Online Network Traffic Classification Algorithm Based on RVM

Zhang Qunhui
Hunan Information Science Vocational College, Changsha, Hunan, 410151, China

Abstract—Since compared with the Support Vector Machine (SVM), the Relevance Vector Machine (RVM) not only has the advantage of avoiding the over-learn which is the characteristic of the SVM, but also greatly reduces the amount of computation of the kernel function and avoids the defects of the SVM that the scarcity is not strong, the large amount of calculation as well as the kernel function must satisfy the Mercer’s condition and that human empirically determined parameters, so we proposed a new online traffic classification algorithm base on the RVM for this purpose. Through the analysis of the basic principles of RVM and the steps of the modeling, we made use of the training traffic classification model of the RVM to identify the network traffic in the real time through this model and the “port number+ DPI”. When the RVM predicts that the probability is in the query interval, we jointly used the "port number" and "DPI". Finally, we made a detailed experimental validation which shows that: compared with the Support Vector Machine (SVM) network traffic classification algorithm, this algorithm can achieve the online network traffic classification, and the classification prediction probability is greatly improved.

Key Words—RVM Modeling, Deep Packet Inspection, Normalization, Predict the Probability

I. INTRODUCTION

The Internet traffic classification is not only the basis of network management and network security, but also an important way to understand the behavior of Internet users and improve the network quality of service [1-2]. In recent years, the traffic classification algorithms based on the machine learning has become a hot research at home and abroad, and such algorithm does not need to detect the load on the application layer, which can protects user’s privacy and classifies the encrypted traffic. Most of the current results use the classification model that has the supervised learning function, which needs a large numbers of training samples with accurate marks to train the model [3-6]. With the escalation of the application protocols and the widely use of the encryption technology, the accurate mark becomes very difficult [7-9].

The real-time traffic classification techniques can classify the online network traffic according to the type of application, which is of great significance for network management, traffic control and the relevant network-related researches [10-13]. Currently, there are four main traffic identification and classification techniques: port identification technology, Deep Packet Inspection (DPI), network behavior-based traffic classification technology and machine learning-based traffic classification technology. In the above methods, in addition to the network behavior-based traffic classification technology, the other methods can all be applied to the classification of the online network traffic, but there are still many inadequacies. With the emergence of the port hopping technology and the camouflage technology, the port identification technology is no longer reliable; DPI needs to depth test each packet so the time and space complexity is higher, which makes it difficult to meet the real-time requirement in the high-speed network; the traffic classification method based on machine learning has the advantage of high classification accuracy, but it needs to be pre-train the classification model and the test time is relatively longer, so it is difficult to be competent the online classification task in the high-speed network [14-17]. As the continually development of the Internet, many new network services (such as P2P, online games, etc.) adopt the dynamic port, the protocol encryption as well as other actions, all of the above actions make the traditional port-based and payload-based traffic classification method can not guarantee the totally precise classification and statistics of the network traffic. In recent years, some scholars have been used the machine-learning method to do the traffic classification research. To classify the network traffic makes the idea of using machine to learn the techniques to be first proposed in the research of intrusion detection; Zander S and others proposed an unsupervised machine-learning method to identify the frameworks of different network applications, in which they used the statistical properties of the flow as the flow characteristics to do the research of automatic classification of network traffic; MOORE AW and others used the Bayesian-based classifier and a large numbers of flow characteristics to do such research. Relevance Vector Machine (RVM) is a new kind of machine learning method which has the same decision form with SVM. By introducing the sparse Bayesian learning theory, it not only has the advantage of avoiding the over-learn which is the characteristic of the SVM, but also avoids the defects of the SVM that the scarcity is not strong, the large amount of calculation as well as the kernel function must satisfy the Mercer's condition and that human empirically determined parameters. Therefore, compared with SVM, RVM is sparser, has shorter test time, and is more suitable for the online classifieds. In addition, RVM uses the probability model to explain the noise in the data, which makes it possess the ability to predict the probability of a result, and to further understand the classification result by the probability

Abbreviations

RVM: Relevance Vector Machine
SVM: Support Vector Machine
DPI: Deep Packet Inspection

forecast. Currently, RVM has been widely used in voice recognition, hyper spectral image classification and other fields and has achieved good classification results.

