	5-way 1-shot	5-way 3-shot
Meta-Learner	41.43%	59.38%
LSTM		
MAML	46.55%	61.29%
Relation Network	47.64%	64.06%
Ours(BiLSTM-Ridge)	49.72%	66.47%

V. CONCLUSION

This paper proposed a Meta learning classification method based on few-shot samples, called BiLSTM-Ridge regression model, which can recognize and learn the characteristics of high threat log samples through less sample learning, and has better generalization detection ability in new unknown test tasks. Specifically, we design a Meta knowledge learner in the Meta training scenario to extract the sample features of high threat logs and improve the adaptability of Meta learners in new tasks. We prove that our method achieves good results on the 5-way 1-shot and 5-way 3-shot problems on the real 335 high threat log data sets. Future work includes applying this model to other fields, including high threat sample detection and classification of unmarked logs, and exploring the migration learning of other massive log detection in the cloud environment.

ACKNOWLEDGMENT

This work has been supported by Information operation and maintenance project of CAS under Grant 2021000146, Youth Innovation Promotion Association CAS.

REFERENCES

- [1] Pan C, Huang J, Gong J, et al. Few-Shot Transfer Learning for Text Classification With Lightweight Word Embedding Based Models[J]. IEEE Access, 2019.
- [2] Aakriti Gupta, Kapil Thadani, et al. Effective few-shot classification with transfer learning. In Proceedings of the 28th International Conference on Computational Linguistics, 2020
- [3] Finn C, Abbeel P, Levine S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks[C]// International Conference on Machine Learning, August 6-11, 2017, Sydney, NSW, Australia. 2017:1856-1868.
- [4] Ravi S, Larochelle H. Optimization as a Model for Few-Shot Learning [C]// International Conference on Learning Representations, April 24-26, 2017, Toulon, France. 2019.

- [5] Munkhdalai T, Yu H. Meta networks[C]//International Conference on Machine Learning, August 6-11, 2017, Sydney, NSW, Australia. 2017: 3393-3943.
- [6] Koch G. Siamese neural networks for one-shot image recognition [D]. University of Toronto, 2015.
- [7] Vinyals O, Blundell C, Lillicrap T, et al. Matching networks for one shot learning[C]//Neural Information Processing Systems. December 5-10, 2016, Barcelona, Spain. 2016: 3630-3638.
- [8] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning[C]// Neural Information Processing Systems. December 4-9, 2017, Long Beach, CA, USA. 2017: 4077-4087.
- [9] Sung F, Yang Y, Zhang L, et al. Learning to compare: Relation network for few-shot learning[C]//IEEE Conference on Computer Vision and Pattern Recognition. June 18-22, 2018, Salt Lake City, UT, United States. 2018: 1199-1208.
- [10] Wang, Xuan, Xu, et al. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN[J]. Expert Systems with Application, 2017.
- [11] Neculoiu P, Versteegh M, Rotaru M. Learning Text Similarity with Siamese Recurrent Networks[C]//Proceedings of the 1st Workshop on Representation Learning for NLP.2016
- [12] Lample G , Ballesteros M , Subramanian S , et al. Neural Architectures for Named Entity Recognition[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016..
- [13] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Herve J ' egou, and Tom ' as Mikolov.2016. Fasttext.zip: Compressing text classification models. CoRR, abs/1612.03651.



Hanji Shen is currently a full-time engineer in Computer Network Information Center, Chinese Academy of Sciences, China. He received M.S. degree in Engineering in the field of Computer Technology from University of Chinese Academy of Sciences, and he is currently studying for a Ph.D. in University of Chinese Academy of Sciences, majoring in Computer Software and Theory. He has been an IEEE member since 2014. His research interest is Cyber Security.



Jun Li is currently a researcher, deputy chief engineer and doctoral supervisor in Computer Network Information Center, Chinese Academy of Sciences, China. He received Ph.D. degree in Computer System Structure from University of Chinese Academy of Sciences. He is one of the earliest experts engaged in computer network technology research in China. His research interests include Network Architecture and Cyber Security.



Xiang Fu is currently a full-time engineer in Beijing Blue Sky Frontier Science and Technology Innovation Center. He received a Doctor of Science degree in the field of computer science from Peking University. His research interest is artificial intelligence.