Journal of Computational Information Systems 7:5 (2011) 1524-1532
Available at http://www.Jofcis.com

Online Wireless Mesh Network Traffic Classification using Machine Learning

Chengjie GU[1,†], Shunyi ZHANG[1], Xiaozhen XUE[2], He HUANG[3]

[1]*Institute of Information Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*

[2]*Department of Computer Science, Texas Tech University, Texas 79415, USA*

[3]*School of Software, Beijing University of Aeronautics and Astronautics, Beijing 100083, China*

Abstract

Internet traffic classification based on flow statistics using machine learning has attracted great attention. Wireless mesh networks (WMNs) facilitate the extension of wireless local area networks into wide areas and have emerged as a key technology for next-generation wireless networking. Online and accurate traffic classification is a key challenge for wirless mesh network management. We propose an online wireless mesh network traffic classfication using C4.5. We evaluate the effectiveness of our proposed method through the experiments on real mesh traffic traces. Experiment results illustrate this method can classify online wireless mesh network traffic with high accuracy.

Keywords: Traffic Classification; Wireless Mesh Network; Machine Learning; Accuracy

1. Introduction

The rapid development of P2P (Peer-to-Peer) applications have enriched the performance of Internet in recent years. The P2P traffic has occupied above 50% of the total Internet traffic, and the ratio is still growing [1]. The large amount of P2P traffic and the rapid growth on the usage of P2P applications have led to network congestion and traffic hindrance because of the excessive occupation of the network bandwidth [2]. Therefore, accurate and online classification of network traffic plays important roles in many areas such as traffic engineering, QoS, and intrusion detection etc.

The earliest and simplest traffic classification approach is port-based traffic classification which consists in examining the port numbers in TCP headers [3]. This method is no longer valid because of the inaccuracy and incompleteness of its classification results. Several payload-based analysis techniques have been proposed to inspect the packets payload searching for specific signatures [4]. Although this solution does can achieve high classification accuracy, it can't work with encrypted traffic or newly P2P applications [5]. At the same time, traffic classification method based on flow statistics shows effective performance in this field. Substantial attention has been invested in data mining techniques and machine learning algorithms using flow features for traffic classification.

† Corresponding author.
Email addresses: jackiee.gu@gmail.com (Chengjie GU).

May, 2011

Wirelss mesh network is generally considered as a type of ad-hoc networks due to the lack of wired infrastructure that exists in cellular or Wi-Fi networks through deployment of base stations or access points [6]. Recently, lots of attention has been paid on wirelss mesh network as a fundamental technology which is one of the most sought-after networks of next generation [7].

While traffic classification methods based on flow statistics offer various degrees of successes, there are several limitations. 1) The existing classifiers cannot classify traffic effectively, because they cannot always capture the initial sub-flow when a communicating node moves in wireless mesh network. 2) Many proposed classifiers can't solve the online traffic classification problems faultlessly.

To address the above-mentioned problems, we take two aspects to improve the accuracy and speed of the machine learning methods for wireless mesh network traffic classification. 1) We propose sub-flow selection with application behaviors, and the method solves a critical problem of how to select appropriate sub-flows for achieving traffic classification in wireless mesh network. 2) In order to achieve early detection, we allow the classifier to classify traffic flows early in the connection using the first p packets of flow. Our online traffic classification should be indispensable to manage the network which has a variety of mobile nodes.

The remainder of this paper is structured as follows. Related work is represented in Section 2. Section 3 describes the relationship between wireless mesh network and traffic calssification. Our proposed online mesh network traffic classification using machine learning is presented in this section 4. Section 5 presents the experimental results and analysis. The conclusions and potential future work are listed in Section 6.

2. Related Work

Machine learning technique which is a powerful tool in data separation in many disciplines aims to classify data based on either a priori knowledge or statistical information extracted from raw dataset. This method can be well suited with Internet traffic classification, as long as the traffic classified into categories that exhibit similar characteristics in parameters. Each individual flow can be associated with a specific application according to its flow features, and consequently traffic classification is achieved without direct inspection of each packet's header or payload by Machine Learning (ML) algorithm. The flow features are statistical patterns, such as packet size, packet inter-arrival time, and packet arrival order.

