**Introduction**

With the rapid growth of the Internet, we need to send and receive massive traffic every day. Most of them will be regular traffic, while some show anomalous status; they may belong to network attacks, spam, or other abnormal behaviours. To identify and predict this traffic, we will try new methods. In this project, we focus on the research and analysis of anomalous traffic detection methods. Our research found two mainstream approaches nowadays to detect traffic: the clustering and the LSTM model. Based on this, we propose fusing two models and trying to get a more effective approach.

We will do a full introduction for our project in five parts. The first part is the introduction of the project. The second part shows the related works we read and referenced. The third part contains our approaches for the project, including the introduction of the model and data we used. The fourth part shows the evaluation, which means applying and mixing two models and getting a relatively higher accuracy result. The final part will conclude our work and what we will do for the future works

**Related work**

By reading a lot of papers, we found three articles as our reference for this project. The images of the three papers are shown below.

The first paper focuses on the K-means clustering algorithm. Their team classifies time interval and destination port number, which results in two-class clusters of normal and abnormal traffic. Their work has been a reference for our project; they found that data preprocessing is necessary and beneficial, so we changed the data type and selected 11 features and labels to process clustering. Their result shows that 15% of traffic data don't direct to the unusual destination port. This also gives us the possibility to use other models to optimise it.

文本

描述已自动生成

The second paper also employs the K-means clustering algorithm for anomalous traffic detection. Their team found that the clustering result may be unstable; this also allows us to optimise the clustering process to make sure it is convergent. And as a result, the false alarm rate remains at a relatively high level.

报纸上的文字

描述已自动生成

The third paper applies the LSTM algorithm to predict the traffic. Their team implemented the LSTM model while finding it difficult to go for a balanced dataset for training, which is also one of the issues we raised. Their result provides the guideline of the model, and by applying the LSTM model, the accuracy and recall rate can reach a high level.

文本

描述已自动生成

After we looked up related work on abnormal network traffic monitoring. We found that the accuracy of network traffic anomaly detection of a single model is already very high. If we want to make a breakthrough on accuracy rate, we must innovate in the method.

**Approach**

Our approach is to use two different models, clustering and LSTM. We try to make a fusion to let these two models forecast the data together to get a higher accuracy rate.

The clustering method is classified as unsupervised learning to mining the structure of the dataset. The image below shows how we apply the clustering method in our project code. We choose the K-means algorithm, select the k value for 5, and then train and predict.

文本

描述已自动生成

The LSTM model is seen as an improved version of the RNN model. RNN cannot cope with two far apart attributes because of its single register, and this problem is called long-term dependence. LSTM, as its name suggests, has long and short term memory capacity and can solve this problem. The image below shows how we use LSTM to create the training and testing dataset, construct the model, set up the optimiser, and train the model.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

**Evaluation**

Our understanding about the topic：When we were thinking about innovation, we looked up related work on abnormal network traffic monitoring. We found that in fact, the accuracy of network traffic anomaly detection of a single model is already very high. If you want to make a breakthrough in accuracy, you must innovate in the method.

We have done many tests on a single model and two models together. Finally, you will intuitively find that using one model alone for network anomaly detection will have a lower average accuracy rate than using the two models together.

文本

描述已自动生成Picture1(k-mean)

文本

描述已自动生成Picture2(LSTM Model)

文本

描述已自动生成

Picture3(integrated model)

We used a simple statistical concept to prove the feasibility of model fusion. If there are now 10 records, the probability of each record being correctly classified is 70%, or a model can classify these 10 records with 70% accuracy. Now fit three equivalent models. In the case of majority voting, for each record, the probability that the three models are correct is 0.7\*0.7\*0.7~=0.34, and the probability that the two models are correct is 0.7\*0.7 \*0.3\*3~=0.44, then if three models with an accuracy rate of 0.7 are used for fusion, in theory, the probability that each record can be classified correctly will increase to 0.78!

Model fusion is feasible and can improve accuracy .

**Conclusion**

Before we started work on our project, we have found several related works about LSTM model and K-mean algorithm. We have a clear understanding of the advantage of the two models. Then we prove the feasibility of our project by a statistic example. Finally, we compare the accuracy between just run LSTM model and run both LSTM model and k-mean algorithm. We found the integrated model have a better accuracy. In the future we may try to integrate different model into our project, and try different dataset to evaluate the detection performance.