ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

Analysis of a Novel Non-Parametric Approach in Traffic Classification Using Correlation (TCC) Information Framework

Dheeraj S. Sadawarte[1], Shital Y. Gaikwad[2]

M. E. (II[nd] Year Student), Department of CSE, MPGI College of Engineering, Kupsurwadi, Nanded. (M.S.), India[1]

Asst. Professor, Department of CSE, MPGI MPGI College of Engineering, Kupsurwadi, Nanded. (M.S.), India[2]

ABSTRACT: Traffic classification is widely used for network management, from security supervising to the quality of service judgment. Recent studies, as a rule, apply machine learning methods to classification method based on statistical properties. The nearest neighbor method (NN) shows excellent classification performance. There are some relevant advantages like, it does not require a training method, excessive use of parameters and, of course, there is adequate to handle many classes. Though, if the size of training data has been compromised, then the performance of the NN classification can be greatly reduced. This paper proposes an innovative non-parametric method of traffic classification, which excellently increases the performance of classification, corresponds to the relevant information in the classification process. Analysis of the new classification method and its effectiveness gain from the theoretical and experimental approach.The contribution work is to implement the combination of K-means and Expectation Maximization (EM) unsupervised clustering algorithm. The list of experiments is working on the real-time traffic datasets to verify the proposed method with TCC. The results show that the performance of traffic classification will be increased according to challenging prospects of very few training samples.

KEYWORDS: Traffic Classification, Network Operations, Security, Correlation, Bag of Flows

I. INTRODUCTION

Classifying network traffic flows by their generation applicationsassumes critical part in network security and management,such as quality of service (QoS) control, legal capture and intrusion detection. There is a need to identify drifts and classify network traffic into various classes and after that to investigate the behavior of traffic to detect attacks. The greater part of the network administrators are stimulated on catching client's information and anticipate about clients conduct for getting network resources by investigating this information. In this procedure network administrators focus on the end purposes of the network yet they nearly disregard the overallbehavior of their networks. Network classification is required to discover the networkfeatures and the overall behavior. By evaluatingthe networktraffic numerous attacks can be caught.

Network classification is necessary to detect normal and anomalous behaviors. Data mining equipment play a very important role in the classification of network traffic. An unsupervised learning method is used for grouping similar traffic flow to characterize network traffic. By use of mixed type attributes like numerical, categorical and hierarchal in clustering algorithm network traffic is characterized. The supervised traffic classification can be split into two parts: parametric classifiers and nonparametric classifiers. Parametric classifiers, such as decision tree, Bayesian network, SVM, and neural networks, require an intensive training procedure for the classifier parameters. Nonparametric classifiers, e.g., k-Nearest Neighbor (k-NN), usually require no training phase and make classification decision based on closest training samples in the feature space. When k=1, the NN-based traffic classifier assigns a testing traffic flow into the class of its nearest training sample.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

This paper proposes a new framework, Traffic Classification using Correlation (TCC) information, to address the problem of very few training samples. The correlation information in network traffic can be used to effectively improve the classification accuracy. The major contributions of this work are summarized as follows:

- Proposes a novel nonparametric approach which incorporates correlation of traffic flows to improve the classification performance.
- Provides a detailed analysis on the novel classification approach and its performance benefit from both theoretical and empirical aspects.
- The performance evaluation shows that the traffic classification using very few training samples can be significantly improved by our approach.

Motivation:
1. The Detection accuracy of anomaly system should maximize.
2. The false positive rate should minimum.
3. The detector generation time should less.

Objectives:
1. To effectively improve the network traffic classification accuracy in a robust way.
2. To correlate information in the traffic flows using flow correlation analysis.
3. To discover correlation information in the traffic flows and incorporates it into the classification process.
4. To measure the efficiency of the BoF discovery method by calculating some statistics of five real-worlds traffic data sets.
5. To compare the NN classifiers with the proposed algorithm this can effectively improve the overall performance of network traffic classification.

II. REVIEW OF LITERATURE

The paper [1] proposes a new analytical model evaluates the performance of communication networks based on the m-port n-tree topology in multi-cluster systems in the presence of the spatio-temporal bursty traffic. The spatial traffic burstiness is caught by the communication locality and the transient traffic burstiness is displayed by the Markov-modulated Poisson process (MMPP) which can capture the properties of bursty message arrivals while remaining systematically tractable. Advantages are: The communication locality can decrease the degrading effects of bursty message arrivals on the network performance of multicluster systems. The proposed model is utilized to research the effect of bursty message arrivals and communication locality on network performance. Disadvantages are: The proposed analytical model does not handle Mesh-of-Tree (MoT) topology.