Machine learning algorithms are generally divided into supervised learning and unsupervised learning. Unsupervised learning essentially clusters flows with similar characteristics together [8-10]. The advantage is that it does not require training, and new applications can be classified by examining known applications in the same cluster. Erman et al. [8] compared the performance of unsupervised machine learning algorithms in traffic classification. Since our main focus is on evaluating the predictive power of a built/trained traffic classifier rather than on detecting new applications or flow clustering. Also, Erman et al. [9] evaluated the performance of two clustering algorithms, namely K-Means and DBSCAN, in Internet traffic classification. The result indicated that K-Means was one of the quickest and simplest algorithms for clustering of Internet flows. Bernaille et al. [10] used a simple K-Means clustering algorithm to perform classification by using only the first five packets of the flow, aiming at applying on the real-time classification. Supervised learning requires training data to be labeled in advance and produces a model that fits the training data [11-13]. Moore et al. [11] used a Naive Bayes classifier which was a supervised

machine learning approach to classifying internet traffic. But only 65% accuracy rate, which was not good enough to classify Internet traffic. Williams et al. [12] conducted a comparison of five machine learning algorithms which were widely used to classify empirical study of Internet traffic. Among these algorithms, C4.5 achieved the highest accuracy in their results. Auld [13] proposed supervised machine learning based on a Bayesian neural network to classify the traffic with higher accuracy and better stability, but it was not capable for real-time applications.

3. Traffic Classification in Wireless Mesh Network

3.1. Wireless Mesh Network

As various wireless networks evolve into the next generation to provide better services, wireless mesh networks (WMNs) have emerged recently. In WMNs, nodes are comprised of mesh routers and mesh clients. Each node operates not only as a host but also as a router, forwarding packets on behalf of other nodes that may not be within direct wireless transmission range of their destinations [14]. A WMN is dynamically self-organized and self-configured, with the nodes in the network automatically establishing and maintaining mesh connectivity among themselves [15]. This feature brings many advantages to WMNs such as low up-front cost, easy network maintenance, robustness, and reliable service coverage.

As illustrated in Figure 1, WMNs consist of two types of nodes: mesh routers and mesh clients. Other than the routing capability for gateway/repeater functions as in a conventional wireless router, a wireless mesh router contains additional routing functions to support mesh networking [16]. To further improve the flexibility of mesh networking, a mesh router is usually equipped with multiple wireless interfaces built on either the same or different wireless access technologies. Compared with a conventional wireless router, a wireless mesh router can achieve the same coverage with much lower transmission power through multi-hop communications. Optionally, the medium access control (MAC) protocol in a mesh router is enhanced with better scalability in a multi-hop mesh environment [17].