Based on the above views, this paper applies the RVM into the online network traffic classification. By introducing the concept of query interval, we studied and analyzed the impact of the prediction of the probability on the classification results, based on which, we proposed a new RVM-based online network traffic classification algorithm. This algorithm takes the sub-traffics consist of several data packets at the beginning stage of the online traffic as the research object and adopt the mixed mechanism of the combination of RVM and “port number + DPI” to identify the network traffic online, which has the features of high classification accuracy and fast processing speed.

II. PROPOSED SCHEME

A. Classification Of Online Network Traffic

For the sake of ease comparison and analysis, this thesis uses two sets of experimental data. The first one is the experimental data set-Moore_Set used by Moore AW et al, which is mainly used for the description of the query interval; the second one is collected from the Network Engineering Laboratory in Hunan University, which is recorded as IN_Set and mainly used to validate the online classification algorithm.

Due to Moore_Set takes the complete TCP traffic as the traffic sample, ignoring a large number of UDP packets existed in the network, especially the P2P applications account for much broadband, whose UDP packets even account for about 60% of P2P packet. Therefore, in this thesis, we adopt the TCP / UDP bi-directional network traffic as the traffic sample according to the quintuple and take an interval T (T = 120s) as the end flag of the stream, and we also collect the traffic characteristics according to the 44 types of network traffic characteristics defined in NetMate to generate the dataset IN_Set.

The IN_Set dataset contains two data sets, the first one is captured on June 22, 2011 from 14:30 to 17:30, recorded as IN_Set1; the second one is captured on October 7, 2012 from 14:30~17:30, recorded as IN_Set2. Two data sets both adopt the method of combining the port number, DPI technology and manual processing to apply identify the data sets, for the traffic can not be identified, they mark them as the unknown traffic. The experimental data does not include the unknown traffic. The statistical information of the data sets are shown in Table 1 and Table 2.

<table>
<thead>
<tr>
<th>Type of flow</th>
<th>Num of flow</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>WWW</td>
<td>70312</td>
<td>48.03%</td>
</tr>
<tr>
<td>FTP</td>
<td>9045</td>
<td>6.39%</td>
</tr>
<tr>
<td>MAIL</td>
<td>7392</td>
<td>5.28%</td>
</tr>
<tr>
<td>DataBase</td>
<td>5589</td>
<td>3.98%</td>
</tr>
<tr>
<td>P2P</td>
<td>40078</td>
<td>28.05%</td>
</tr>
<tr>
<td>GAME</td>
<td>937</td>
<td>0.76%</td>
</tr>
<tr>
<td>SERVER</td>
<td>471</td>
<td>0.35%</td>
</tr>
<tr>
<td>Unknown</td>
<td>11402</td>
<td>0.19%</td>
</tr>
<tr>
<td>Total</td>
<td>13945</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table II. The statistics of the IN_Set2 Dataset

<table>
<thead>
<tr>
<th>Type of flow</th>
<th>Num of flow</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>WWW</td>
<td>72314</td>
<td>49.09%</td>
</tr>
<tr>
<td>FTP</td>
<td>8239</td>
<td>5.68%</td>
</tr>
<tr>
<td>MAIL</td>
<td>7193</td>
<td>4.94%</td>
</tr>
<tr>
<td>DataBase</td>
<td>6501</td>
<td>4.57%</td>
</tr>
<tr>
<td>P2P</td>
<td>39854</td>
<td>25.32%</td>
</tr>
<tr>
<td>GAME</td>
<td>1086</td>
<td>0.71%</td>
</tr>
<tr>
<td>SERVER</td>
<td>471</td>
<td>0.35%</td>
</tr>
<tr>
<td>Unknown</td>
<td>12409</td>
<td>8.70%</td>
</tr>
<tr>
<td>Total</td>
<td>14923</td>
<td>100%</td>
</tr>
</tbody>
</table>

B. Query Interval

For the binary classification, RVM can not only get the binary output, but also the prediction probability of the results. The more prediction probability is close to 0 or 1, the higher the accuracy of its prediction will be. For the prediction value of a certain interval, the prediction accuracy decreases significantly and there is a big uncertainty in the classification results, and the accuracy of the classification is doubtful, so it can be called a query interval.