The paper [2] reveals the three sources of the discriminative power in classifying the Internet application traffic: (i) ports, (ii) the sizes of the first one-two (for UDP flows) or four-five (for TCP flows) packets, and (iii) discretization of those features. First preprocess the dataset with the Entropy-based discretization method with the Minimum Description Length criterion (Ent-MDL), after apply machine learning algorithms on the discretized dataset, and then compare their performance with those of the same algorithms with the original non-discretized data. Advantages are: The entropy-based discretization is the essential technique for accurate traffic classification achieves high and similar accuracy 93%. The entropy-based discretization is more effective on ports than packet size information. Disadvantages are: The classification accuracy decreases when the first few packets are missed; thus, for accurate classification, application traffic flows should be captured from the beginning, not in the midst.

The paper [3] presents a novel and practical IP traceback system, Flexible Deterministic Packet Marking (FDPM). A novel and practical packet marking traceback system, incorporating a flexible mark length strategy and flexible flow-based marking scheme, is proposed. FDPM is suitable for not only tracing sources of DDoS attacks but also DDoS detection. Advantages are: FDPM provides more flexible features to trace IP packets than other packet marking schemes, and can obtain better tracing capacity. No additional bandwidth requirement and less computation overhead.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

The paper [4] represents a simple but effective procedure for the optimization of the working parameters of a statistical network traffic classifier. The proposed classification method is based on statistical pattern recognition from which derived the definitions of features, patterns and classes. The optimization process on one hand can avoid the network administrator a manual and imprecise tuning procedure of the classification engine. Advantages are:Improves the precision of the classifier.

The [5] paper describes an approach to traffic classification based on SVM. Apply one of the approaches to solving multi-class problems with SVMs to the task of statistical traffic classification, and describe a simple optimization algorithm that allows the classifier to perform correctly with as little training as a few hundred samples. SVM-based traffic classifier integrates the SVM ‘‘one-against-all” approach to solve multi-class problems when needed. Advantages are: The accuracy of the classifier is very good with 90%. Disadvantages are: Needs to deal with out-of-order packets, packet loss, and fragmentation in a robust way.

The paper [6] looks at emerging research into the application of Machine Learning (ML) techniques to IP traffic classification - an inter-disciplinary blend of IP networking and data mining techniques. It provides the rationale for IP traffic classification in IP networks, review the state-of-the-art approaches to traffic classification, and then review and critique emerging ML-based techniques for IP traffic classification. Advantages are: Efficient use of memory and processors for the classification. A model may be considered portable if it can be used in a variety of network locations and robust if it provides consistent accuracy in the face of network layer perturbations such as packet loss, traffic shaping, packet fragmentation, and jitter.

The paper [7] works on critically revisit traffic classification by conducting a thorough evaluation of three classification approaches, based on transport layer ports, host behavior, and flow features. The effectiveness of port-based classification in identifying legacy applications is still impressive, further strengthened by the use of packet size and TCP flag information. The host behavior-based methods such as BLINC can be augmented with per-flow based classification process to increase byte accuracy. The SVM consistently achieved the highest accuracy. Advantages are: 1. CoralReef: Fast, easy-to-use, and scalable. Good at identifying conventional applications (except Passive FTP). 2. BLINC: Good at identifying P2P flows. Detects new as well as encrypted applications. 3. Flow features based learning: Automatic learning. Highly accurate and robust to link and traffic conditions. Disadvantages are: 1. CoralReef: Port-masquerading applications 2. BLINC: Tuning overheads. Strongly depends on the topological locations and traffic mixes. Low bytes accuracy. 3. Flow features based learning: High computational overheads (SVM, Neural Net., and k-NN) and memory requirements (C4.5).

The [8] paper proposes a methodology that classifies (or equivalently, identifies) network flows by application using only flow statistics. Our methodology, based on machine learning principles, consists of two components: a learner and a classifier. It proposes a semi-supervised framework for classifying network traffic using only flow statistics. Advantages are: Achieves high flow and byte accuracy. A variety of applications, including P2P, HTTP, FTP, and email can be successfully classified. Robust classifiers can be built that are immune to transient changes in network conditions.

