Cooperative Statistical Traffic Big Data Classification with Machine Learning

Roshini. G, Nathiya. M, Keerthana. M, Ms. A.Sujethira (AP/CSE)
Department Of Computer Science and Engineering
Paavai College of Engineering, Namakkal

***Abstract*:- The continuous collection of traffic data through the network leads to big data problems caused by the volume, variant, and speed characteristics of big data. Learning network properties requires machine learning techniques that capture global knowledge about traffic patterns. The big data properties lead to significant system challenges in the implementation of machine learning frameworks. In this system described about the difficulty and challenges of behavior big data classifications using geometric representation learning techniques and modern big data networking technologies. Cooperative protocol for the establishment of these clusters and for the cooperative transmission of data. We gain the upper limit of the capacity of the protocol and analyze the end-to-end robustness of the protocol against data packet loss as well as the trade-off between power consumption and error rate. The analysis results are used to compare the energy savings and the end-to-end robustness of our protocol with two non-cooperative methods. In particular, this article discusses the problems associated with the combination of supervised learning techniques, representation learning techniques, and machine-based lifelong learning techniques and big data technologies**

Keywords: Learning network, big data, machine learning frameworks, Cooperative protocol, power consumption, supervised learning techniques, machine-based lifelong learning techniques.

1. INTRODUCTION

Big data is currently defined by three data characteristics: volume, grade, and speed. It means that at a time when the amount of data, the variety and the speed of the data are increased, the current techniques and technologies may not be able to process the storage and processing of the data. At this point, the data is defined as big data. In big data research, the term "big data analytics" is defined as the process of analyzing and understanding the properties of large-size datasets by extracting useful geometrical and statistical patterns. Ideally, these three features of a data set increase the complexity of the data and cause the current techniques and technologies to stop functioning as expected within a given processing time. Many applications suffer from the big data problem, including network traffic risk analysis, geographic classification, and business forecasting. Network intrusion detection and prediction are time-critical applications that require high-efficiency big-data techniques and technologies to solve the problem on the fly. The new technologies can help to make big data analytics for different applications. The Hadoop Distributed File Systems (HDFS), Cloud Technology and Hive Database technologies can all be combined to address issues such as big data classification. However, the applications that require continuous growth in the Big Data domain, including the Intrusion Prediction System and Geospatial, can suffer significantly from big data problems. This article discusses some of the issues and challenges associated with integrating advanced network technologies and machine learning techniques to solve big data classification problems for network intrusion prediction.

2. RELATED WORK

2.1 A survey of techniques for Internet traffic classification using machine learning

T. T. T. Nguyen and G. Armitage was says the Real-time traffic classification has potential to solve difficult network management problems. Network managers need to know traffic characteristics. Traffic classification can be useful in QoS provisioning Real-time traffic classification is core to QoS-enabled services and automated QoS architectures. RT traffic classification allows to respond to network congestion problems. Automated intrusion detection systems. Detect patterns indicative of denial of service attacks. Trigger automated re-allocation of network resources. Identify customer use of network resources which contradicts operator's term of service. Clarification of ISP obligations with respect to "lawful interception" of IP traffic

2.2 Survey on Threats and Attacks on Mobile Networks

S. Mavoungou, G. Kaddoum, M. Taha, and G. Matar described Since the 1G of mobile technology, mobile wireless communication systems have continued to evolve, bringing into the network architecture new interfaces and protocols, as well as unified services, high data capacity of data transmission, and packet-based transmission (4G). This evolution has also introduced new vulnerabilities and threats, which can be used to launch attacks on different network components, such as the access network and the core network. These drawbacks stand as a major concern for the security and the performance of mobile networks, since various types of attacks can take down the whole network and cause a denial of service, or perform malicious activities. In this survey, we review the main security issues in the access and core network (vulnerabilities and threats) and provide a classification and categorization of attacks in mobile network. In addition, we analyze major attacks on 4G mobile networks and corresponding countermeasures and current mitigation

solutions, discuss limits of current solutions, and highlight open research areas.

