Classification model for imbalanced traffic data based on secondary feature extraction

Jian Shen1, Jingbo Xia2, Yong Shan1, Zekun Wei1

1 School of Information and Navigation, Air Force Engineering University, Xian 710077, People’s Republic of China
2 Tan KahKee College, Xiamen University, Xiamen 363105, People’s Republic of China

Abstract: The non-equilibrium of network traffic data brings about the non-equilibrium of classification. Feature extraction is an effective method to reduce data dimensions, while it can intensify the influence of non-equilibrium further. A secondary feature extraction algorithm of multidimensional assessment is proposed in this study. The features of network traffic are evaluated in different dimensions to provide the basis for feature extraction. Furthermore, a model dealing with imbalanced data is proposed based on secondary feature extraction and sampling. The model combines the benefits of dimension reduction and redistribution. The experiment results show that the proposed model can not only increase classification accuracy and decrease non-equilibrium, but also enhance the performance of different classification algorithms.

1 Introduction

As the foundation of network cognition, management and optimising, network traffic classification is making a significant difference in resource scheduling, safety analysis and future tendency prediction [1]. The non-equilibrium and diversity properties of network traffic bring about the imbalance of traffic data [2]. This may mislead the classification result because of less training of minority categories and low classification accuracy [3–7]. However, the data of minority categories also has great importance, and the inaccurate classification of it may lead to serious consequences. As the pre-processing of traffic classification, feature extraction plays the role of dimension reduction, which can effectively meet the issue of data classification with nowadays explosively increasing of network traffic [8–13]. Even so, after feature extraction, the non-equilibrium of classification is not improved but gets worse because of the decrease of feature information quantity. Therefore, it is an essential challenge to reduce the non-equilibrium of network traffic classification on the base of feature extraction.

Due to the quantity and continuous increasing properties of internet traffic, it consumes tremendous time to process the data. To solve the problem, the strategy of secondary feature extraction divides database into several parts, so that the techniques of distributed computing and cloud computing can be taken into reality in the field of traffic classification. The strategy can compress the consuming time and enhance the real-time capability simultaneously. Gao et al. [14] proposed the divide–conquer and voting (DV) strategy, which divides database into several subsets to select features separately, and gathering the final feature subset through the means of voting. Shao et al. [15] put forward a feature extraction method combined with multiple algorithms to simplify classification model and reduce consuming time. Liu et al. [16] presented a method of iterative feature selection, which utilises some successful feature selection methods, such as information gain and $\chi^2$ statistics, to iteratively select features and perform clustering at the same time. Arthur et al. [17] came up with a new methodology for traffic identification through the combination of four classification algorithms.

In light of that, we suggest a secondary feature extraction algorithm of multidimensional assessment (MA). The index of classification capability of feature is defined to estimate every feature in multiple dimensions. With the defined index, the non-equilibrium property of traffic classification is decreased during the process of secondary feature extraction. Moreover, a model dealing with imbalanced data combines with secondary feature extraction and sampling (SFES) is proposed. The model redistributes database by sampling on the base of secondary feature extraction, and raises the classification performance further.

2 Non-equilibrium of network traffic

2.1 Non-equilibrium of network traffic distribution

Network traffic has the property of non-equilibrium in space distribution. Based on a series of research, it is found that 10% of the IP addresses contribute 90% of traffic [18]. Moreover, the businesses of these IP users are not fixed relatively, and the traffic of various businesses differs a lot. As a result, network traffic distributes imbalance in space.

Network traffic has the property of locality in time distribution, which is supposed to be one of the most important network traffic properties. Carey Williamson suggested that network traffic structure was not random and its distribution was affected by the business behaviour of users in application layer [19]. Jeffrey C. Mogul [20] pointed out the fact that most of the data in network belonged to the process that had just sent or received packages lately, and the property of locality still existed in microsecond range. On the other hand, the impulsiveness of traffic also leads to traffic gathering. Therefore, the same kind of traffic relatively concentrates in a specified period of time.

This paper applies Moore database into the experiment [21]. The database contains ten traffic categories, including WWW, MAIL, FTP-CONTROL, FTP-PASV, ATTACK, P2P, DATABASE, FTP-DATA, MULTIMEDIA and SERVICES. There are 115,521 flows with 248 features in the database. Moore database is divided equally into five subsets according to the arriving time of flows, so as to compare the proportion of ten categories in different subsets.