The paper [9] represents the Bayesian framework using a neural network model that allows identification of traffic without using any port or host [Internet protocol (IP) address] information; Though this technique uses training data with categories derived from packet content, training and testing were done using features derived from packet streams consisting of one or more packet headers. Advantages are: Bayesian trained neural network is able to classify flows, based on header-derived statistics and no port or host (IP address) information, with from 95% to 99% accuracy. A small number of features carry high significance regardless of this classification scheme. Disadvantages are: Some issues are the selection of optimum features.

In [10] paper, proposes and evaluates a machine learning approach for identifying and grouping network traffic according to traffic classes (e.g., Web, P2P, FTP, Others) at egress and ingress points of core networks. An algorithm is capable of estimating statistics from unidirectional traces such as number of bytes and the number of packets of the unseen portions of the flow. Advantages are: Better classification performance is achieved when statistics for the server-to-client direction are used than when statistics for the client-to-server direction are used. Disadvantages are: The collection of the server-to-client statistics may not always be feasible.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

III. PROPOSED SYSTEM

Network traffic Classification is an important task in operational large-scale informatics networks. The Network should be able make sure that the network is ready to transfer traffic properly altogether with conditions of the security levels of the organization. Machine learning methods try attempts to solve intelligence problems without the needs of human expertise, however they are not perfect. This paper proposed a novel technique for classifying network traffic using machine learning approaches.This novel technique presents a new framework which calls Traffic Classification using Correlation information or TCC for short. A novel nonparametric approach is also proposed to effectively incorporate flow correlation information into the classification process. Fig. 1 shows the proposed system model. In the preprocessing, the system catches IP packets crossing a computer network and develops traffic flows by IP header examination. A flow consists of successive IP packets having the same five-tuple: {src_ip, src_port, dst_ip, dst_port, protocol}. After that, a set of statistical features are extracted to represent each flow. In this paper, we use "bag of flows" (BoF) to model correlation information in traffic flows. The BoF model-based classification approach describes that a BoF can be classified by aggregating the prediction values of flows produced by the NN classifier. The proposed classification approach can apply different aggregation strategies, which have different meanings.

- AVG-NN: combines multiple flow distances to make a decision for a BoF
- MIN-NN: selects a minimum flow distance to make a decision for a BoF
- MVT-NN: combines multiple decisions on flows to make a final decision for a BoF.

AVG-NN shows better performance than MIN-NN and MVT-NN in terms of overall performance, perexperiment performance, and per-class performance. TCC is better thanthe existing traffic classification methods since it shows the capacity of applying flow correlation to effectively enhance traffic classification performance.

The contribution work is to implement K-means + Expectation Maximization unsupervised clustering algorithm. In the training phase it randomly generate K cluster. In the classification phase these randomly generated clusters further classify the unknown traffic using Euclidean distance formula. The Expectation Maximization (EM) follows an iterative approach, sub-optimal, which tries to find the parameters of the probability distribution that has the maximum likelihood of its attributes. This approach gives good accuracy and overcome the drawback of the NN-classifier.

