CHAPTER 5

Discussion And Future Works

5.1 Introduction

This chapter summarizes the results of the project on detecting traffic anomalies in IoT devices using the (UNSW) BoT-IoT dataset. Through a scientific experimental process, including data collection, exploratory data analysis, preprocessing, and the application of various machine learning algorithms, the dataset ultimately led to the development of a machine learning method that effectively detects traffic anomalies in IoT devices. Additionally, it briefly discusses the experimental directions for future projects and the potential improvements to enhance the quality and accuracy of the experiments, aiming to achieve more precise analysis. Therefore, this study adopts a structured approach from data processing to model evaluation, with the goal of making a significant contribution to the detection of traffic anomalies in IoT devices.

5.2 Summary

This study selected the UNSW BoT-IoT dataset, which is more suitable for this experimental dataset, through a comparative analysis of various advantages across four different datasets. The dataset was visually explored to provide detailed information, which serves as a basis for subsequent preprocessing. After preprocessing, including removing duplicate columns, missing values, and low-correlation columns, and converting data types, a dataset suitable for machine learning was obtained.

This project employs three methods—Logistic Regression, Support Vector Machine (SVC), and Naive Bayes—for experimentation. The experimental data shows that Naive Bayes outperforms Logistic Regression and SVC in the normal class samples (True Positive Rate (TP) = 100%, False Positive Rate (FP) = 0%). This indicates that among these three models, Naive Bayes is superior and is more suitable for detecting abnormal traffic in IoT devices compared to the other two models.

From the success of the project, we can draw the following conclusions：

1、Determination of data set: It is particularly important to select the data set suitable for the project for the experiment;

2、Exploratory data analysis: It is indispensable to conduct preliminary analysis and understanding of the data set. This step not only provides preliminary information for the project, but also provides a basis for subsequent data preprocessing;

3、Data preprocessing: This step is indispensable and particularly important. Reasonable and effective processing of the data set not only makes the experiment more accurate, but also makes the experiment more efficient.

4、Model selection: Logistic regression, SVC and naive Bayes are three classical machine learning classifiers, which have the advantages of easy interpretation, low implementation threshold and suitable for network security/intrusion detection tasks.

Overall, the project successfully achieved the goal of detecting abnormal traffic in IoT devices in a structured and data-driven manner. In addition, the project also proved that Logistic Regression is a relatively good detection machine learning model.

5.3 Future Works

While the project has provided many valuable insights, there are several areas that could be further developed to improve the quality of future analysis. Here are some suggestions:

1、Given the large size of the dataset, only a 5% random sample (2.93 million data points) was used for the experiment. This random sampling may introduce biases, leading to an imbalanced dataset. As a result, the model might tend to predict all samples as the majority class (the attack class), thereby reducing its ability to accurately predict normal class samples：a) In future studies, the dataset can be re-sampled, and if feasible, a larger dataset can be used; b) The classification threshold can be adjusted to enhance sensitivity to normal class samples; c) A weighted loss function can be employed, giving higher weight to minority classes during training;

2、This study uses the traditional model, but the deep learning-based model, such as the model using Lightweight GNN + Autoencoder, will better process the data and provide higher accuracy in the future;

3、Clarity of Models In data-driven decision-making, it is important to develop interpretable models. Further research could explore why the model makes certain predictions, so that the results are easier for policymakers to understand.

The aforementioned steps will provide room for further research in this project, to broaden its scope and enhance the accuracy and relevance of the results. The current project has already provided a viable method for detecting abnormal traffic

in IoT devices; thus, further developments are expected to have a more significant impact on this area in the future.