Fig. 1 shows the extreme non-equilibrium of traffic distribution. More than 90% of the traffic concentrates in one or two categories. However, the proportion of some categories is <0.5%. Besides, some minority categories concentrate in specified subsets, while seldom appear in others. The result shows evident gathering phenomenon and locality characters of traffic.
which is applied to secondary feature extraction of network traffic in this paper, provides fundamental basis for distributed computing and cloud computing to meet the challenges of big data and multi-services. The divide–conquer strategy parts the database $T$ into several subsets $\{T_1, T_2, \ldots, T_n\}$ first, and then extracts features of these subsets separately with the corresponding searching and estimating algorithms.

The distribution of minority categories is different in every data subset, so the features extracted from different subsets by the same algorithm differ a lot. From the analysis above, it can be seen that the distribution of minority categories is concentrated relatively in a certain range, and the divide–conquer strategy may increase the proportion of minority class in the specific subsets. As a result, the features extracted from subsets can be more representative for the certain minority categories.

### 3.2 Feature extraction and gathering

The paper combines best first (BF) searching algorithm and correlation based feature selection (CFS) estimating algorithm to select features for the first time, generating a series of feature subsets. Due to the fact that the feature gathering collection $F^*$ is assembled from feature subsets, there are still too many features in the collection, which is not the original purpose of feature extraction. So, the secondary feature extraction is urgently required to simplify the feature gathering collection and create the final representative feature subset to classify the database.

### 3.3 Secondary feature selection

Secondary feature extraction can simplify the statistical numeric characteristics and eliminate the redundant information. The classification capability is not only bound up with the information content of features, but also the features correlation within each other. Hence, the estimating index of classification capability of feature is proposed to multidimensionally assess every category of traffic. The capability of feature $f$ to classify a kind of data from the others is defined as follows:

$$I_d(f) = P^N(f) - P^{N-1}(f|f)$$

(1)

where $N$ is the total number of features in the feature gathering collection, $P^N(f)$ is the classification accuracy of a kind of data based on the feature gathering collection and $P^{N-1}(f|f)$ is the classification accuracy of a kind of data based on the feature gathering collection except for feature $f$. From formula (1), it can be seen that the classification capability of feature indicates the information content of feature that is not overlapped by the other features. This proposed index can avoid the computing of feature information content and correlation. The first selected feature should meet the requirement below:

$$f(1) = \arg \max_{f_1} \left\{ \sum_{a \in N} I_d(f_a) \right\} = \arg \max_{f_1} \left\{ \sum_{a \in N} [P^N(f) - P^{N-1}(f|f_a)] \right\} = \arg \min_{f_1} \left\{ \sum_{a \in N} [P^{N-1}(f|f_a)], n \in [1, N] \right\}$$

(2)

During the procedure of the following feature extraction, the point of extraction aims at the category that is classified most inaccurately, and that category is the target one to be improved. The database is classified through the selected feature collection $F$.

The classification accuracy of category $b$ is the lowest when $P_d(f) = \min \{ P_1(f), P_2(f), \ldots, P_M(f) \}$, where $M$ is the total number of categories. The feature whose capability of classification to category $b$ is the highest should be selected into the final feature collection. That means the selected feature should fulfill the requirement of the following formula:
Input: $T$ is the database of network traffic, $F$ is the feature collection.
Output: $F'$ is the final feature subset.

1) Divide database $T$ into $n$ subsets, $\{T_1, T_2, ..., T_n\}$
2) For $T_i \subseteq T$ do
3) $F_i = (BF + CFS)(T_i, F)$
4) end for
5) $F' = F_1 \cup F_2 \cup \cdots \cup F_n$
6) for $f' \in F'$ do
7) $P(i) = \sum_{x \in M} I_x(f'_i)$
8) end for
9) if $P(x) = \max\{\sum_{x \in M} I_x(f'_i)\}$ then
10) for $f'_i \in F'$ do
11) end for
12) if $P(F') = \min P_i(F'), m \in M$
13) if $I_x(f'_i) = \max I_x(f'_i)$
14) if $P(F') = \max I_x(f'_i)$
15) $F' = F' \cup f'_i$
16) end if
17) end if
18) end if
19) end for
20) end for
21) end if
22) return $F'$

**Fig. 3 Algorithm flow**

$$ f(x) = \arg \max_{f_s} \{I_x(f_a)\} $$

$$ = \arg \max_{f_s} \{[P^{N - x + (F - P^{-1}(F)|f_a)])] \} $$

$$ = \arg \min_{f_s} \{[P^{N - x + (F - P^{-1}(F)|f_a)])], n \in [1, N]\} $$(3)

From the analysis above, it is seen that the aim of feature extraction every time is to do the utmost to increase the classification accuracy of the category that is classified most inaccurately. However, there are still some problems that will interfere with the process of algorithm because of the correlationship and redundancy among different features. To solve these problems, a set of rules are established as follows:

**Rule 1**: If the classification accuracy of the target category does not rise with the increasing of feature number, choose the category whose classification accuracy is barely higher than the original one to be the new target category.