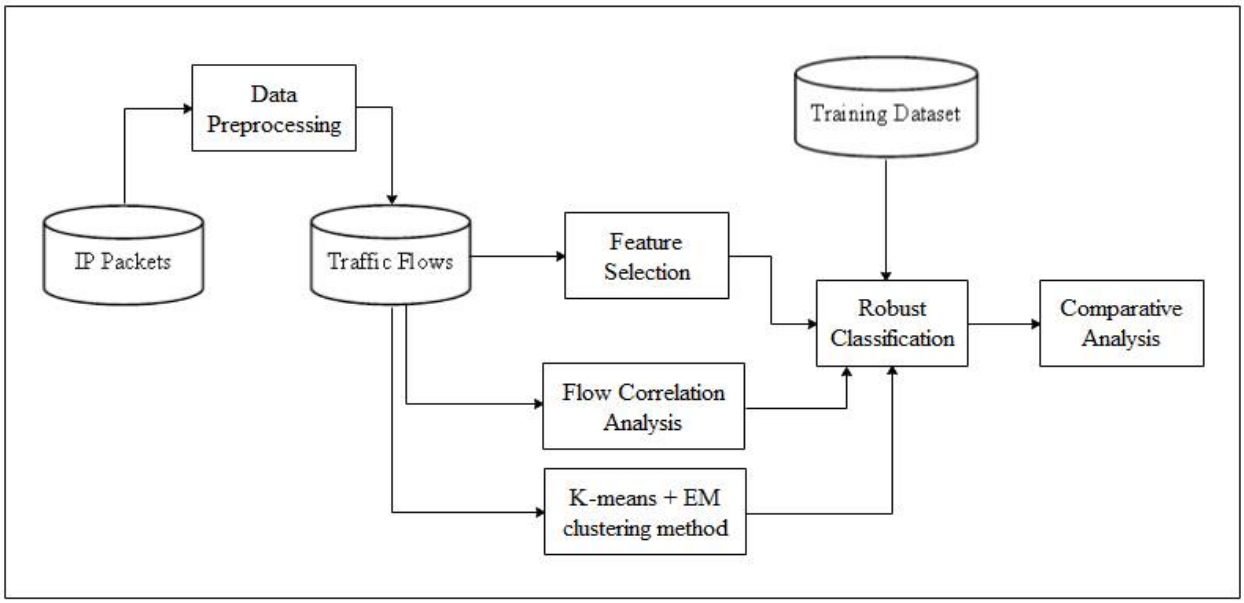

Fig. 1 Proposed Architecture

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

Advantages:

1. According to the proposed methods of the overall performance of traffic classification is better than the NN classifier.
2. The proposed methods can improve the classification accuracy in a robust way and consistent improvement is achieved.
3. The proposed methods can improve the F-measure of every class and significant improvements are obtained in most classes.
4. AVG-NN shows better performance than MIN-NN and MVT-NN in terms of overall performance, per experiment performance, and per-class performance.
5. TCC is superior to the existing traffic classification methods since it demonstrates the ability of applying flow correlation to effectively improve traffic classification performance.
6. The combination of K-means and Expectation Maximization unsupervised clustering method gives higher accuracy than the TCC.
7. K-means and EM produces good quality of clusters for various traffic applications.

IV. MATHEMATICAL MODEL

1. BoF:

A "Bag-of-flows" (BoF) consists of some correlated traffic flows which are generated by the same application.
A BoF can be described by

a. $$Q = \{X_1, X_2, \dots, X_n\} \quad (1)$$

Where X_i is a feature vector representing the ith flow in the BoF Q.
Q denotes the correlation among n flows.

2. NN Classifier:

The prediction value of a flow x produced by the NN classifier is defined using its minimum distance to the training samples of class ω.

$$d_x = \min_{x' \epsilon \omega} \|x - x'\|^2 \quad (2)$$

3. AVG-NN Classifier:

The distance of a query BoF Q to class ω is obtained by aggregating the flow distances with the "average" operation as follows:

$$d_Q^{avg} = \frac{1}{\|Q\|} \sum_{x \epsilon Q} d_x \quad (3)$$

Finally, the flows in Q are classified into the class with the minimum distance of Q.

4. MIN-NN Classifier:

$$\omega^* = \arg\min_{\omega} \left(\min_{x \epsilon Q} \min_{x' \epsilon \omega} \|x - x'\|^2 \right) \quad (4)$$

5. MVT-NN Classifier:

The binary prediction values produced by the NN classifier to conduct classification of BoFs.

$$\omega_x^* = \arg\min_{\omega} \left(\min_{x' \epsilon \omega} \|x - x'\|^2 \right) \quad (5)$$

We define the vote of a flow x for class ω as

$$v_\omega(x) = \begin{cases} 1, & for\ \omega\ is\ \omega_x^* \\ x, & else \end{cases} \quad (6)$$

$$\omega^* = \arg\max_{\omega} \left(\sum_{x \epsilon Q} v_\omega(x) \right) \quad (7)$$

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

Algorithms:

Proposed Algorithm (K-means + EM)

Steps:

1. Initially, randomly assigns k cluster centers
2. Iteratively refine the clusters based on two steps
3. Expectation step: Assign each data point X_i to cluster C_i with the following probability

$$P(X_i \epsilon C_i) = P\left(\frac{C_k}{X_i}\right) = P(C_k)P(\frac{X_i}{C_k}) / P(X_i)$$