In order to research the scope of query interval and its impact on the classification accuracy, the experiments in this thesis all take the Moore_Set, and IN_Set data sets as the experimental subjects. In the first set of experiments, we regard the entry01 as the training set, and the remaining nine data sets as the test sets; in the second set of experiments, we respectively divided the IN_Set1 and IN_Set2 into two groups, the former group is as the training set, and the latter group as the test set. We use the CSF algorithm to select the traffic characteristics and randomly sample the training sets to create a classification model. The two sets of experimental results both take the mean value as shown in Figure 1.
the ordinate represents the misclassification probability on the corresponding prediction results probability. As can be see from Figure 1, when the classification results prediction probability $p$ is in the interval $[0, 0.1]$, the is the lowest (in the Moore Set, the mis-prediction probability is only 0.3%, and in IN Set is 0.6% ), which means the classification accuracy is the highest; in the $[0.9, 1]$ interval, the mis-prediction probability is relatively low (in the Moore Set , the mis-prediction probability is 4.85%, in IN Set is 3.74% in ).In other probability intervals, the prediction probability is significantly increased. In the Moore Set data, there are three intervals whose mis-prediction probability have been more than 40%; in the IN Set data, there are four, in which the error rate of the $[0.3, 0.4], [0.6, 0.7]$ and $[0.7, 0.8]$ intervals are above 50%. Under the normal circumstances, when the error rate is about 50%, the prediction correctness is equal to the guess, which means this prediction can be seen as invalid.

The further research on the distribution of the classification error rate on the prediction probability as shown in Figure 1 finds that in the above two experiments, the classification results whose prediction probability is in the $[0, 0.1]$ and $[0.9, 1]$ intervals accounts for 90.47% (Moore Set) or 94.58% (IN Set) of the total number of samples and the classification accuracy in these two intervals is up to 98.79% (Moore Set) or 98.02% (IN Set). In addition, the classification accuracy in the query interval $[0.1, 0.9]$ is only 65.72% (Moore Set) or 40.12% (IN Set), which is far to meet the demand of the classification accuracy. Therefore, we define the interval $[0.1, 0.9]$ as the query interval.

C. Classification Algorithm

Based on the above characteristics, in order to obtain a higher classification accuracy, this thesis proposes a new mixed traffic classification algorithm, whose flow chart is shown in Figure 2. The algorithm consists of two parts of the offline and online, in which the former is based on the offline packet to select the appropriate stream characteristics and matching characteristics and using the RVM training traffic classification model; the latter online capture the packets to form a network stream, through the RVM classification model and “port number + DPI ” to identify the network stream in the real-time. When the RVM prediction probability is in the query interval, jointly use the “Port Number” and “DPI” to improve the classification accuracy. Some of the online steps of the algorithm are as follows:

a. Capture the data packet online to form the network stream and extract the characteristics of the sub-stream, and store the first 20 packets of the network stream into the off-line packet memory;

b. Store the selected network sub-stream into the cache list, and after the cache meets a certain time limit, input the network sub-stream in the list to the RVM classification model, then clear the cache;

c. The RVM classification model outputs the prediction probability and judges if the prediction probability is in the query interval $[0.1, 0.9]$; if it is, then execute according to step 4; if not, completely adopt the classification results;

d. If the network sub-stream port number is in the commonly used port mapping table and there is no corresponding feature field in the matching feature library, we can mark the apply type according to the port number; if there is , we can mark the apply type according to the feature field ; otherwise mark it as unknown.

The abstract representation of network traffic: in an interval, there exist the same source IP, destination IP, source port, destination port, and the network packet sequence in transport layer protocol , which is known as network traffic, and these five attributes are also known as the quintuple. The stream statistics feature is the statistical characteristics based on the units of the stream, including the expectation and variance of the size of the packet of the stream and the interval of the packet reaching time.