**Rule 2**: If the average classification accuracy of all the categories decreases with the increasing of feature number, choose the feature whose classification capability ranks only second to the original one to be the new target feature.

**Rule 3**: If the classification capability of several features is simultaneously to be the maximum, choose the feature whose average classification accuracy of all the categories is the highest to be the target feature.

### 3.4 Algorithm flow
See Fig. 3.

### 4 Model of SFES
From the analysis above, it can be seen that the proposed secondary feature extraction algorithm of MA can solve the problem of imbalanced classification to some degree. However, through secondary feature extraction, the non-equilibrium of data distribution still exists in space, and there is still room to improve the balance of classification. The model based on SFES is proposed in this paper to promote the classification accuracy of the minority categories by redistributing the database. The architecture of SFES model is shown in Fig. 4.

The conventional method to deal with imbalanced data is sampling. Under-sampling and over-sampling are two representative means. Under-sampling balances data of different categories by deleting the majority category samples, while over-sampling realises the same function by coping the minority category samples. However, over-sampling brings about additional training data which will extend the time of building classifiers, and the exact copy of sample may cause over-fitting. So under-sampling is the better choice [24], and it is adopted in this proposed model as the means to balance samples of different categories. SFES model combines the advantages of secondary feature extraction and under-sampling. On one hand, it reduces data dimensions substantially through secondary feature extraction. On the other hand, it balances data distribution in space and further decreases samples and training complexity.

### 5 Experiment and results

#### 5.1 Data pre-processing

The experiment divides database Moore equally into five subsets according to the order of arriving time, and extracts features through the means of BF&CFS. The feature extraction result is shown in Table 1. The numbers in Table 1 are the feature references.

Feature gathering collection is the summarisation of the above feature subsets, that is $\{4, 50, 51, 71, 72, 78, 81, 91, 108, 109, 119, 137, 148, 149, 152, 155, 180, 202\}$. The final feature subsets, which are secondary extracted from the feature gathering collection separately by MA, DV and BF&CFS algorithms, are shown in Table 2.

Then the experiment classifies different traffic categories with the feature subsets extracted by these three algorithms. The ten-fold cross-validation method is used to estimate the data. The result takes the average of the experiment which is conducted ten times.

#### 5.2 Performance of MA algorithm

The experiment adopts Naivebayes whose classification result is the most imbalanced to be the classifying algorithm, and then compares the accuracy changes of different feature extraction algorithms as the number of selected features increases. BF&CFS algorithm does not select features one by one, but select the final feature subset directly by estimating the value of feature collection. The number of features through BF&CFS algorithm does not change over time. So the experiment just compares MA algorithm and DV algorithm. The results of Moore database are shown in Figs. 5 and 6.

From the average value of different categories classification accuracy, we can see that the accuracy rises quickly along with the increasing of feature number. The classification accuracy of the two algorithms is the same when the first feature is extracted, because the two algorithms select the same first feature. In other circumstances, the average accuracy of MA algorithm is obviously higher than that of DV algorithm.
From the comparison of the variance of different categories classification accuracy, we can see that MA algorithm decreases the non-equilibrium of classification in general while DV algorithm does not. The variance of different categories classification accuracy of MA algorithm is less than that of DV algorithm, except for the situation where the number of extracted features is two. That is because two extracted features are not enough to label all the attributions of every category. The secondly selected feature does not only promote the category that is classified least accurately, but also all the others, and the promotion could be even more intensive. The result is that the non-equilibrium of classification is more obvious temporarily.

However, the non-equilibrium of classification declines quickly as the number of features increases.

To analyse the applicability of the proposed algorithm, another database NSL-KDD is applied to the experiment [25]. NSL-KDD is the database which was processed and stored by using approximately five million network connections at DARPA. There are 25,192 connections with 41 features in the database with binary labels which are NORMAL and ANOMALY. The final feature subsets extracted by the three algorithms are shown in Table 3.

The experiment results of NSL-KDD database are shown in Figs. 7 and 8. It can be seen that the advantage of MA algorithm is more obvious on NSL-KDD database. Only two features are extracted by MA algorithm, which have better classification performance on the indexes of accuracy and variance. With the increasing of the extracted features’ quantity, the average value of different categories classification accuracy of the features extracted by MA algorithm rises notably, and the variance decreases greatly. However, the performance of DV algorithm is only promoted obviously when the fourth feature is added to the feature subset. Although DV algorithm extracted more features to enhance the classification performance, it still cannot gain an advantage over MA algorithm. It is clear that the proposed MA algorithm can improve the classification accuracy and decrease the non-equilibrium.