4. Maximization step: Use the probability estimates from above to re-estimate (or refine) the model parameter.

$$m = \frac{1}{N}\sum_{i=1}^{N}\frac{X_i P(X_i \epsilon C_k)}{\sum_{j=1}^{k} P(X_i \epsilon C_j)}$$

V. EXPERIMENTAL RESULT

In this section, we measure the performance analysis usingthe NIMS dataset for traffic classification. The performance measure using the overall accuracy, F-measure and execution time of existing system like MIN-NN classifier, AVG-NN classifier and MVT-NN classifiers. Also compare performance with the proposed K-means + EM classifier. The NIMS dataset features are shown below:

TABLE I NMIS Dataset Features

S. No.	Feature	Description
1	src_ip	Source IP address
2	src_port	Source port number
3	dst_ip	Destination IP address
4	dst_port	Destination port number
5	min_fpktl	Minimum of forward packet length
6	mean_fpktl	Mean of forward packet length
7	max_fpktl	Maximum of forward packet length
8	std_fpktl	Standard deviation of forward packet length
9	min_bpktl	Minimum of backward packet length
10	mean_bpktl	Mean of backward packet length
11	max_bpktl	Maximum of backward packet length
12	std_bpktl	Standard deviation of backward packet length
13	min_fiat	Minimum of forward inter-arrival time
14	mean_fiat	Mean of forward inter-arrival time
15	max_fiat	Maximum of forward inter-arrival time
16	std_fiat	Standard deviation of forward inter-arrival times
17	min_biat	Minimum of backward inter-arrival time
18	mean_biat	Mean backward inter-arrival time
19	max_biat	Maximum of backward inter-arrival time
20	std_biat	Standard deviation of backward inter-arrival times
21	duration	Total duration
22	proto	Protocol
23	total_fpackets	Total number of packets in forward direction
24	total_fvolume	Total number of bytes in forward direction
25	total_bpackets	Total number of packets in backward direction
26	total_bvolume	Total number of bytes in backward direction
27	class	Class

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

For experimental set up, use Windows 7 operating system, Intel i5 processor, 4 GB RAM, 200GB Hard disk, Eclipse Luna JDK 7 tool. To calculate the results, NIMS data set is used. In training dataset, there are 11 types of are included.For performance comparison, four classification methods are implementedusing the NMIS dataset, MIN-NN, AVG-NN, MVT-NN and proposed K-means + Expectation Maximization unsupervised clustering algorithm. The 1st three methods can incorporatecorrelation information into the classification process whereas the proposed method does not take into account correlation information in traffic flows.

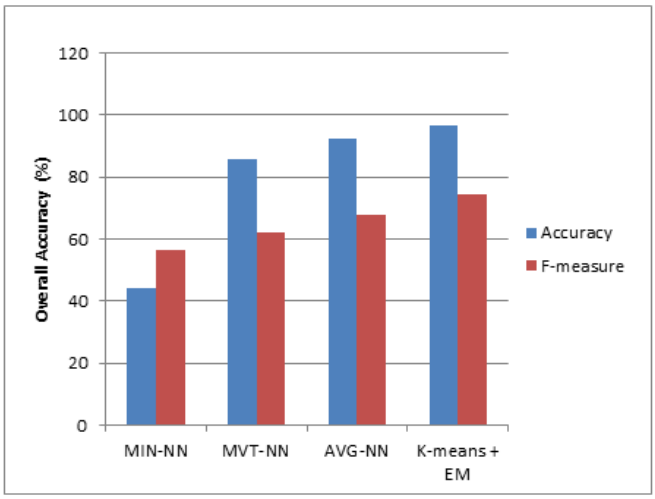

Fig.2 Comparison analysis on overall accuracy and f-measure

The Fig. 2 shows the AVG-NN classifier's accuracy is greater than MIN-NN and MVT-NN classifiers. The K-means + EM classifier's overall accuracy is greater than AVG-NN classifier.