2.3 Flexible deterministic packet marking: An IP traceback system to find the real source of attacks

Y. Xiang, W. Zhou, and M. Guo was said Now a day, protection is essential part of our daily life and it becomes easier because many factors working against the intruders, hackers, criminal. Very popular security system i.e. alarm and camera systems protect secure places like banks. Network security is to protect data during their transmission. While transmitting the message between two users, the unauthorized user intercepts the message, alters its contents to add or delete entries, and then forwards the message to destination user. For finding the real source of IP packets we have to use an IP traceback system. In this paper we are presenting an IP traceback system. Flexible Deterministic Packet Marking (FDPM). FDPM belongs to the packet marking family of IP traceback system. There are two main characteristics of FDPM: first it uses the flexible mark length strategy; second it can also change its marking rate according to the routers which are participating in the transmission by flexible flow based marking scheme. FDPM is very efficient for finding the real source of attack than the other traceback systems and with the low resource requirement on routers. Internet Protocol (IP) traceback is the technology to give security to internet and secure from the internet crime. IP traceback system is also called as Flexible Deterministic Packet Marking (FDPM) which builds such a defense mechanism which has ability to find out real source of attacks on packets that traverse through the network. While a number of other trace back schemes exist, FDPM provides innovative features to trace the source of IP attack and can obtain better tracing capability than others. In this paper we are concentrating on how packet marking is done and how we trace the source of attack so firstly the whole message is splits into number of packets. Then all Packets are marked on marker side according to marking Scheme algorithm. If intruder intrudes and gets access of the packets and modify them then with the help of reconstructor we reconstruct the file at the receivers end. Finally receiver reconstructs the file and gets IP address of sender and hacker Using IP spoofing Technique, MAC address and Location of an intruder also.

2.4 A Heterogeneous Service-Oriented Deep Packet Inspection and Analysis Framework for Traffic-Aware Network Management and Security Systems

M. A. Ashraf, H. Jamal, S. A. Khan, Z. Ahmed, and M. I. Baig was said the A variety of Web-based applications, mobile apps, and other over the top data services with affordable 3G/4G enabled smart devices are major factors for enormous increase in heterogeneous data traffic at enterprise and mobile networks. This creates challenges regarding traffic management and requires traffic-aware intelligent network management to deliver sustained quality of experience for subscribers. Deep packet inspection and analysis (DPIA) provides base platform for development of traffic-aware intelligent network management and security systems. However, computationally complex DPIA-related packet processing for high speed data traffic makes these systems expensive. Furthermore, conventionally these traffic-aware network management and security systems are deployed in enterprise networks with independent and dedicated DPIA-related processing resources and require multiple copies of passively provisioned high speed data from network, while performing similar DPIA operations over the same data again and again. This duplicate deployment of expensive software and hardware resources for DPIA processing eventually results in higher capital expenditures as well as operational expenditures for network operators. We have proposed a novel service-oriented framework for heterogeneous deep packet inspection and analysis (SoDPI) that simultaneously provides diversified DPIA services to multiple client applications for network management and security operations in high-speed networks. Proposed framework provides flexible and comprehensive API-based service interface for client applications to register required DPIA services. SoDPI framework implementation is based on commodity hardware and deploys shared set of DPIA-related packet processing components, requiring only single copy of passive data provisioned from network. Experimental evaluations show that novel SoDPI framework requires considerably reduced amount of software and hardware resources to fulfill heterogeneous DPIA packet processing requirements for multiple client applications in comparison with conventional network management and security applications with dedicated DPIA components. This results in lower cost impacts for network operators with more network manageability.

3. PROPOSED SYSTEM

The continuous collection of traffic data by the network leads to Big Data problems that are caused by the volume, variety and velocity properties of Big Data. The learning of the network characteristics requires machine learning techniques that capture global knowledge of the traffic patterns. In our model of cooperative transmission, every node on the path from the source node to the destination node becomes a cluster head, with the task of recruiting other nodes in its neighborhood and coordinating their transmissions. Consequently, the classical route from a source node to a sink node is replaced with a multi-hop cooperative path, and the classical point-to-point communication is replaced with many-to-many cooperative communication.