Here is a known network stream i.e. Marked sample sets $x = \{x_1, x_2, \ldots, x_n\}$ in which $x_i = (x_i, x_{i,1}, x_{i,2}, \ldots, x_{i,n})$, $x_{ij}$ is the $j$th statistical characteristics of the network stream. The corresponding protocol category of the $n$ network stream samples is $\{c_1, c_2, \ldots, c_1, c_n\}$, the value of $c^*\in \{c_1, c_2, \ldots, c_1, c_n\}$. When the value of $c^*_i$ is the same as $c^*_j$, it represents that the two network stream protocols are the same. The aim of the traffic classification is to train the classification model $f : x \rightarrow c$ and to judge the protocol class $c_i$ of the network stream $x$.

III. EXPERIMENTAL RESULTS

A. Data Preprocessing

The real-time traffic classification technology research aims at finding a method to identify and classify the network traffic apply type online. The traffic characteristics proposed in the Netmate will not be fully applicable to the real-time traffic classification, because some characteristics cannot be obtained by adding up the first $n$ packets in the network stream. Table 3 lists the
main traffic characteristics that can be used for real-time traffic classification, whose biggest feature is that it can be obtained through the online calculation at any time, and the time and space complexity are low.

Screen the traffic characteristics of the data set by using the method of CFS (Correlation-Based Feature Selection). Select 5 traffic characteristics which have the strongest correlation and the weakest cross-correlation, as shown in Table 4:

<table>
<thead>
<tr>
<th>TABLE 4. THE FEATURE SUBSET SELECTED BY THE CFS ALGORITHM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHARACTER</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

From the previous experiments, we found that if directly using the raw data for training and classification, the classification efficiency and the accuracy will be low, and the related vectors will be relatively more. There are two main reasons. Firstly, in the statistical process of the traffic characteristics, different characteristics adopt different units of measurement, and there are big differences in the size of each variable, which means if directly using the raw data for the classification, it may lead to a loss of information and cause the instability of numerical calculation. Secondly, most of the traffic sample data are excessive concentrated, while the degree of deviation of a small amount of data is too big, which results in the RVM is difficult to produce effective related vectors to distinguish between different types of data. Therefore, we need to preprocess the traffic characteristics vector.

First of all, consider the linear normalization method to normalize it, but the experimental result shows that this method is not an effective solution to the problem that the deviation degree of a small amount of data is too big. However, use a logarithmic function to do the normalization, the zero element and the minimum elements in the data will make the degree of deviation between the data much bigger. Therefore, to reduce the degree of deviation of the data and distribute the data in the samples as reasonable as we can, we use the following method:

\[
Y_{(k+1)} = \frac{1}{(k+1)} \sum_{i=0}^{k} f(e_{i1}, e_{i2}, \ldots, e_{iM}) e_{i1} e_{i2} \ldots e_{ihk+1}
\]

\[
, 0 < a < 1, i = 1, 2, \ldots, k + 1
\]

\[
x_g = (x_1, x_2, \ldots, x_g, \ldots, x_m)
\]

\[
y(c_i | \beta_i) = \sum_{j=0}^{n} \pi(c_i | \mu_i, \sum j) \quad (8)
\]

\[
y = \left( n, \frac{x - x_{\min}}{x_{\max} - x_{\min}} \right)^t \quad (9)
\]

Normalize the data. The experiment shows that when \(n=13\), we can get the satisfactory classification accuracy, meanwhile the number of the relevant vectors is acceptable.

B. Result

For a certain type of samples, the recall rate is recorded as:

\[
f(e_1, e_2, \ldots, e_M) = f(0, 0, 0, \ldots, 0)
\]

\[
+ \sum_{i=0}^{H} f_i (0, 0, 0, \ldots) R_i + \frac{1}{2} \sum_{i=0}^{H} f_{i,j} (0, 0, 0, \ldots) R_i R_j + \ldots
\]

\[
+ \sum_{i=0}^{H} f_{i,j} \sum_{i=0}^{H} \sum_{i=0, j, k \neq i} (0, 0, 0, \ldots) R_{i} R_{j} R_{k} R_{i} + P_{i, j, k, l}
\]

\[
recall_i = \frac{TR_i}{TR_i + FN_i}
\]

Recall all the samples, the overall accuracy is:

\[
E^2 \left( \frac{S^2_{\alpha} (f)}{var \left( \frac{S^2_{\alpha} (f)}{H} \right)} \right) \approx TR \cdot TB \cdot C_{\alpha}^2 (f)
\]

\[
s = \sum_{i=1}^{m} \sum_{i=1}^{n} \frac{TR_i}{TR_i + FN_i}
\]

\[
overall accuracy = \frac{\sum_{i=1}^{m} \sum_{i=1}^{n} TR_i}{\sum_{i=1}^{m} (TR_i + FN_i)}
\]

The overall accuracy has been widely adopted by relevant researchers, and it reflects the proportion of the classification model correctly predicted number of samples in the total number of the prediction. In this thesis, we mainly adopt the overall accuracy as an evaluation tool.