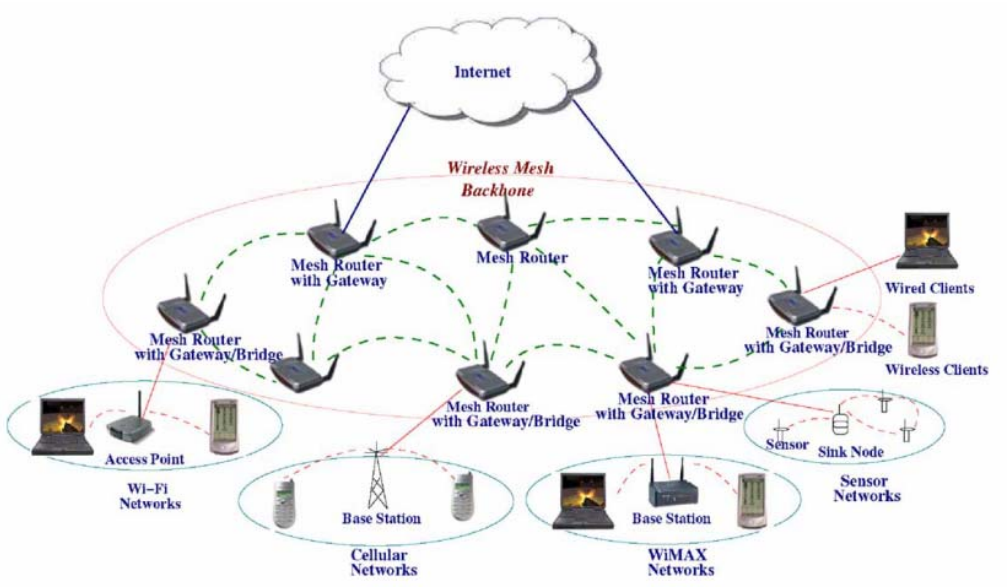

Fig.1 WMNs Infrastructure Architecture

3.2. Traffic Classification in Wireless Mesh Network

Different from other ad hoc networks, most applications of WMNs are broadband services with various QoS requirements. Thus, in addition to end-to-end transmission delay and fairness, more performance metrics such as delay jitter, aggregate and per-node throughput, and packet loss ratios, must be considered by communication protocols. Traffic classification of wireless mesh network should take into account the factors of packet loss ratio and connection loss ration.

Until recently, most published works about online traffic classification have relied on features of initial sub-flow. Those works assume the initial sub-flow is captured and available for traffic classification. However, this assumption is not accepted in wireless mesh network because of seamless mobility. In the case of changing its point of attachment caused by movement of a node with communication, those classifiers begin to capture packets composing a flow at a point in time when the flow is already in progress. Consequently those classifiers cannot capture the initial sub-flow. Desirable classifier should thus be capable of traffic classification based on not only initial sub-flow but also various types of sub-flow.

4. Online Wireless Mesh Network Traffic Classification using Machine Learning

4.1. Sub-flow Selection

Internet traffic classification schemes operate on the notion of network flows. A flow is defined to be as a series of packet exchanges between two hosts, identifiable by the 5-tuple (source address, source port, destination address, destination port, transport protocol), with flow termination determined by an assumed timeout or by distinct flow termination semantics. For each flow, network monitors can record statistics such as duration, bytes transferred, mean packet interarrival time, and mean packet size. Sub-flow is N consecutive packets taken from a flow. An initial sub-flow is the initial *N* consecutive packets from the point where communication established.

The applications provide users with various types of service. We define these events of each application as application behaviors that are expressible like finite state machine. Each application starts from initial behavior and accepts various events, as instructions changing behaviors. Whenever the application accepts the event, it changes to the next behavior and executes actions associated with the behavior.

We assume that a sub-flow, taken from the transition point of the application behavior, is effective for the achievement of the desirable classifier. The rationale behind possibility of the classifier is the following. A sub-flow is N consecutive packets taken from a full-flow; therefore, two or more sub-flows can be taken from a flow. The sub-flow features have ability to distinguish each application, because the sub-flow contains a sequence exchanging control packets that are pre-defined messages in each application. We devise the classifier that added the function of sub-flow selection with application behaviors.

4.2. Machine Learning Algorithm

Machine Learning is prevailing in traffic classification because of its being independent from the port and payload information. These are some different machine learning algorithms, such as C4.5, Support Vector Machine (SVM), Naive Bayesian and Random Forest. In order to identify wireless mesh network traffic, we choose the C4.5 which has high TP rate and low FP [18].

C4.5 is a decision tree based identification algorithm. A decision tree is a hierarchical data structure for

implementing a divide-and-conquer strategy. It is an efficient non-parametric method that can be used both for identification and regression. In non-parametric models, the input space is divided into local regions defined by a distance metric. In a decision tree, the local region is identified in a sequence of recursive splits in smaller number of steps. A decision tree is composed of internal decision nodes and terminal leaves. Each node m implements a test function with discrete outcomes labeling the branches. This process starts at the root and is repeated until a leaf node is hit. The value of a leaf constitutes the output.

4.3. Online Wireless Mesh Network Traffic Classification Architecture

Classifiers based on machine learning use a training dataset that consists of *N* tuples (x_i, y_i) and learn a mapping $f(x) \to y$. In the traffic classification context, examples of attributes include flow statistics such as duration and total number of packets. In our supervised Internet traffic classification system, Let $X = \{x_1, x_2, \cdots, x_n\}$ be a set of flows. A flow instance x_i is characterized by a vector of attribute values, $x_i = \{x_{ij} \mid 1 \le j \le m\}$, where *m* is the number of attributes, and x_{ij} is the value of the j^{th} attribute of the i^{th} flow. Also, let $Y = \{y_1, y_2, \cdots, y_q\}$ be the set of traffic classes, where *q* is the number of classes of interest. The y_i can be classes such as “P2P”, “WWW”, and “FTP”.

