THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Classification of Traffic using Supervised Learning Approach and its Building Time

P. Pinky
Assistant Professor, Department of Information Technology
SNS College of Engineering, Coimbatore, Tamil Nadu, India

Abstract:
In a large network, it is a tedious process to classify the traffic. Classify the traffic helps to provide Quality of Service(QoS) to the application and is also used to detect intruder. Inorder to provide best service to the user, the Machine learning techniques are used to prioritize the traffic or to classify the traffic. The machine learning techniques are of two types supervised and unsupervised method. This paper explains the review of supervised learning method.

Key works: *Machine Learning, Supervised Learning, Traffic classification*

1. Introduction

In Networks all the systems are connected together and the service will be provided by means of ISP. The data transmission in the network is administered by end-to-end transmission protocols such as TCP and UDP without network monitoring, auditing and control over the traffic. Because of this, unauthorized traffic may pass through the network. Hence IP traffic classification comes to a part to detect the traffic and classify it based on the application. The classification of traffic is mainly useful for automated intrusion detection; identifying patterns of denial of service attacks etc. So, real-time traffic classification has the potential to solve many network management problems. In earlier days the IP traffic classification is based on the deep packet inspection of each packet's contents. The deep packet inspection mechanisms use signature analysis to understand and verify different applications. Signatures are distinctive and they are related with every application [1, 2]. The classification engine then evaluates the traffic next to this reference to make out the exact application. Regular updates are requisite to maintain current with new applications when the payload format changes.

But the payload based inspection is unworkable when the payload is encrypted and the signature often changes for every application. So it requires more memory to store the reference database for each application's unique characteristics. Port based classification was also widely used in traffic classification. This approach is based on the association between the transport layer's port number and the corresponding application. Historically, many applications use well-known port on their local host as a meeting point to which other hosts may start the communication. However, this approach has limitations. Firstly, some applications may use ports other than its well known ports and some applications may not have the IANA registered ports. Also in some cases server ports are dynamically allocated as needed. So port based and payload based techniques are fading in these days [3]. The auspicious tactic that has lately accepted is machine learning techniques.

2. Machine Learning

In this paper, we present the machine learning technique as a best method for traffic classification. Machine learning [3] is used to relate flow instances into distinct classes of network traffic. Every flower is labelled by an array of statistical features and a related feature value. The statistical features such as mean packet length, mean packet size, inter packet length etc computed over numerous packets. Each feature reveals discrete values of the feature which is dependent on the class of traffic in the network to which it fit in.

The ML algorithms are divided into two categories called supervised and unsupervised. Unsupervised algorithms [4] collect flow of traffic into distinct clusters based on the identical values of feature. Thos algorithm does not have a capacity of prior learning of exact class of traffic. In supervised learning [5] the traffic class must be find out in advance. The classification design that has been created by means of training instances has capable to envisage the recent hidden instances by seeing the feature values of anonymous flows. Thus machine learning algorithms are greatly useful in IP traffic classification and in identification using statistical features. It can also make use of supervised and unsupervised learning algorithms to classify the known traffic flows and anonymous traffic flows.

3. Supervised Learning

In Supervised learning, the class of traffic must be identified before it gets to be classified. The classification model that has been built using the training set of instances can able to predict the new instances by probing the feature values of unknown flows. The supervised learning uses weka to implement the algorithms. These algorithms' performance is calculated in terms of classification speed and the model building time.

3.1. Bayesian Network

Bayesian Network [6] is one of the supervised techniques used to classify the traffic. Bayesian Network is otherwise called as Belief Networks or Causal Probabilistic Networks. It depends on a Bayesian Theorem of probability theory to generate information between nodes and it gives the relationship between nodes even if the nodes are ambiguous. It is a graphical based probabilistic model that signifies randomvariables and their conditional probabilities.

Bayesian Network is composed of a directed acyclic graph of nodes that represent features or classes and links that represents the relationship between nodes. It also includes set of conditional probability tables which determines the strength of the links. Each node has a probability table that defines the probability distribution for the node if it has parent nodes. The probability distribution is unconditional when the node has no parents and conditional if it has one or more parents.

Bayesian Network makes easy to study about the causal relationship between variables. In clear, the causal relationship is explained as follows: the node A and B is connected to the node C by means of links. So the node A and B is the parent node of C and C is the child node. The parent node of A and B symbolizes the causal factors of the node C. The conditional probability between the parent node and the child node is represented by the conditional probability tables. Bayesian Network is difficult to explain the conditional probability tables.

3.2. C4.5 Decision Tree

C4.5 Decision Tree algorithm [7] creates a tree structured model where the nodes in the tree represent features and the branches represents values which connects features. A leaf node represents the class which terminates nodes and branches. The decision tree has been built with the root as a starting point and continuous down to its leaves. To classify the object, we begin at the root of the tree then compute the test and proceeds towards the branch which yields a suitable outcome. This process prolongs until the leaf is met. If a class named by the leaf is identified then the object belongs to that class. The class instance can be determined by examining the path from nodes and branches to the terminating leaf. If all classes of an instance belong to the same class then the leaf node is labeled with that class. Otherwise the decision tree algorithm uses divide and conquer method which is used to divide the training instance set into non-trivial partitions until every leaf contain instances of only one class or until further partition is not possible.