The experiment compares the performance of the three algorithms mentioned above to observe the classification result of all the data categories in both of the databases. As shown in Fig. 9, MA algorithm has relative higher accuracy than the other two, and the accuracy of six categories by MA algorithm is the highest.
However, the classification accuracy of ATTACK category is quite low by whichever algorithm. The reason of that is the data amount of ATTACK is excessively rare, and the features which can label the ATTACK category effectively are not extracted into the feature gathering collection during the procedure of feature extraction at the first time. As a result, no algorithm classifies this category accurately.

Table 4 shows the comparison of the collected performance. It is clear that MA algorithm has the highest average value and lowest variance of different categories classification accuracy on both of the databases, which means the proposed algorithm can not only promote classification capability but also decrease the non-equilibrium. Meanwhile, MA algorithm has broader coverage of cases on both of the databases, which demonstrates that the rules of the algorithm are more effective.

5.3 Performance of SFES model

The experiment applies the three above-mentioned algorithms into SFES model, and three new algorithms of SFES BF&CFS, SFES DV and SFES MA are generated. The comparison of performance between the new algorithms and the original ones is shown in Fig. 10. It can be seen that all the average accuracy value of the three algorithms is promoted, and the variance is decreased at the same time. That means SFES model can not only improve the classification capability, but also decrease the non-equilibrium obviously.

Receiver operating characteristic (ROC) curve is the method of patterning to show the true positive rate and false positive rate of classification model. The closer the curve approaches to the top left corner, the better the performance of the classifier is. To quantitatively assess the classification model, the area under the curve (AUC) of ROC is adopted as the evaluating index. Obviously, the value of AUC ranges from 0 to 1, and the bigger the value is the better the performance of the classifier is. From Fig. 11, it can be seen clearly that SFES_BF&CFS performs much better than BF&CFS. However, the other curves of ROC are very close to that of the original ones. Thus, they should be analysed quantitatively. The AUC values of all the algorithms are shown in Table 5. It can be concluded that SFES model can improve the performance of different algorithms in varying degrees.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Final feature subsets extracted from NSL_KDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order</td>
<td>MA</td>
</tr>
<tr>
<td>1</td>
<td>flag</td>
</tr>
<tr>
<td>2</td>
<td>dst_host_srv_diff_host_rate</td>
</tr>
<tr>
<td>3</td>
<td>logged_in</td>
</tr>
<tr>
<td>4</td>
<td>srv_serror_rate</td>
</tr>
<tr>
<td>5</td>
<td>diff_srv_rate</td>
</tr>
<tr>
<td>6</td>
<td>dst_host_srv_diff_host_rate</td>
</tr>
<tr>
<td>7</td>
<td>flag</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4</th>
<th>Comparison of performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BF&amp;CFS</td>
</tr>
<tr>
<td>average value of accuracy</td>
<td>0.4386</td>
</tr>
<tr>
<td>variance of accuracy</td>
<td>0.3782</td>
</tr>
<tr>
<td>coverage of cases</td>
<td>0.96656</td>
</tr>
</tbody>
</table>
6 Conclusion

To decrease the non-equilibrium of traffic classification, we put forward MA algorithm and SFES model based on secondary feature extraction. The proposed MA algorithm divides database into several subsets and improves the performance of secondary feature extraction by estimating every feature separately in the dimensions of categories. Meanwhile, SFES model, which is based on SFES, combines divide-conquer and dimension reduction properties of secondary feature extraction with the function of redistributing the imbalanced data of sampling. The experiment shows that SFES model can reduce complexity and non-equilibrium of classification algorithm at the same time, and promote the classification capability in general. Our work in this paper is mainly focused on increasing the equilibrium of

![Fig. 10 Performance of all the algorithms](image)

![Fig. 11 ROC curves of all the algorithms](image)

<table>
<thead>
<tr>
<th>Table 5</th>
<th>BF&amp;CFS</th>
<th>DV</th>
<th>MA</th>
</tr>
</thead>
<tbody>
<tr>
<td>original model</td>
<td>0.869667</td>
<td>0.977211</td>
<td>0.98226</td>
</tr>
<tr>
<td>SFES model</td>
<td>0.981608</td>
<td>0.979689</td>
<td>0.98371</td>
</tr>
</tbody>
</table>
classification of all the categories and promoting the classification performance. As future work, it is possible to research the feature extraction methods to identify the specific categories for the purposes of network management and safety protection.

7 References