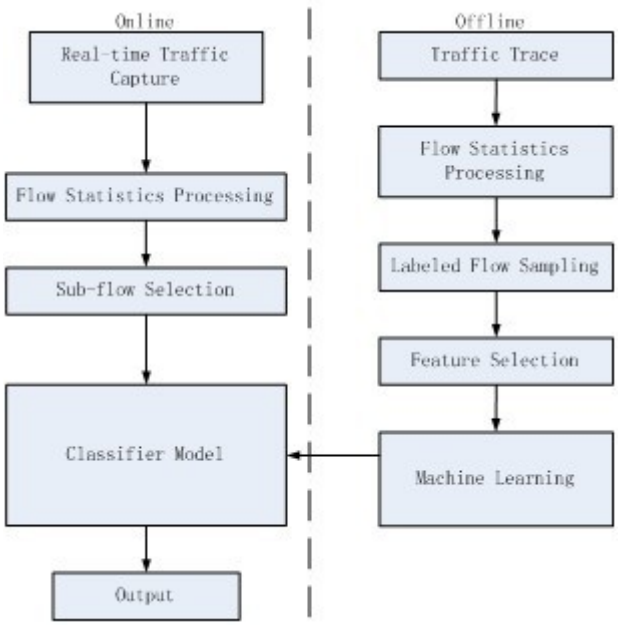

Fig.2 Architecture of Online Wireless Mesh Network Traffic Classification using Machine Learning

As illustrated in Figure 2, the architecture of online wireless mesh network traffic classification using machine learning includes two portions: off-line ML modeling and on-line ML classification. In the stage

of off-line ML modeling, traffic trace should be labeled to different application using L7-Filter or Hand-classification method. Training subsets are uniformly sampled for the specific ML classification. Once the datasets are ready, the system carries out feature selection to eliminate the redundant and irrelevant features to get optimal feature subset. In the stage of online ML classification, the flows which can be realtime collected are trained using the selected ML classifier. Flow statistics are computed in terms of each selected feature and stored them to the corresponding database in the flow preprocessing module. The sub-flow selection module can implement sub-flow selection. The output classifier model produced by off-line ML modeling is applied to classify the captured traffic. Eventually, the classification output would be applied to network activities such as network surveillance, QoS.

5. Experimental Results and Analysis

5.1. Empirical Traces

This subsection describes the empirical traces in our work. We construct a pint-sized wireless mesh network in Nanjing University of Posts and Telecommunications. To facilitate our work, we collect traces on the campus and residential area. The Campus_Set was collected from 9 am to 10 am on June 28, 2009. The Residential _Set was collected from 6 pm to 7 pm on June 28, 2009. To simplify the presentation, we group the applications by category. For example, the P2P category includes all identified P2P traffic from protocols including PPlive, PPstream, BitTorrent, Gnutella, Xunlei and KaZaA. Table 1 summarizes the applications found in the Campus traces and Residential traces.

Table 1 Statistics of Empirical Traces

Type of Flow	Application Names	Campus_Set		Residential _Set	
		Num of flow	Percent(%)	Num of flow	Percent(%)
WWW	http, https	46384	48.77	23215	21.04
MAIL	Imap, pop3, smtp	9326	9.81	3827	3.47
BULK	ftp	13729	14.44	16308	14.75
DATABASE	oracle, mysql	5462	5.74	2103	1.91
SERVER	ident,ntp,x11,dns	2099	2.21	962	0.87
P2P	kazaa, bittorrent	13753	14.46	37917	34.36
MEDIA	real, media player	4301	4.52	18342	16.64
GAME	half-life	46	0.05	7681	6.96
Total	32 applications	95100	100	110355	100

Prior to the ML modeling, feature selection can be executed off-line. Feature selection is an important step to machine learning which is the process of choosing a subset of original features. It can optimize for higher learning accuracy with lower computational complexity by removing irrelevant and redundant features. In order to find the best feature subset for traffic classification, we use the Sequential Forward Selection (SFS) method in the paper.