Bernaille pointed out that if use the traffic characteristics of the first \(m\) packages of the network stream, we can achieve the goal that distinguish the
traffic applied model. In order to select the appropriate m, this thesis adopts the IN–Set data sets as the experimental subject and divides IN_Set1a and IN_Set2 into four groups of IN_Set1a, IN_Set1b, IN_Set2a and IN_Set2b, in which IN_Set1a and IN_Set2a are as the training sets, IN_Set1b and IN_Set2b are as the corresponding test sets and carries out two sets of experiments. Take m = 10, 11, ..., 20, extract the first m packages to generate a stream sample, then randomly sample the training sets and establish the classification model. The two sets of experimental results all take the mean value. The experiment result is show in Figure 3:

![Figure 3. Classification accuracy and the confident accuracy](image)

In the figure, “Overall Accuracy” represents the overall classification accuracy by using the RVM classification network sub-stream, “Confident Accuracy” represents the confident accuracies of those whose classification results prediction probabilities is in the interval of [0,0.1] and [0.9,1]. We can see from the figure 3 that the overall classification accuracy steadily rises, and when m=17, the highest classification accuracy is 96.12%, then it slightly declines to stable. The confident accuracy trembles and rises, and when m=18, it reaches its highest value-98.25%. In order to get a higher classification accuracy in the following mixed classification algorithm, we select m=18 to be the packet of the network sub-stream in this thesis.

Classify the experimental data IN_Set1 and IN_Set2 by using the online traffic classification algorithm proposed in Section C, and the classification accuracies are shown in Table5 respectively.

From the Table5, we can find that the algorithm has achieved a good classification result with the overall classification accuracy of about 98%. Besides, the identification accuracies of the applied traffic which has a bigger proportion in the data are relatively higher, which are about 98%. In which the WWW applied traffic is as high as 99%; on the contrary, the identification accuracies of the applied traffic which has a smaller proportion in the experimental data are relatively lower, in which the identification data of GAME is lower than 80%. The reasons are: firstly, in traffic classification accuracy of this part in the RVM traffic classification process is not high; secondly, the GAME traffic is composed of the UDP packet which is difficult to identify by the method of “port number+ DPI”.

<table>
<thead>
<tr>
<th>data set</th>
<th>FTP</th>
<th>www</th>
<th>MAIL</th>
<th>P2P</th>
<th>Game</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN_Set1</td>
<td>98.1</td>
<td>99.0</td>
<td>94.8</td>
<td>98.4</td>
<td>78.9</td>
<td>98.0</td>
</tr>
<tr>
<td>IN_Set2</td>
<td>96.0</td>
<td>98.8</td>
<td>96.3</td>
<td>98.3</td>
<td>70.0</td>
<td>97.8</td>
</tr>
</tbody>
</table>

IV. CONCLUSION

RVM not only can provide binary output, but also can provide the probability prediction. This thesis analyzes the prediction probability of the classification results and its influence on the classification accuracy. We found that the classification accuracy in the query interval (in this experiment is [0.1, 0.9]) is difficult to meet the accuracy requirements of network traffic classification. Therefore, this thesis proposes a hybrid online network traffic classification algorithm based on the RVM prediction probability. The algorithm captures online 18 packets as the network sub-stream at the beginning of the network traffic, and uses the classification model trained by RVM to classify the sub-stream. If the output prediction probability is in the query interval [0.1, 0.9], re-identify by using the combination of the “port number+ DPI ”, if or, fully adopt. The experiment shows that this algorithm can achieve the online network traffic classification, and the classification accuracy is relatively higher.

REFERENCES