2012 International Conference on Network and Computational Intelligence (ICNCI 2012)
IPCSIT vol. 46 (2012) © (2012) IACSIT Press, Singapore

Searching for Online Traffic Classification

Xiaomei Yu[1,2+], Zhenxiang Chen[1,2], Lizhi Peng[1,2], Shupeng Zhao[1,2] and Keke Liu[1]

[1] Shandong Provincial Key Laboratory of Network Based Intelligent Computing, Jinan, 250022, China
[2] School of Information Science and Engineering, University of Jinan, Jinan, 250022, China

Abstract. Real-time traffic classification is vital to solve difficult network management problems. In recent years, machine learning (ML) has proved to be an effective technology for traffic classification. However, almost of the proposed methods are not suitable for real-time traffic classification owing to oppressive cost of computation and storage. In this paper, two machine learning algorithms (C4.5 decision tree and semi-supervised clustering based on K-Means) are estimated. Experimental results demonstrate that, C4.5 decision tree can effectively identify unknown traffic samples while the training data set is enough. Semi-supervised learning need more time to identify unknown traffic although it needn't more labeled data and can find new application. To online traffic classification, it is hard to keep efficiency by semi-supervised learning. So, it can design perfect online traffic classification system by combining supervised learning and semi-supervised learning.

Keywords: Traffic classification, Real time, Semi-supervised learning, Supervised learning

1. Introduction

With the rapid development of Internet, more and more applications expend network resource, especially P2P [1]. These applications constitute a significant share of the total traffic on the Internet and increase network complexity due to the enormous volume of traffic. Therefore, accurate online traffic classification plays important roles in network management such as traffic monitoring, predict patterns and trends of network resource usage [2].

Traditional methodologies for classifying network traffic like port-based and payload-based methods are becoming ineffective, due to more and more applications utilize dynamic port numbers and encryption techniques [3]. In recent years, ML techniques have been used for traffic classification researches and have been proven to be promising technology [4], which classifies traffic with statistics characteristic.

In online classification content, classifier must make advisably decision before a flow is gone. However, most of existing ML classification techniques is not suitable for online traffic classification by using full flow statistics, such as total transferred bytes. Therefore, the online classification still challenge to network management and it should meet the key criteria, such as real time, high accuracy, early detection and low complexity. In order to find the suitable method for online traffic classification, semi-supervised clustering based on K-Means [5] is used and compared with C4.5 decision tree algorithm [6]. Accuracy, latency and cost are analyzed on the same dataset and features.

2. Methodology

2.1. C4.5 Decision Tree (C4.5)

+ Corresponding author. Tel.: +15275165256,13583155823,; fax: +0531-87968014
E-mail address: nic_yuxm@ujn.edu.cn, czx@ujn.edu.cn, nic_zhaosp@ujn.edu.cn, plz@ujn.edu.cn,

C4.5 is well-known as a discriminative decision tree algorithm, which creates the regulation model based on a tree structure [7]. A test node in the tree represent feature, with branches linked to a sub-tree. A leaf representing the class constitutes the output. To classify instances using C4.5, the leaf node is searched for from the root of the tree (the regulation modes). This process will go iteratively into a sub-tree, until it reaches a leaf node with the predicted class.

When building a model, the training set S is consisted of a set of instances which have a fixed set of features $(A_1, \ldots, A_k)^T$ and a class C. The class C represents the application of the network traffic and has the values ($c_1, c_2, \ldots, c_m$). Each feature A_q represents the flow statistics and has the values ($a_1, a_2, \ldots, a_n$). The information gain ratio is used to decide which feature should be chosen as a test node, it reflects the correlation between a feature A_q and a class label C, which is calculated by the equation (1).

$$G_{gainratio}(C \mid A_q) = \frac{-\sum_{i=1}^{m} p(c_i)\log_2 p(c_i) - (-\sum_{j=1}^{n} p(a_j)\sum_{i=1}^{m} p(c_i/a_j)\log_2 p(c_i/a_j))}{-\sum_{i=1}^{m} p(c_i)\log_2 p(c_i)} \quad (1)$$

Where $p(c_i) = P[C = c_i]$, $p(a_j) = P[A_q = a_j]$ and $p(c_i/a_j) = P[C = c_i \mid A_q = a_j]$.

