ISSN NO. 2320-5407

Journal homepage:http://www.journalijar.com
Journal DOI:10.21474/IJAR01

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

MACHINE LEARNING BASED NETWORK TRAFFIC CLASSIFICATION INCLUDING ZERO DAY TRAFFIC AS A CLASS.

UdayakumarBasavaraj Yalawar[1], and Kameswari K[2].

1. Department of information science and engineering, VTU University SJBIT, Bangalore.
2. Assistant professor**,** Department of information science and engineering, VTU University SJBIT, Bangalore.

Manuscript Info

Manuscript History:

Received: 18 March 2016
Final Accepted: 23 April 2016
Published Online: May 2016

Key words:
SVM, Semi-supervised Method, Zero-Day Traffic.

Corresponding Author

UdayakumarBasavaraj Yalawar.

Abstract

Support Vector Machines (SVM)represent one of the most promising Machine Learning (ML) tools that can be applied to the problem of traffic classification in IP networks. In the case of SVMs, there are still open questions that need to be addressed before they can be generally applied to traffic classifiers. Identifying and categorizing network traffic by application type is challenging because of the continued evolution of applications, especially of those with a desire to be undetectable. The diminished effectiveness of port-based identification and the overheads of deep packet inspection approaches motivate us to classify traffic by exploit To tackle this critical problem, we propose a novel traffic classification scheme which has the capability of identifying zero-day traffic as well as accurately classifying the traffic generated bypre-defined application classes.

Introduction:-

Classification of traffic can help identify different applications and protocols that exist in a network, which is a basic tool for network management [1]. For example, most of QOS control mechanisms has a traffic classification module in order to properly prioritize different applications across the limited bandwidth. In addition, to implement appropriate security policies, it is essential for any network manager to obtain a proper understanding of the applications and protocols in the network traffic. In the last decade, traffic classification has absorbed much attention in the industry. Identifying network traffic using port numbers was the norm in the recent past. This approach was successful because many traditional applications use port numbers assigned by or registered with the Internet Assigned Numbers Authority. The accuracy of this approach, however, has been seriously dented because of the evolution of applications that do not communicate on standardized ports. Many current generation P2P applications use ephemeral ports, and in some cases, use ports of well-known services such as Web and FTP to make them indistinguishable to the port-based classifier.

Techniques that rely on inspection of packet contents have been proposed to address the diminished effectiveness of port-based classification. These approaches attempt to determine whether or not a flow contains a characteristic signature of a known application. Studies show that these approaches work very well for today's Internet traffic, including P2P flows. In fact, commercial bandwidth management tools use application signature matching to enhance robustness of classification Nevertheless, packet inspection approaches pose several limitations. First, these techniques only identify traffic for which signatures are available. Maintaining an up-to-date list of signatures is a daunting task. Recent work on automatic detection of application signatures partially addresses this concern. Second, these techniques typically employ "deep" packet inspection because solutions such as capturing only a few payload bytes are insufficient or easily defeated (See Section 4.5 for empirical evidence of this.). Deep packet inspection places significant processing and/or memory constraints on the bandwidth management tool. On our

network, for example, we have observed that during peak hours, effective bandwidth is often limited by the ability of the deployed commercial packet shaping tool to process network flows. Finally, packet inspection techniques fail if the application uses encryption. Many Bit Torrent clients such as Azureus, µtorrent, and Bit Comet already allow use of encryption.

In our work, we develop and evaluate a technique that enables us to build a traffic classifier using flow statistics from both labeled and unlabeled flows. Specifically, we build the learner using both labeled and unlabeled flows and show how unlabeled flows can be leveraged to make the traffic classification problem manageable. This semi supervised approach to learning a network traffic classifier is one key contribution of this work. There are three main advantages to our proposed semi-supervised approach. First, fast and accurate classifiers can be obtained by training with a small number of labeled flows mixed with a large number of unlabeled flows. Second, our approach is robust and can handle both previously unseen applications and changed behavior of existing applications. Furthermore, our approach allows iterative development of the classifier by allowing network operators the flexibility of adding unlabeled flows to enhance the classifier's performance.

Material and Methods:-

This section presents a robust traffic classification scheme to deal with zero-day applications. . There are three important modules in the proposed framework:

- Unknown discovery.
- "Bag of flows" (BoF)-based traffic classification.
- System update.