5.2. Evaluation Metrics

To measure the performance of our proposed method, we use three metrics: *accuracy*, *precision* and *recall*.

$$accuracy = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \times 100\% \quad (1)$$

Accuracy is the ratio of the sum of all True Positives to the sum of all the True Positives and False Positives for all classes. We apply this metric to measure the accuracy of a classifier on the whole trace set. The latter two metrics are to evaluate the quality of classification results for each application class.

$$precision = \frac{TP}{TP + FP} \times 100\% \quad (2)$$

Precision of an algorithm is the ratio of True Positives over the sum of True Positives and False Positives or the percentage of flows that are properly attributed to a given application by this algorithm.

$$recall = \frac{TP}{TP + FN} \times 100\% \quad (3)$$

Recall is the ratio of True Positives over the sum of True Positives and False Negatives or the percentage of flows in an application class that are correctly identified.

5.3. Comparing Performance among Different Technology

We compare classification accuracy of online wireless mesh network traffic classification using machine learning with the other classification approaches solely based on port-based approach and payload-based approach. For our experiments, we classify network traffic using flows from Campus_Set.

Table 2 Classification Performance among Different Traffic Classification Method

Type of Flow	Port-based		Payload-based		ML-based	
	Recall	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*
WWW	70.09	71.97	76.88	77.26	87.98	91.77
MAIL	69.25	68.64	76.04	73.22	87.13	87.65
BULK	64.94	61.15	68.22	65.92	79.37	80.43
DATABASE	67.13	65.57	70.28	69.49	83.26	83.95
SERVER	67.36	64.62	72.97	70.31	84.04	84.91
P2P	59.89	57.25	69.05	64.77	80.15	79.39
MEDIA	70.41	62.08	73.81	66.53	84.93	83.04
GAME	58.54	54.23	67.43	62.15	78.53	76.65

Accuracy is a critical requirement of traffic classification that can be measured in terms of recall and precision, both of which are important. As illustrated in table 2, we can see that the classification precision which uses port-based methods to classify network traffic is 71.97%, 57.27% for WWW and P2P application respectively. At the same time, the precision of payload-based classification which extracts payload signatures to identify specific protocol is 77.26%, 64.79% for WWW and P2P application respectively. In

addition, in order to classify network traffic, we choose the machine learning method using C4.5 which has 91.77%, 79.39% for WWW and P2P application respectively. Compared with the other classification approaches, this method proposed in the paper can achieve higher overall recall and precision. For a range of network applications, online wireless mesh network traffic classification based on machine learning can meet the key requirement, such as high accuracy.

5.4. Impact of Number of Packets for Statistics on Classification Accuracy

A fundamental challenge in the design of the online wireless mesh network traffic classification is to classify a flow as soon as possible. Unlike offline classification where all flow statistics are available a priori, we only have partial information on the flow statistics in the real-time context.

In order to classify the network applications associated with a flow as early as possible, we follow the idea presented in Bernaille et al. [10] and conduct experiments to determine the appropriate packet number p. The statistics information of several packets in each flow could distinguish network traffic from Internet traffic accurately with the least p. For our experiments, we classify network traffic using flows from Campus_Set and Residential_Set respectively.

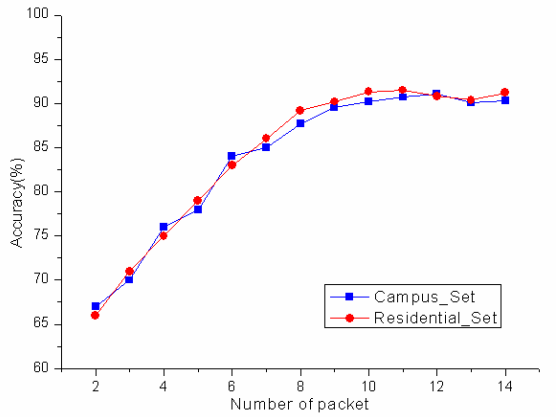

Fig.3 Impact of Number of Packets for Statistics on Classification Accuracy

The experimental result shown in Figure 3 indicates that this method could achieve better accuracy when we choose 11 packets for statistics. It is noticeable that the statistics of the first 11 packets could classify traffic with high accuracy in Campus_Set and Residential_Set. At the same time, the classification accuracy improves marginally using more than 11 packets. Considering our goal to classify wireless mesh network traffic as fast as possible with high accuracy, we choose 11 packets for flow feature statistics.