4. METHODOLOGY

4.1 Network Construction

They can be integrated to build a large and flexible network topology with a storage infrastructure that can change adaptively based on the need of the Big Data processing requirements. However this integrated model will bring several challenges that must be handled efficiently. We construct a network topology to register the nodes. In the Big Data network, numerous nodes are interconnected and exchange data or services directly with

each other. All systems have Connection with other systems. System details are maintained in the server system. It provides connection to the node whenever there is a request from another node. It's possible for a client to get more than one connection to the server. Create packet with IP header, data, and packet length. It receives the packets from source and analyzes the packet header.

4.2 Upload & Send Files to Users

Every node on the path from the source node to the destination node becomes a cluster head, with the task of recruiting other nodes in its neighborhood and coordinating their transmissions. Consequently, the classical route from a source node to a sink node is replaced with a multihop cooperative path, and the classical point-to-point communication is replaced with many-to-many cooperative communication. In this module, server can upload the files in the database.

4.3 Best Path Estimation

Every node on the path from the source node to the destination node becomes a cluster head, with the task of recruiting other nodes in its neighborhood and coordinating their transmissions. Consequently, the classical route from a source node to a sink node is replaced with a multihop cooperative path, and the classical point-to-point communication is replaced with many-to-many cooperative communication. The path can then be described as "having a width," where the "width" of a path at a particular hop is determined by the number of nodes on each end of a hop. Each hop on this path represents communication from many geographically close nodes, called a sending cluster, to another cluster of nodes, termed a receiving cluster. The nodes in each cluster cooperate in transmission of packets, which propagate along the path from one cluster to the next.

4.4 Bayesian Traffic Classification

Bayesian algorithms (i.e. classification algorithms) can help to classify the network traffic data for intrusion prediction. Many supervised learning algorithms have been developed for the classification of network intrusion traffic among them the Bayesian classification received a greater attention. However the computational cost of the Bayesian is in general higher than many other classification techniques. To ease this problem many Bayesian techniques have been subsequently developed in the machine learning research . Due to their computational complexity, they are not suitable for Big Data analytics. However the classification accuracies that the Bayesian techniques give are excellent.

4.5 Apply CRT

The security mechanism in cloud technology is generally weak. Hence tampering of data at the public cloud is inevitable and it is a big concern. Finding a robust security mechanism for the purpose of using we implement the CRT (Chinese Reminder Theorem) for Data packet id, in which a node starts at a random position, Here we are Apply Prime Numbers to the Data packets for some Security Purpose. For these purpose intruder won't identify the data packets order.

- Throughput
- Packet delivery fraction
- End to End delay
- Normalized routing load

6. EXPRIMENTAL RESULTS

To evaluate the accuracy of the proposed method, we choose to use two network traffic data sets for experiments and compare the performance of the proposed method and other traffic classification methods. Our results show that the proposed method is resilient to mislabeled data.