Unknown discovery:-

We propose a two-step method of unknown discovery to extract zero-day traffic samples from a set of unlabeled network traffic crucial to the RTC scheme. The first step is the -means based identification of zero-day traffic clusters. The second step is zero-day sample extraction using random forest. Given the relabeled training sets and an unlabeledset, we roughly filter out some zero-day samples out by using a semi-supervised idea for the first step. The labeled and unlabeled samples are merged to feed the clustering algorithm, -means. The -means clustering aims to partition the traffic flows into clusters, to minimize the within-cluster sum of squares. The traditional -means algorithm uses an iterative refinement technique. Given an initial set of randomly selected centroids, the algorithm proceeds by alternating between the assignment step and the update step.

Bag-of-flows based traffic classification:-

For robust traffic classification, we further propose a new classification method that considers flow correlation in realworldnetwork traffic and classifies correlated flows togetherrather than in single flows. Given the Pre-labeled training sets and the zero-day sample set produced byte module of unknown discovery, we can build classifier for the -class classification. Is able to categorize zero-day traffic into a generic unknown class. Following our previous work [4], we incorporate flow correlation into the traffic classification process in order to significantly improve identification accuracy. Flow correlation can be discovered byte 3-tuple heuristic. That is, in a short period of time, the flows sharing the same destination IP, destination port, and transport protocol are generated by the same application/protocol.