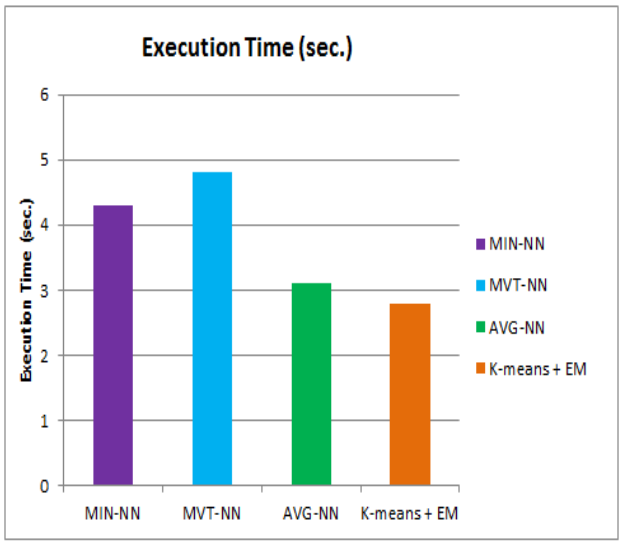

Fig.3 Comparison analysis on execution time of traffic classification

The Fig. 3 shows the existing and proposed classifier algorithm analysis on execution time of traffic classification. The MVT-NN classifier's execution time is more than other classifiers. The AVG-NN classifier's execution time required is

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

less than MIN-NN and MVT-NN classifiers. The proposed classifier algorithm requires execution time is very less than other classifiers.

V. CONCLUSION AND FUTURE WORK

A novel nonparametric technique, TCC proposes to research correlation data in real traffic data and incorporate it into traffic classification. It represents a comprehensive analysis on the system framework and performance benefit from both theoretical and experimental approaches, which strongly support the proposed approach. There are three new classification methods, AVG-NN, MIN-NN, and MVT-NN, are proposed for interpretation, which can incorporate correlation information into the class prediction for improving classification performance. The combination of K-means and Expectation Maximization (EM) unsupervised clustering algorithm gives higher accuracy than the TCC. K-means and EM produces good quality of clusters for various traffic applications.The proposed approach can be used in a wide range of applications, such as automatic recognition of unknown applications from captured network traffic and semi-supervised data mining for processing network packets.

REFERENCES

[1] Y. Wu, G. Min, K. Li, and B. Javadi, "Modelling and Analysis of Communication Networks in Multi-Cluster Systems Under Spatio-Temporal Bursty Traffic," IEEE Trans. Parallel Distributed Systems, vol. 23, no. 5, pp. 902-912, May 2012, http://dx.doi.org/10.1109/TPDS.2011.198.
[2] Y.-s. Lim, H.-c. Kim, J. Jeong, C.-k. Kim, T.T. Kwon, and Y. Choi, "Internet Traffic Classification Demystified: on the Sources of the Discriminative Power," Proc. Sixth Int'l Conf. (Co-NEXT '10), pp. 9:1-9:12, 2010.
[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 Distributed Systems, vol. 20, no. 4, pp. 567-580, Apr. 2009.
[4] M. Crotti, F. Gringoli, and L. Salgarelli, "Optimizing Statistical Classifiers of Network Traffic," Proc. Sixth Int'l Wireless Comm. And Mobile Computing Conf., pp. 758-763, 2010.
[5] A. Este, F. Gringoli, and L. Salgarelli, "Support Vector Machines for TCP Traffic Classification," Computer Networks, vol. 53, no. 14, pp. 2476-2490, Sept. 2009.
[6] T.T. Nguyen and G. Armitage, "A Survey Of Techniques for Internet Traffic Classification Using Machine Learning," IEEE Comm. Surveys Tutorials, vol. 10, no. 4, pp. 56-76, Oct.-Dec. 2008.
[7] H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee, "Internet Traffic Classification Demystified: Myths, Caveats, and the Best Practices," Proc. ACM CoNEXT Conf., pp. 1-12, 2008.
[8] J. Erman, A. Mahanti, M. Arlitt, I. Cohen, and C. Williamson, "Offline/Realtime Traffic Classification Using Semi-Supervised Learning," Performance Evaluation, vol. 64, nos. 9-12, pp. 1194-1213, Oct. 2007.
[9] T. Auld, A.W. Moore, and S.F. Gull, "Bayesian Neural Networks for Internet Traffic Classification," IEEE Trans. Neural Networks, vol. 18, no. 1, pp. 223-239, Jan. 2007.
[10] J. Erman, A. Mahanti, M. Arlitt, and C. Williamson, "Identifying and Discriminating between Web and Peer-to-Peer Traffic in the Network Core," Proc. 16th Int'l Conf. World Wide Web, pp. 883-892, 2007.