On the other hand, the process of building model iteratively looks for the best feature to partition the data. The one with highest information gain ratio will be chose as the test node, until the node becomes a leaf node. To classify the instance using C4.5, it just needs to compare the features of the test instance to the node of the tree. Identifying traffic used C4.5 has the low computational cost and is realized simply [6].

2.2. Semi-supervised clustering based on K-Means

The semi-supervised clustering based on K-Means based on clustering algorithms consists of two steps. Firstly, K-Means clustering algorithm is employed to partition objects into k clusters from a training data that consists of abundant unlabeled objects and few labeled data. The K-Means is selected due to it is a simple and fast clustering algorithm [8]. Secondly, the labeled data are used to map clusters to the applications. The goal of semi-supervised clustering algorithms is that fast and accurately classify traffic. The semi-supervised clustering based on K-Means can be summarized:

It randomly initial cluster centroid according to the assigned number of clusters. Assign each to the closest cluster centroid by measuring similarity, which is computed by the equation (2). Given the object X, which is described by a set of features $X=(x_1, x_2, \ldots, x_n)^T$ .Y is represent the cluster centroid $Y=(y_1,y_2,\ldots,y_n)^T$.

$$D(X,Y) = \left(\sum_{i=1}^{n}(x_i - y_i)^2\right)^{1/2} \quad (2)$$

The new cluster center is calculated based on the new clusters. Repeat the above steps until the convergence criterion is met. It is evaluated by the square error E, which is computed as the equation (3). Given the k is the number of cluster; m is the number of the object that belongs to the cluster C_j.

$$E = \sum_{j=1}^{k}\sum_{i=1}^{m}\left|D(X_i, C_j)\right|^2 \quad (3)$$

Mapping clusters to applications. The output of the clustering algorithm is a set of clusters, which is consisted of labeled and unlabelled flows. Probabilistic assignment is used to find the mapping from clusters to labels. Given the y_i (i = 1, 2, …, q) is the labeled applications and C_k (k = 1, 2, … , m) is the cluster. The decision function for mapping the label y to the classifying sample is the maximum probabilistic equation (4).

$$y = \arg\max_{y_1,\ldots,y_q}(P(Y = y_i \mid C_k)). \quad (4)$$

The “Unknown” application type is assigned to the cluster, which doesn’t have any labeled samples. Thus the “Unknown” cluster can be used to analyze the new application or the special application.

3. Experiment

Our goal is to investigate an algorithm which is suitable for online traffic classification by analyze the performance of our chosen ML algorithms. In this paper, we use a single dataset and fixed features to test the different algorithm.

3.1. Data set

In order to evaluate the performance of the chose methods for online classification, the public dataset is used, which was used in previous investigations by the authors [9]. The traffic set consists of a full 24 h, week-day period. It was captured in different years and two different sites that has over a thousand local users and captures full-duplex traffic at the site border to the Internet. These flow traces were labeled with a corresponding application category. In this paper, we focus on seven different applications: WEB, BULK, MAIL, ATTACK, P2P, DATABASE and SERVICES.

There are millions of network flows in the public trace files used in this work. We randomly sample some flows to constitute our dataset. In the experiments, we only analyze the TCP flows. When testing the C4.5 algorithm, we use 10-fold cross validation to generate the training set and the test set. It is important to note that in each case, there is no overlap between the training set and the testing set. As we know the class of each flow within the dataset, we labeled a part of them as "unknown" flows to test the semi-supervised clustering method based on K-Means. The dataset is consisted of labeled flows and the unknown flows. When evaluating the performance of the classifier, we just compare the predicted class with the known class.

3.2. Features

Each flow is descriptive by a number of characters and exhibits different feature values depending on the category to which it belongs. For online classification, traffic features should be calculated on the fly and will not consume much time. So we used the feature which had been used in the previous work [9] and has been prove to be the suitable features for online classification. The features used in this paper are summarized: count of all packets with push bit set in TCP header,the total number of bytes sent in initial window, average segment size, median of total bytes in IP packet, count of packets with at least 1 byte of TCP data payload, variance of total bytes in packets, minimum segment size observed, total numbers of RTT samples found,count of all packets with push bit set in TCP header, server port and client port.