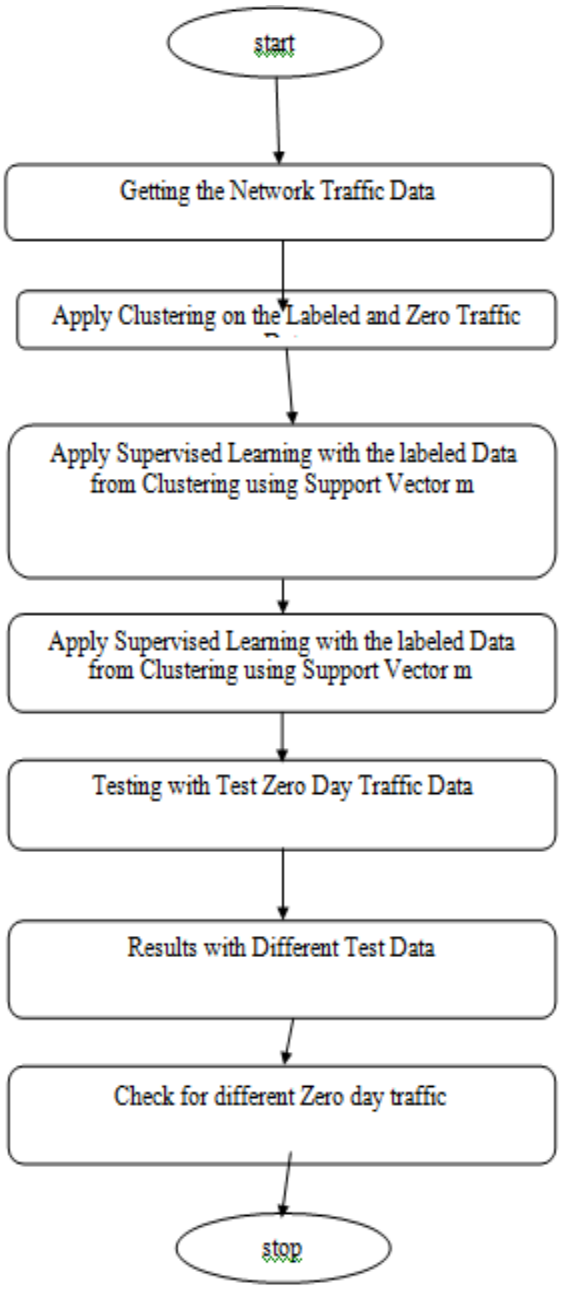


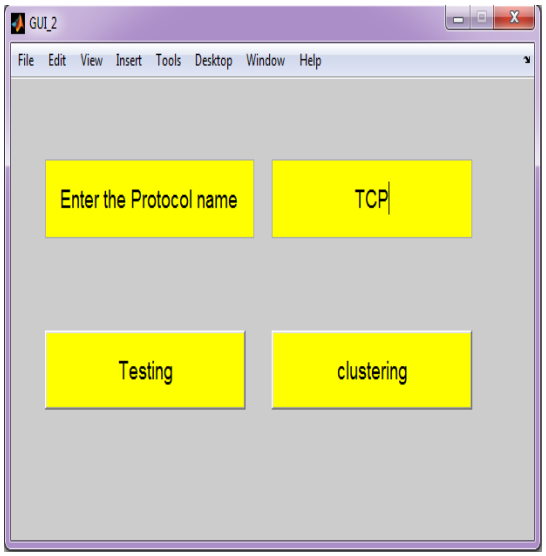
Figure 1: Flow diagram of robust traffic classification.

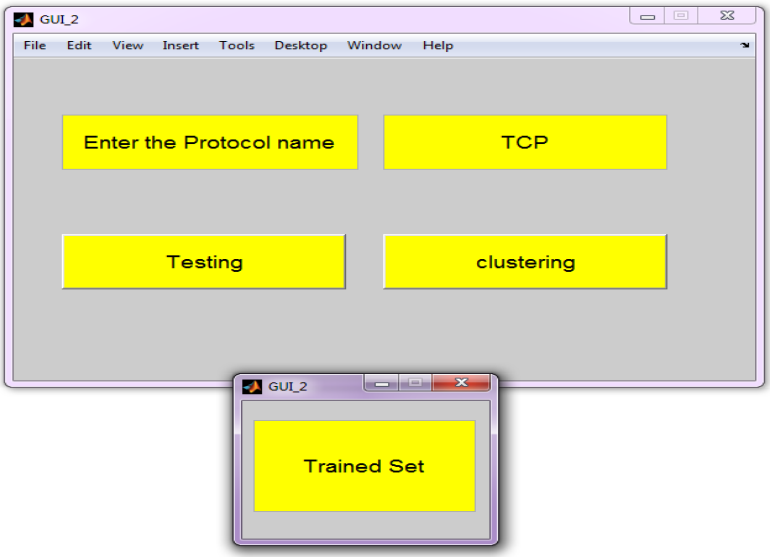
System Update:-

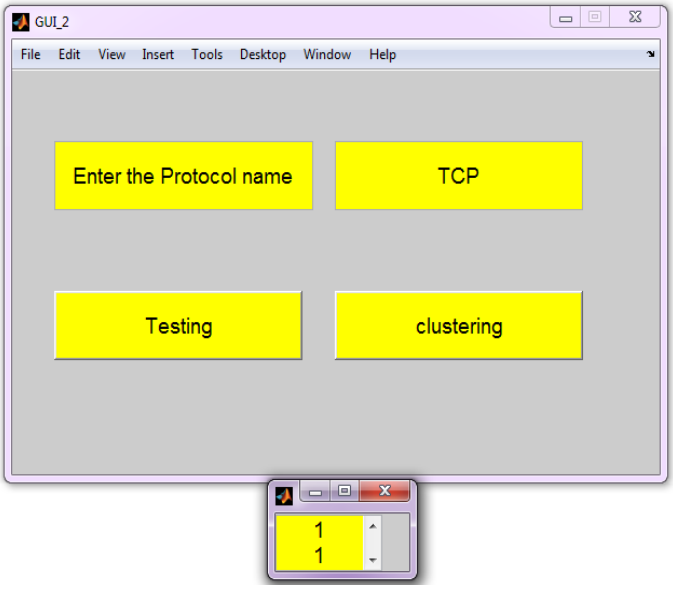
The figure 1 shows the robust traffic classification with unknown discovery and BoF-based traffic classification, the proposed scheme has identified zero-day traffic when performing traffic classification. The module of system updates proposed to achieve fine-grained classification of zero-day traffic. The purpose is to learn new classes in identified zero-day traffic and to complement the system's knowledge. The capability of learning new classes makes the proposed scheme different to the conventional traffic classification method. Given a set of zero-day traffic, which is the outcome of BoF-based traffic classification, we perform –means clustering to obtain the clusters. For each cluster, we randomly select several sample flows (e.g., three) for manual inspection. To guarantee high purity of new training sets, the consensus strategy is adopted to make a prediction. If all the selected flows indicate a new application/protocol, we create a new class and use the flows in the cluster as its training data. For a new class that has been created during the system update, the flows in the cluster will be added to the training setoff that class.