The decision tree algorithm is steadfast to classify and it is understandable because the data is split into nodes and branches. C4.5 is one of the most accurate classifiers and fastest classification speed.

3.3. Naïve Bayes Tree

The Naïve Bayes Tree (NBTree) [8] is a combination of Decision Tree and Naïve Bayes classifier. The NBTree algorithm is labeled as a decision tree which has nodes and branches and it is also defined as a Bayes classifier on the leaf nodes. The accuracy of both Naïve Bayes and decision tree are not good enough. This paper has shown that the NBTree algorithm is more accurate than C4.5 or Naïve Bayes on certain datasets. Like most other tree based classifiers NBTree also has branches and nodes. The algorithm in [8] is mainly concerned with evaluating the utility of a split of each attribute. Utility for a particular node is calculated by making the data discrete and using the method called 5-fold cross validation for which Naïve Bayes is used to estimate the accuracy. The weighted sum of the utility of the nodes is considered as a utility of the split which is considered significant if there are at least 30 instances for the node and the relative error minimization is greater than five percent. The instances are divided based on the highest utility among all the attributes, if it is better than current node utility. If there is no such better utility a Naïve Bayes classifier is created for the current node.

3.4. Naïve Bayes

The Bayesian theorem is the basis of Naïve Bayes algorithm [9] and this method is based on probabilistic knowledge. The Naïve Bayes classifier takes a sign from unrelated attributes to rustle up final prediction to classify the attributes. The Naïve Bayes classification uses Bayes rule to evaluate the conditional probability by examining the association between each attribute value and the class [5].

$$\boxed{P(A|B)} = \frac{P(B|A)P(A)}{P(B)}$$

Where A is class and B is fixed attribute value. To get the probability of an object which belongs to class A using these conditional probabilities multiplied together. Naive Bayes classifiers calculate the probabilities of a feature which is having a feature value. The frequency distribution cannot calculate the probability of a continuous feature if they have a large number of values. Instead it can be achieved by modeling features for the continuous probability distribution or discrete values. The classifier's performance was going down when it's considered full flow features. Instead, sub-flows are used to enhance the performance.

Algorithm	Feature	Model Building Time
Bayesian Network [6]	• Packet length (min, mean, max, std deviation) • Inter-Arrival time (min, max, mean, std deviation)	less
C4.5 Decision Tree [7]	• Statistical features such as Packet Length, Inter-Packet Length and Inter Packet Arrival time.	less
Naïve Bayes Tree [8]	• Inter-Packet Arrival time for both the direction • Packet length for both the direction • calculate statistics for all features	more
Naïve Bayes [9]	• Inter-Packet Arrival time for both the direction • Inter packet length variation for both the direction • IP packet length • calculate statistics for all features	less

Table 1: Summarized Results of Supervised Learning Algorithm

S.No	Algorithm	Building Time
1	Naïve Bayes	0.03
2	C4.5	0.02
3	BayesNet	0.04
4	Naïve Bayes Tree	0.9

Table 2: Building Time of Supervised Learning Algorithm

4. Conclusion

The traffic classification acts as a crucial part to categorize the traffic. So far the port based method and payload based method were used but it was not up to the mark to categorize the traffic in a large network . Hence Machine Learning techniques were used to classify it. In this paper gives a review of one of the machine learning technique (i.e) supervised method. This method can able to find the new instances in the unknown flows by using the algorithms. The supervised learning algorithm's efficiency was calculated in terms of model building time.

5. References

1. Patrick Schneider. TCP/IP Traffic Classification Based on Port Numbers. Division Of Applied Sciences, Cambridge, MA 02138.
2. Patrick Haffner, Subhabrata sen, Oliver Spatcheck, Dongmei Wang. 2005. ACAS: Automated Construction of Application Signatures in Proc. ACM SIGCOMM MineNet.
3. Nigel Williams, Sebastian Zander, Grenville Armitage. 2006. A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP traffic Flow Classification in proc ACM SIGCOMM
4. Erman, J., Mahanti, A and Arlitt, M. 2006. Internet Traffic Identification Using Machine Learning in proc. IEEE GLOBECOM.
5. Li, W and Moore, A.W. 2007. A Machine Learning Approach for Efficient Traffic Classification in proc Comput.Telecommun.Syst.
6. Kohavi, R., Quinlan, J.R., Klosgen, W and Zytkow, J. 2002. Decision Tree Discovery, Handbook Data Mining Knowledge.
7. Kohavi, R. 1996. Scaling Up the Accuracy of Naïve-Bayes Classifiers: a Decision –Tree Hybrid in proceedings of 2nd International Conference on Knowledge Discovery and Data Mining.
8. Moore, A and Zuev, D. 2005. Internet Traffic Classification Using Bayesian Analysis Techniques in SIGMETRICS'05, Banff, Canada.
9. Bouckaert, R. 2005. Bayesian Network Classifiers in Weka. Technical Report, Department of Computer Science, Waikato University