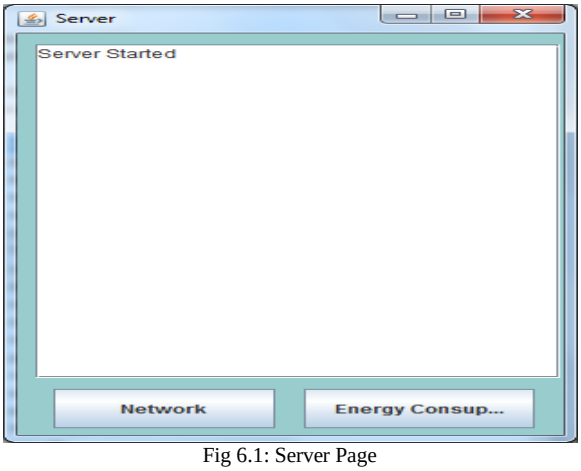

Fig 6.1: Server Page

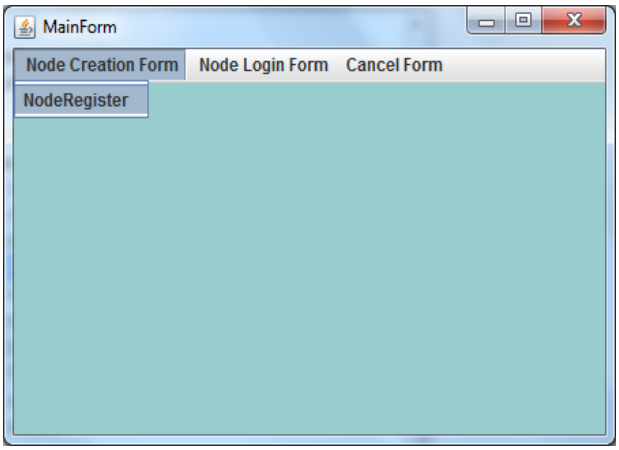

Fig 6.2: Node Register Page

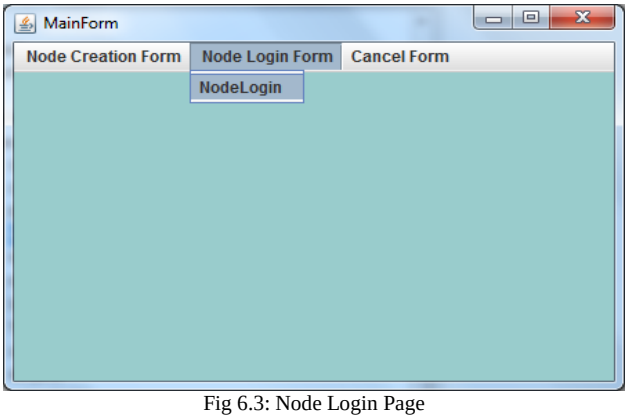

Fig 6.3: Node Login Page

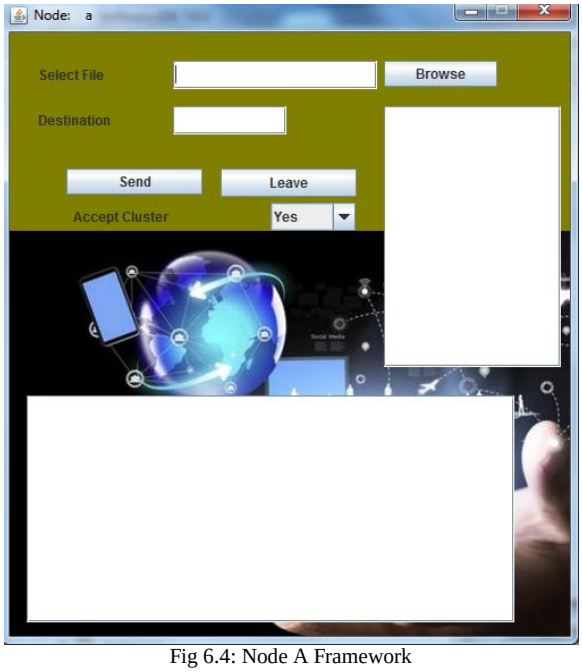

Fig 6.4: Node A Framework

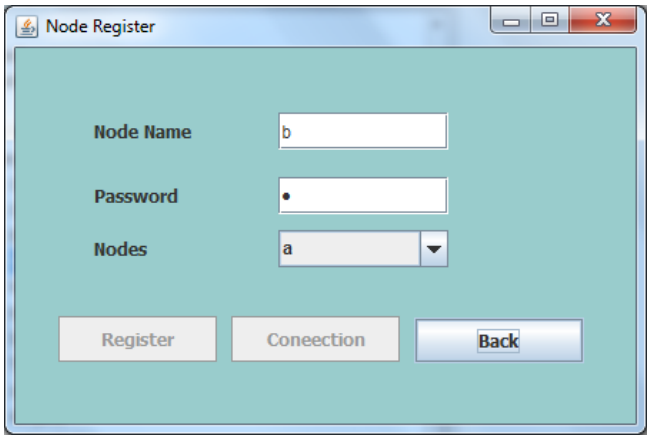

Fig 6.5: Node b connection with node a

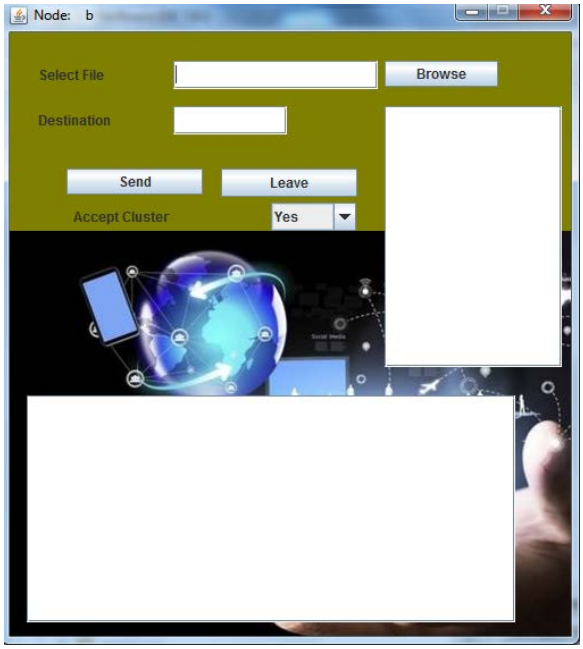

Fig 6.6: Node B Framework

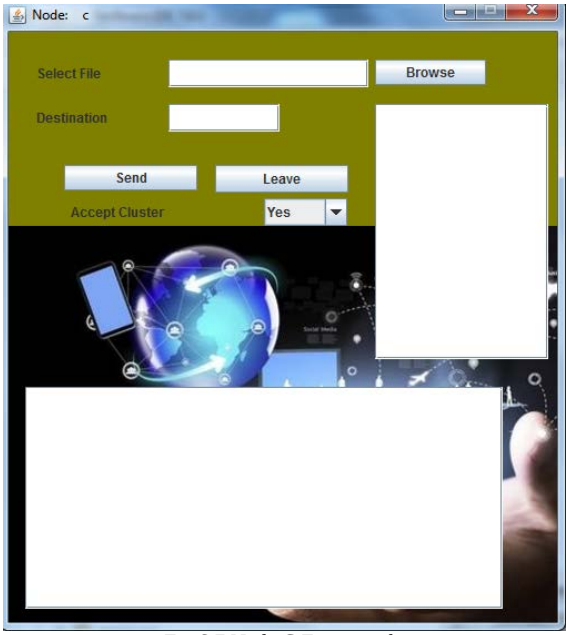

Fig 6.7 Node C Framework