3.3. Evaluation Techniques

- Precision and Accuracy

The classification model is evaluated by the conventional machine learning metrics such as precision and accuracy [4]. Precision of algorithm is the ratio of the number of class members classified correctly over the total number of instances classified as class members. This metric describes the classification capability to identify objects correctly. Recall is the ratio of the number of class members which are classified correctly over the total number of class members. This value represents the classification capability to determine misclassified members are something else. Overall Accuracy is the overall ratio of correctly classified instances over the total number of instances.

- Computational performance

For online classification, the computational performance is important due to thousands of simultaneous networks flows need to be identified. So we use the term computational performance, which is described by two additional metrics: build time and classification speed. Build time refers to the time required to train a classifier on a given dataset. Classification speed describes the number of classification that can be performed each second.

- Robustness

In the real-world traffic, the user behavior reflected in traffic may vary dramatically owing to different conditions and different periods [10]. For the online classification, if it can be used in different network locations, should be considered. The online classifier also can effectively identify the emergence of new traffic applications. So the robustness of the classifier should regard as a metrics to evaluate the classification performance.

3.4. Experiment toolkit and platform

The experiment platform is general-purpose PC which carries on Windows 7 operating system, its CPU is Intel Core(TM) 2-6300, dominant frequency is 1.88 GHz, and Physical memory is DDR-667 2GByte. The machine-learning models are implemented in C++ language.

4. Results and analysis

4.1. Comparing Algorithm Classification Accuracy

- Accuracy of the C4.5 algorithm

To test and evaluate the C4.5 algorithm we use 10-fold cross validation. In this process the dataset is divided into 10 subsets. Each time, one of the 10 subsets is used as the test set and the other 9 subsets form the training set. The performance is calculated across all 10 experiments. When it is repeated 10 times, C4.5 algorithm obtains the average accuracy is 85.12%, the average recall and precision of each classification is shown in the Fig 1.

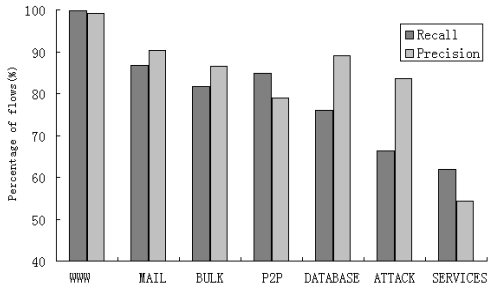

Fig. 1: Per-application precision and recall of C4.5

The results show that the average precision and recall of WWW are 99.82% and 99.06% respectively, which indicates that WWW can be effectively identified. At the same time, the average identification accuracy of SERVICES is 57 %, which means that the method is hard to accurately classify SERVICES. This situation may be caused by lacking enough training samples, which are just 555 instances in the training dataset. Instead, the WWW has 10000 instances which take over 36% in the training dataset. It suggests that the scarce training samples of the application category make the precision low.

- Accuracy of semi-supervised clustering method based on K-Means

Due to the semi-supervised clustering method based K-Means algorithms, the number of clusters impacts the accuracy of clustering and the time complexity. So in this experiment, the number of clusters and the number of labeled samples, those are both varied. We changed the number of clusters from 100 to 500, and varied the number of labeled samples in the dataset from 10% to 50%.

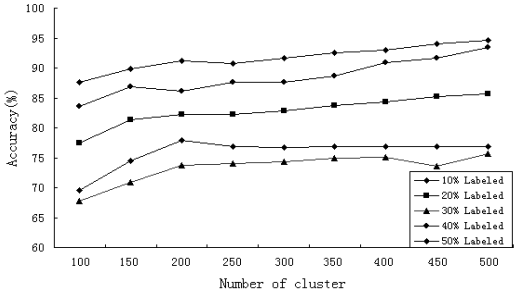

Fig. 2: Accuracy of the semi-supervised clustering method

As illustrated in Fig 2, sample accuracies improve with the number of clusters increases. At the same time, the accuracy varies with the changing percentage of the labeled flows. From the results, we can observe that for a fixed number of labeled flows and a number of unlabeled flows can achieve high accuracy. For example, the accuracy can achieve 86.85%, when the training data sets with 40% labeled flows and the number of cluster is 150.when the K is 500 with 50% labeled flows, the accuracy can achieve 94.61%.