6. Conclusion

Wireless mesh networks (WMNs) are communications network made up of radio nodes organized in a mesh topology. Wireless mesh networks traffic, being aggregated from a large number of end users, changes infrequently. It causes the most challenging problem in network traffic classification. In this paper, we propose online wireless mesh network traffic classification using machine learning. We show that our method can achieve high classification accuracy. This method is also suitable for online network traffic classification. Our experimental results so far are promising for this research direction in wireless mesh

network, and several opportunities exist for future work. Supervised machine learning method is the requirement on a large number of labeled training samples. We can propose semi-supervised classification method to solve this issue for wireless mesh network traffic classification.

Acknowledgement

This work is supported by National High-Tech Research and Development Plan (863) of China (No.2009AA01Z212, No.2009AA01Z202), Natural Science Foundation of China (No.61003237), Natural Science Foundation of Jiangsu Province (No.BK2007603), High-Tech Research Plan of Jiangsu Province (No.BG2007045).

References

[1] B.Marco, Mellia, Antonio Pescape and LucaSalgarelli, Traffic classification and its applications to modern networks, *Computer Networks*, 53(6):759-76, 2009.

[2] Soysal, Murat,Schmidt, Ece Guran.Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison. *Performance Evaluation*, 67(6):451-467, 2010.

[3] Karagiannis T, Roido A, Aloutsos M, Laffy K. Transport layer identification of P2P traffic. *Proceedings of the 2004 ACM SIGCOMM Internet Measurement Conference*, pages 121–134, 2004.

[4] Haffner P, Sen S, Spatscheck O, Wang D. ACAS: Automated Construction of Application Signatures. *ACM SIGCOMM'05 Workshops*, pages 197–202, 2005.

[5] Moore AW, Papagiannaki K. Toward the accurate identification of network applications. *Passive and Active Measurement Workshop*, pages 41–54, 2005.

[6] Akyildiz Ian F, Wang Xudong, Wang Weilin.Wireless mesh networks: A survey. *Computer Networks*, 47(4):445-487, 2005.

[7] Zeng Guokai, Wang Bo, Ding Yong, Xiao Li. Efficient multicast algorithms for multichannel wireless mesh networks. *IEEE Transactions on Parallel and Distributed Systems*, 21(1):86-99, 2010.

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

[9] J.Erman, M. Arlitt, and A. Mahanti. Traffic classification using clustering algorithms.*Proceedings of SIGCOMM workshop on Mining network data*, pages 281-286, 2006.

[10] Bernaille L, Teuxeira R, Akodkenous I, Soule A, Slamatian K. Traffic classification on the fly. *ACM SIGCOMM Computer Communication Review* 36: 23–26, 2006.

[11] A.W. Moore and D. Zuev. Internet traffic classification using bayesian analysis techniques. *Proceedings of ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pages 50-60, 2005.

[12] 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.

[13] T.Auld, A.W.Moore, and S.F. Gull. Bayesian neural networks for internet traffic classification. *IEEE Transaction on Neural Network*, 18(1):223-239, 2007.

[14] Li Bo,Zhang Qian, Liu Jiangchuan and Wang, Chonggang. Advances in wireless mesh networks. *Mobile Networks and Applications*, 13(1):1-5, 2008.

[15] Bejerano Yigal, Han Seung-Jae, Kumar Amit. Efficient load-balancing routing for wireless mesh networks *Computer Networks*, 51(10):2450-2466, 2007.

[16] Yanchao Zhang, Yuguang Fang. A secure authentication and billing architecture for wireless mesh networks. *Wireless Networks*. 13(5):663- 678, 2007.

[17] Ian Akyildiz, Xudong Wang. Cross-layer design in wireless mesh networks. *IEEE Transaction on Vehicular Technology*. 57(2):1061- 1076, 2007.

[18] Yongli Ma, Zongjue Qian, Guochu Shou, Yihong, Hu. Study of information network traffic identification based on C4.5 algorithm. *International Conference on Wireless Communications, Networking and Mobile Computing,* 2008.