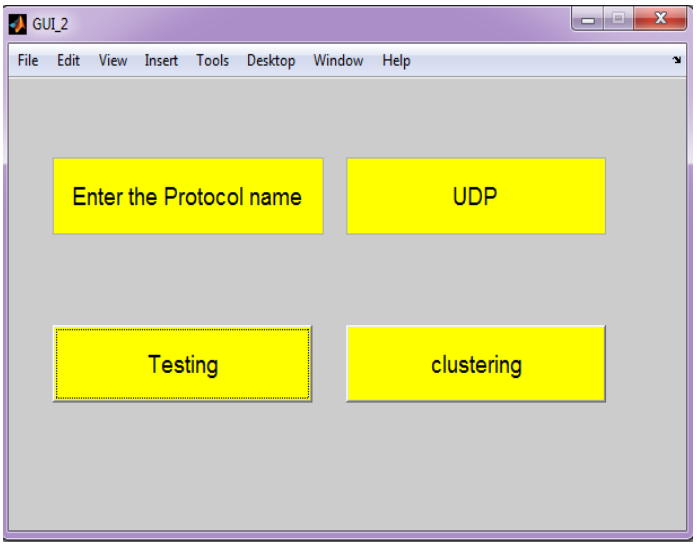
Result and Discussion:-

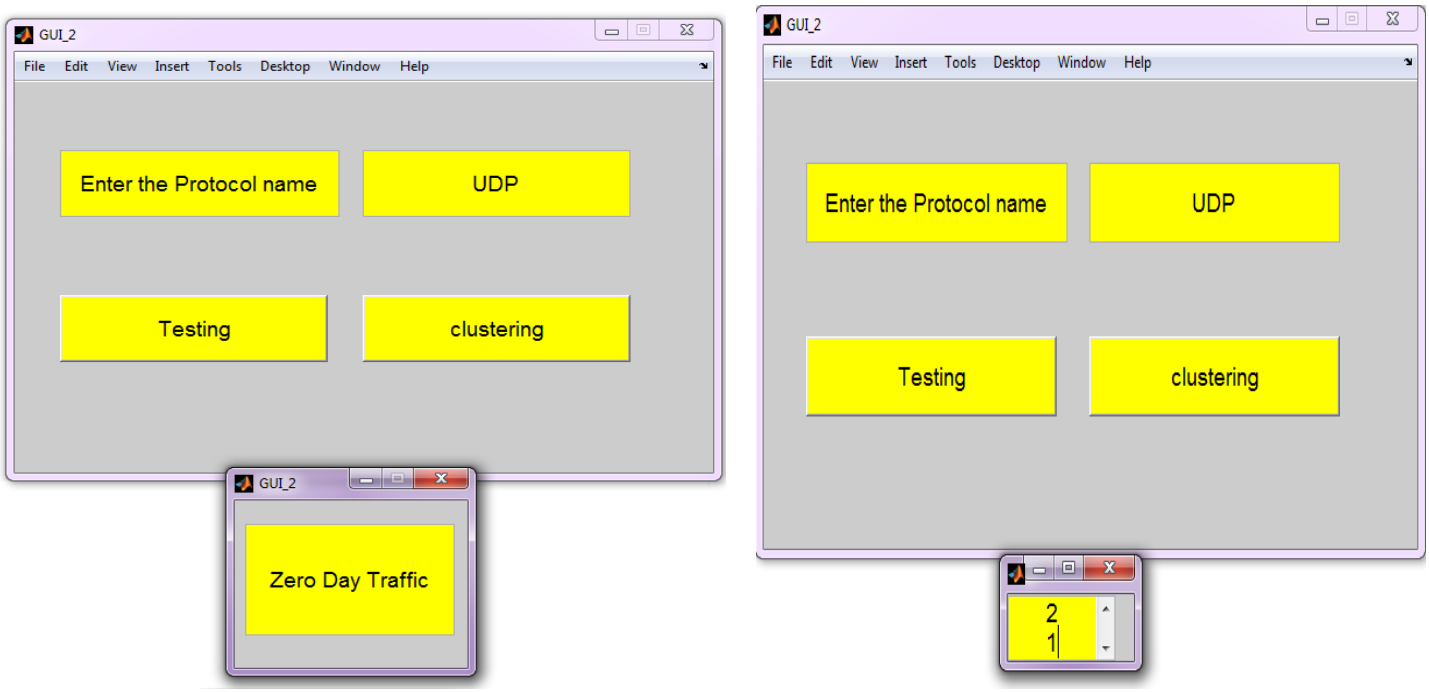
Snapshots:-

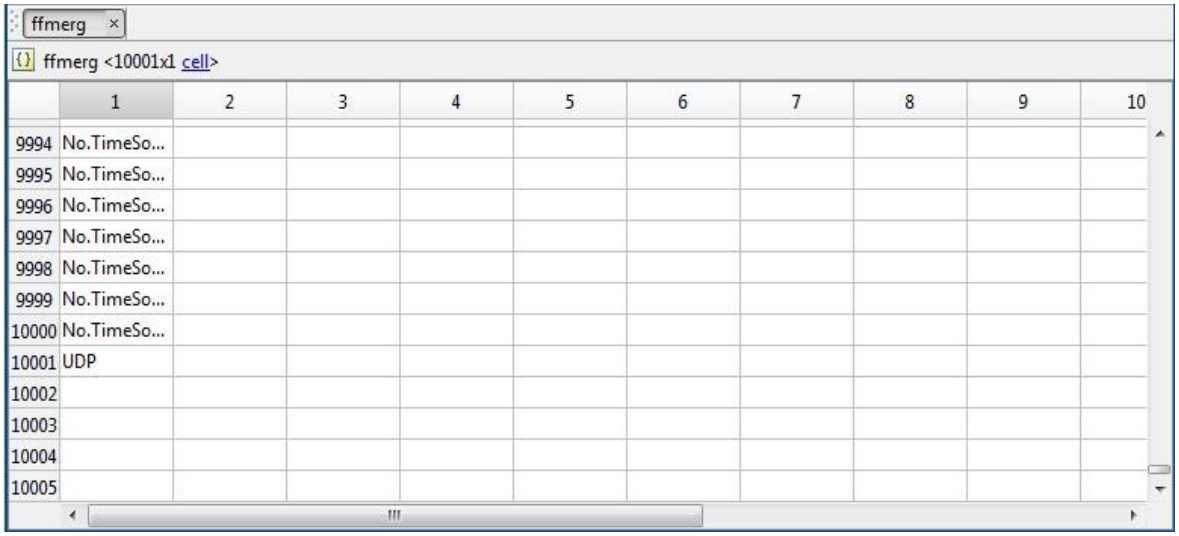












The above snapshot shows the work flow of my project. At first it is tested for the trained data, if it is available in the database it shows that it is a trained data otherwise if the searched text is not available in the database, then it shows that as the zero day traffic. It creates the new cluster class for the zero day traffic. If you again entering same data at the second time it shows that it is trained data because it is stored in the database.

References:-

1. **Moore and D. Zuev**, "Internet traffic classification using Bayesian analyses techniques," Perform. Eval. Rev., vol. 33, no. 1, pp. 50–60, 2013.
2. **T. Auld, A. Moore, and S. Gull**, "Bayesian neural networks for Internet traffic classification," IEEE Trans. Neural Newt., vol. 18, no. 1, pp. 223–239, Jan. 2012.
3. **Este, F. Gringoli, and L. Salgarelli,** "Support vector machines for TCP traffic classification," Compute. Newt. vol. 53, no. 14, pp. 2476–2490, 2012.
4. **J. Erman, A. Mahanti, M. Arlitt, I. Cohen, and C. Williamson,** "Offline/ real time traffic classification using semi-supervised learning," Perform. Eval. vol. 64, no. 9, pp. 1194–1213, 2011.
5. **Bernaille, L., Teixeira, R., Salamatian, K.:** Early application identification. In: Proc. of ACM International Conference on Emerging Networking Experiments and Technologies (CoNEXT) 2008, Lisboa, Portugal (2006)
6. **Nguyen, T.T.T., Armitage, G.:** A survey of techniques for internet traffic classification using machine learning. IEEE Communications Surveys & Tutorials 10 (2006)
7. **Wright, C., Monrose, F., Masson, G.:** HMM profiles for network traffic classification (extended abstract). In: Proc. O Workshop on Visualization and Data Mining for Computer Security (VizSEC/DMSEC), Fairfax, VA, USA (2004).