International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Review on Network Traffic Classification

Kuldeep Singh[1], Manoj Kumar[2]

[1]SGT Institute of Engineering and Technology, Gurgoan, Haryana, India

[2]Assistant Professor (CSE), SGT Institute of Engineering and Technology, Gurgoan, Haryana, India

Abstract: ***Traffic classification has wide applications in network management, from security monitoring to quality of service measurements. Recent research tends to apply machine learning techniques to flow statistical feature based classification methods. The nearest neighbor (NN)-based method has exhibited superior classification performance. It also has several important advantages, such as no requirements of training procedure, no risk of over fitting of parameters, and naturally being able to handle a huge number of classes. However, the performance of NN classifier can be severely affected if the size of training data is small. In this paper, we propose a novel nonparametric approach for traffic classification, which can improve the classification performance effectively by incorporating correlated information into the classification process. We analyze the new classification approach and its performance benefit from both theoretical and empirical perspectives. A large number of experiments are carried out on two real-world traffic data sets to validate the proposed approach. The results show the traffic classification performance can be improved significantly even under the extreme difficult circumstance of very few training samples.***

Keywords: SVM, P2P, TCC, TCP, AVG-NN, MINNN ANDMVT-NN

1. Existing System

As reported in the NN classifier can achieve superior performance similar to that of the parametric classifiers, SVM and neural nets. They are the top three out of seven evaluated machine learning algorithms. In contrast to the parametric classifiers, the NN classifier has several important advantages. For example, it does not require training procedure immunizes over fitting of parameters and is able to handle a huge number of classes. In this point of view, the NN classifier is more suitable for traffic classification in current complex network environment. However, the performance of the NN classifier is severely affected by a small size of training data which cannot accurately represent the traffic classes. We have observed that the classification accuracy of the NN-based traffic classifier decreases by approximate 20 percents when the number of training samples reduces from 100 to 10 for each class. Other supervised classification methods, such as SVM and neural nets, are not robust to training data size either. Machine learning can automatically search for and describe useful structural patterns in a supplied traffic data set, which is helpful to intelligently conduct traffic classification. However, the problem of accurate classification of current network traffic based on flow statistical features has not been solved. The flow statistical feature-based traffic classification can be achieved by using supervised classification algorithms or unsupervised classification.

Disadvantages

- A novel classification approach and the theoretical analysis are proposed in presents a large number of experiments and results for performance evaluation.
- The supervised traffic classification methods analyze the supervised training data and produce an inferred function which can predict the output class for any testing flow.

2. Proposed System

The problems suffered by payload-based traffic classification, such as encrypted applications and user data privacy, Moore and applied the supervised naive techniques to classify network traffic based on flow statistical features. Evaluated the supervised algorithms including naive Bayes with discretization, naive Bayes with kernel density estimation, C4.5 decision tree, Bayesian network, and naive Bayes tree. Nguyen and Armitage proposed to conduct traffic classification based on the recent packets of a flow for real-time purpose. Extended the work of with the application of Bayesian neural networks for accurate traffic classification. used unidirectional statistical features for traffic classification in the network core and proposed an algorithm with the capability of estimating the missing features. Proposed to use only the size of the first packets of an SSL connection to recognize the encrypted applications proposed to analyze the message content randomness introduced by the encryption processing using Pearson's chi-Square test-based technique. The probability density function (PDF)-based protocol fingerprints to express three traffic statistical properties in a compact way. Their work is extended with a parametric optimization procedure.

Advantages

- These works use parametric machine learning algorithms, which require an intensive training procedure for the classifier parameters and need the retraining for new discovered applications.
- Evaluated three supervised methods for an ADSL provider managing many points of presence, the results of which are comparable to deep inspection solutions.
- Applied one class SVMs to traffic classification and presented a simple optimization algorithm for each set of SVM working parameters proposed to classify P2P-TV traffic using the count of packets exchanged with other peers during the small time windows.

3. Modules Description

Supervised Methods

The supervised traffic classification methods analyze the supervised training data and produce an inferred function which can predict the output class for any testing flow. In supervised traffic classification, sufficient supervised training data is a general assumption. To address the problems suffered by payload-based traffic classification, such as encrypted applications and user data.

Unsupervised Methods

The unsupervised methods (or clustering) try to find cluster structure in unlabeled traffic data and assign any testing flow to the application-based class of its nearest cluster. The proposed to group traffic flows into a small number of clusters using the expectation maximization (EM) algorithm and manually label each cluster to an application.

A Traffic Classification Approach with Flow Correlation

The presents a new framework which we call 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.

Correlation Analysis

The correlated flows sharing the same three-tuple are generated by the same application. For example, several flows initiated by different hosts are all connecting to a same host at TCP port 80 in a short period. These flows are very likely generated by the same application such as a web browser. The three-tuple heuristic about flow correlation has been considered in several practical traffic classification schemes proposed a payload based clustering method for protocol inference, in which they grouped flows into equivalence clusters using the heuristic. tested the correctness of the three-tuple heuristic with real-world traces.

Computational Performance

The computational performance includes learning time, amount of storage, and classification time. First, the NN classifier does not really involve any learning process, which is shared with our proposed methods. However, other supervised methods, such as neural nets and SVM, need time to learn parameters for their classification model. Second, the proposed methods use the nearest neighbor rule which requires storage for all training data samples. However, the amount of storage is tiny if the training data size is small.

System Flexibility

The proposed system model is open to feature extraction and correlation analysis. First, any kinds of flow statistical features can be applied in our system model. In this work, we extract unidirectional statistical features from full flows. The statistical features extracted from parts of flows can also be used to represent traffic flows in our system model. Second, any new correlation analysis method can be embedded into our system model. We introduce flow correlation analysis to discover correlation information in traffic flows to improve the robustness of classification. In this paper, a three-tuple heuristic-based method is applied discover flow correlations which are modeled by BoFs. We presented the comprehensive analysis from theoretical and empirical perspectives, which are based on the BoF model instead of the three-tuple method.

4. Conclusion

In this paper, we investigated the problem of traffic classification using very few supervised training samples. A novel nonparametric approach, TCC, was proposed to investigate correlation information in real traffic data and incorporate it into traffic classification. We presented a comprehensive analysis on the system framework and performance benefit from both theoretical and empirical perspectives, which strongly support the proposed approach. Three new classification methods, AVG-NN, MINNN, and MVT-NN, are proposed for illustration, which can incorporate correlation information into the class prediction for improving classification performance. A number of experiments carried out on two real-world traffic data sets show that the performance of traffic classification can be improved significantly and consistently under the critical circumstance of very few supervised training samples. 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] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, “BLINC: Multilevel Traffic Classification in the Dark,” Proc ACM SIGCOMM, vol. 35, pp. 229-240, Aug. 2005.
[2] 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.
[3] 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.
[4] 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.
[5] 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.
[6] A.W. Moore and D. Zuev, “Internet Traffic Classification Using Bayesian Analysis Techniques,” ACM SIGMETRICS Performance Evaluation Review (SIGMETRICS), vol. 33, pp. 50-60, June 2005