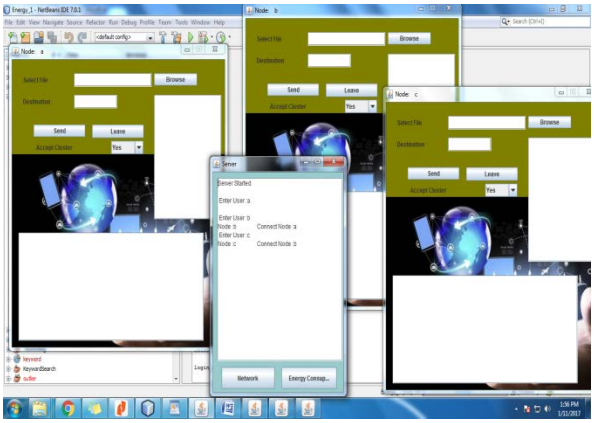

Fig 6.8: Energy Frame works

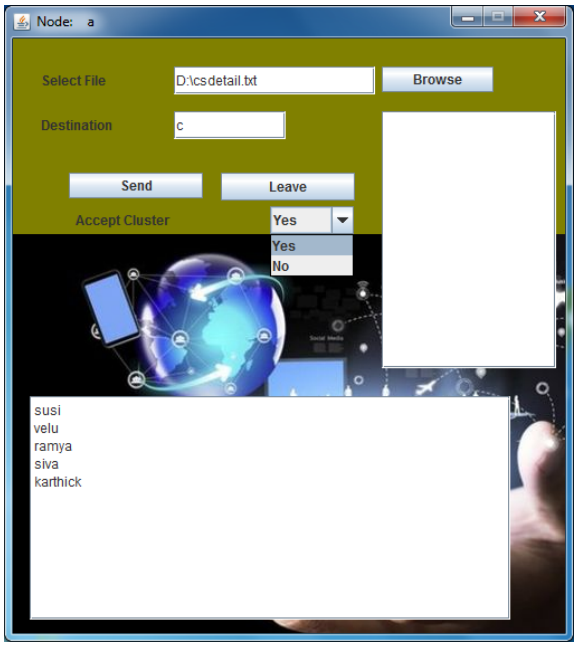

Fig 6.9: Find the destination and accepted cluster

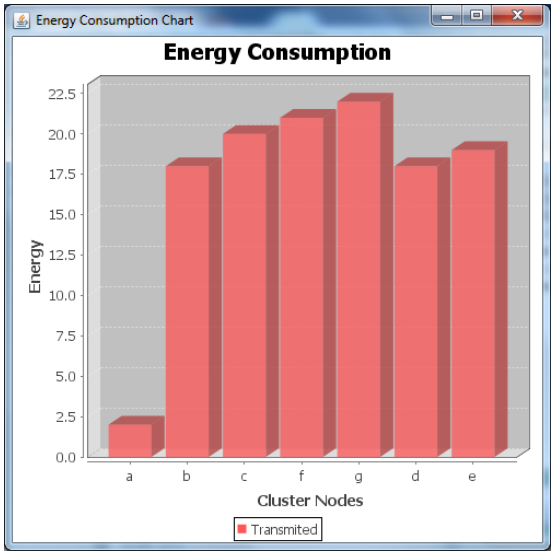

Fig 6.10: Energy Minimization Graph

7. CONCLUSION

The continuous collection of traffic data by the network leads to Big Data problems that are caused by the volume, variety and velocity properties of Big Data. The learning of the network characteristics requires machine learning techniques that capture global knowledge of the traffic patterns. In our model of cooperative transmission, every node on the path from the source node to the destination node becomes a cluster head, with the task of recruiting other nodes in its neighborhood and coordinating their transmissions. Consequently, the classical route from a source node to a sink node is replaced with a multi-hop cooperative path, and the classical point-to-point communication is replaced with many-to-many cooperative communication.