4.2. Comparing Algorithm Computational Performance

Considering the requirements of the online classification, the classification method must be practicable to facilitate online real-time identification on high speed links for large traffic volumes with low computational complexity. We therefore focus on the classification speed of the algorithms. The average training time of the C4.5 is 908s and the testing time is 0.45s. The time cost of the semi-supervised clustering method based on K-Means is varied from 216s to 652s with the increase of the number of the clusters.

While C4.5 algorithm builds the classification model, it costs a lot of time to compute information gain ration. But this algorithm can fast (6,087 classifications per second) identify the test samples after the model has been built. The semi-supervised clustering method based on K-Means can't identify the samples until the clustering is end and this process needs much time. From the result of the experiment, we can conclude that the test performance of C4.5 is better than the semi-supervised clustering method.

4.3. Robustness

C4.5 based on supervised learning must be trained using the training data before it is used to identify the test traffic. The similarity between the test data and the training data related to the accuracy of the classifier. When C4.5 used to the online classification, if the test network environment is different to the training traffic, the result is invalid for network management. Moreover, it also cannot find the new application. However, the semi-supervised clustering method based on K-Means can do this due to it don't need to train firstly and it can identify the network traffic according the labeled flow in this network environment.

5. Conclusions and future work

The goal of this work is find the suitable method to be used on the online classification. In this paper we have demonstrated the performance of C4.5 and semi-supervised clustering method for classifying the network traffic. The experiment results demonstrated that the C4.5 can fast identify the traffic with the desired accuracy. But it requires a large amount of labeled data to train classifier and they cannot discover new applications. Moreover, the labeled data are also hardly obtained. The semi-supervised clustering method not only accurately classifies but also can find new application. It requires few labeled flows and gives higher accuracy. However, the time cost is too much to be implemented on practical circumstance. So we can conclude that different method can be used in different network circumstance and can design perfect online traffic classification system by combining supervised learning and semi-supervised learning. In the future work, we concentrate on realizing the online traffic classification in the dynamic traffic.

6. Acknowledgments

This work is supported by national natural science fund under Grant No. 60903176, youth and mid-life scientist's award fund in Shandong province under Grant No. BS2009DX037, Shandong province natural science fund under Grant No. ZR2010FQ028, the Natural Science Foundation of Shandong Province of China under Grant No. 2011ZRB019A7 and Youth science and technology star fund of Jinan No.TNK1108.

7. References

[1] H. Schulze and K. Mochalski. Ipoque Internet Study 2008/2009. *Available from: <http://www.ipoque.com/>*.

[2] Soysal, Murat,Schmidt, E. Guran. Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison. *Performance Evaluation*. vol 6, pp 451-467 2010.

[3] T. Karagiannis, A. Broido, M. Faloutsos, and K. claffy. Transport Layer Identification of P2P Traffic. In *IMC'04", Taormina, Italy,* October pp.25-27. October 2004.

[4] T. Nguyen, and G. Armitage. A Survey of Techniques for Internet Traffic Classification using Machine Learning. *IEEE Communications Surveys & Turorials.* vol 4, pp 56-76, 2008.

[5] Erman, A. Mahanti, M. Arlitt, I. Cohen, and C. Williamson. Offline/real-time traffic classification using semi-supervised learning. *Technical report.* University of Calgary,2007.

[6] N. Williams, S. Zander, and G. Armitage. A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification. *ACM SIGCOMM Computer Communication Review.* 30(5),5-

16(2006).

[7] Y. Ma, Z. Qian, G. Shou, Yihong, Hu. Study of information network traffic identification based on C4.5 algorithm. *2008 International Conference on Wireless Communications, Networking and Mobile Computing*. 2008

[8] Erman, M. Arlitt, and A. Mahanti. Traffic classification using clustering algorithms. *In Proceedings of SIGCOMM workshop on Mining Network Data*. 281-286(2006).

[9] W. Li, M. Canini, A. Moore, R. Bolla. Efficient application identification and the temporal and spatial stability of classification schema. *Computer Network.* 53(6), 790 – 809(2009).

[10] Q. Sun, X. Huang, Y. Ma. A Dynamic Online Traffic Classification Methodology based on Data Stream Mining. *In: WRI World Congress on Computer Science and Information Engineering.* vol. 1, pp. 298–302 (2009).