REFERENCE

[1] T. T. T. Nguyen and G. Armitage, "A survey of techniques for Internet traffic classification using machine learning," IEEE Commun. Surveys Tuts., vol. 10, no. 4, pp. 56–76, 4th Quart., 2008.
[2] S. Mavoungou, G. Kaddoum, M. Taha, and G. Matar, "Survey on Threats and Attacks on Mobile Networks," IEEE Access, vol. 4, no. , pp. 4543– 4572, 2016.
[3] Y. Xiang, W. Zhou, and M. Guo, "Flexible deterministic packet marking: An IP traceback system to find the real source of attacks," IEEE Trans. Parallel Distrib. Syst., vol. 20, no. 4, pp. 567–580, Apr. 2009.
[4] M. A. Ashraf, H. Jamal, S. A. Khan, Z. Ahmed, and M. I. Baig, "A Heterogeneous Service-Oriented Deep Packet Inspection and Analysis Framework for Traffic-Aware Network Management and Security Systems," IEEE Access, vol. 4, pp. 5918–5936, 2016.
[5] J. Zhang, X. Chen, Y. Xiang, W. Zhou, and J. Wu, "Robust Network Traffic Classification," IEEE/ACM Trans. Netw., vol. 23, no. 4, pp. 1257– 1270, Aug. 2015.
[6] S. Tatinati, K. C. Veluvolu and W. T. Ang, "Multistep Prediction of Physiological Tremor Based on Machine Learning for Robotics Assisted Microsurgery," IEEE Trans. Cybern., vol. 45, no. 2, pp. 328–339, Feb. 2015.
[7] D. Kelly and B. Caulfield, "Pervasive Sound Sensing: A Weakly Supervised Training Approach," IEEE Trans. Cybern., vol. 46, no. 1, pp.123–135, Jan. 2015.
[8] A. W. Moore and D. Zuev, "Internet traffic classification using Bayesian analysis techniques," in Proc. SIGMETRICS, vol. 33, pp. 50–60, Jun. 2005.
[9] T. Auld, A. Moore, and S. Gull, "Bayesian neural networks for internet traffic classification," IEEE Trans. Neural Netw., vol. 18, no. 1, pp. 223– 239, Jan. 2007.
[10] A. Este, F. Gringoli, and L. Salgarelli, "Support vector machines for TCP traffic classification," Comput. Netw., vol. 53, no. 14, pp. 2476– 2490, 2009.
[11] B. Hull´ar, S. Laki, and A. Gyorgy, "Early identification of peer-to-peer traffic," in Proc. IEEE Int. Conf. Commun., pp. 1–6, 2011.
[12] T. T. T. Nguyen and G. Armitage, "Training on multiple sub-flows to optimize the use of machine learning classifiers in real-world IP networks," in Proc. 31st IEEE Conf. Local Comput. Netw., pp. 369–376, 2006.
[13] P. Bermolen, M. Mellia, M. Meo, D. Rossi, and S. Valenti, "Abacus: Accurate behavioral classification P2P-TV traffic," Comput. Netw., vol. 55, no. 6, pp. 1394–1411, 2011.
[14] J. Zhang, Y. Xiang, Y. Wang, W. Zhou, Y. Xiang, and Y. Guan, "Network traffic classification using correlation information," IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 1, pp. 104–117, Jan. 2013.
[15] E. Glatz and X. Dimitropoulos, "Classifying internet one-way traffic," in Proc. ACM SIGMETRICS/PERFORMANCE Joint Int. Conf. Meas. Model. of Comput. Syst., pp. 417–418, 2012.
[16] Y. Jin, N. Duffield, J. Erman, P. Haffner, S. Sen, and Z.-L. Zhang, "A modular machine learning system for flow-level traffic classification in large networks," ACM Trans. Knowl. Discov. Data, vol. 6, no. 1, pp. 4:1–4:34, Mar. 2012.
[17] A. Callado, J. Kelner, D. Sadok, C. Alberto Kamienski, and S. Fernandes, "Better network traffic identification through the independent combination of techniques," J. Netw. Comput. Appl., vol. 33, no. 4, pp. 433–446, 2010.
[18] G. Xie, M. Iliofotou, R. Keralapura, M. Faloutsos, and A. Nucci, "Sub- Flow: Towards practical flow-level traffic classification," in Proc. IEEE INFOCOM, pp. 2541–2545, 2012.
[19] M. AlSabah, K. Bauer, and I. Goldberg, "Enhancing Tor's performance using real-time traffic classification," in Proc. ACM Conf. Computer and Comm. Security, pp. 73–84, 2012.
[20] R. Wang, L. Shi, and B. Jennings, "Ensemble classifier for traffic in presence of changing distributions," in Proc. IEEE Symposium Comput. Commun., pp. 629–635, 2013.
[21] L. Grimaudo, M. Mellia, and E. Baralis, "Hierarchical learning for fine grained internet traffic classification," in Proc. IEEE Int. Conf. IWCMC, pp. 463–468, 2012.
[22] T. T. T. Nguyen, G. Armitage, P. Branch, and S. Zander, "Timely and continuous machine-learning-based classification for interactive IP traffic," IEEE/ACM Trans on Nets., vol. 20, no. 6, pp. 1880–1894, Dec. 2012.
[23] M. Jaber, R. G. Cascella, and C. Barakat, "Using host profiling to refine statistical application identification," in Proc. IEEE INFOCOM, pp. 2746–2750, 2012.
[24] J. Erman, A. Mahanti, M. Arlitt, I. Cohenz, and C. Williamson, "Semisupervised network traffic classification," SIGMETRICS Perform. Eval. Rev., vol. 35, no. 1, pp. 369–370, 2007.
[25] X. Li, F. Qi, D. Xu, and X. Qiu, "An internet traffic classification method based on semi-supervised support vector machine," in Proc. IEEE ICC. pp. 1–5, 2011.