Volume 3, Issue 12, December 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey on Recent Traffic Classification Techniques Using Machine Learning Methods

M. Tamilkili

Dept of CSE& Karunya University

India

Abstract-Recent research tends to apply machine learning techniques to classify network traffic using flow statistical features and IP packet payload. This survey paper looks into the use of Machine Learning (ML) algorithms for traffic classification in the period 2009 to 2013. It provides motivation for the application of ML techniques to network traffic classification, and reviews significant work in this area. The recent techniques focus on statistical features of network packets to provide security. A review of ML techniques has been done

Keywords— Machine learning, Traffic classification, feature extraction, signature generation, cluster aggregation.

I. INTRODUCTION

Network Traffic Classification is a technique used to classify network traffic based on features passively observed in the network according to specific classification goals. It has a wide range of applications in network security and management, such as quality of service (QOS) control, lawful interception (LI) and intrusion detection. In general traffic classification based on well known port-numbers are not effective as observed from previous studies [1].

The network security challenges in recent years are to impose privacy regulations constraining the ability of third parties to inspect payload in any situation even if it is lawful to do so and the packet information needs to be updated regularly. In most of the cases, the traffic is classified based on the source of application, payload content and some flow statistical features. The traditional well known port based methods are ineffective therefore the statistical features based methods can be used. Examples where traditional classification techniques are increasingly losing their effectiveness are in cases where explicit ‘‘anti-classifier” obfuscation techniques are employed by the user applications. The packet inspection for traffic classification [1] which is port based traffic classification, and payload based traffic classification have their own limitations. The rest of the paper is organized as follows Section II outlines the basic requirements of traffic classification. Section III outlines the Machine learning approaches. Section IV outlines the review of ML-Based traffic classification techniques. Section V concludes the paper.

II. BASIC REQUIREMENTS

Flows: These are represented by 5 tuples, {source_ip, destination_ip, source_port, destination_port, protocol} and are classified as unidirectional, bidirectional and full flows. Flow statistical Features: Features are properties or characteristics of flows calculated over multiple packets. Some of the properties are maximum or minimum packet length in each directions, minimum or maximum packet arrival time, minimum or maximum number of bytes transferred in forward and backward directions.

Performance metrics: Terms used are True Positive (TP), True Negative (TN), and False Positive (FP), and False Negative (FN), accuracy (flow and byte), precision, recall, classification error, F-measure [10].Common metrics to find classifier’s accuracy are as follows [2]:

- True Positive: The number of features belonging to class Y classified as class Y.
- False Positive: The number of features of other classes incorrectly classified as belonging to class Y.
- True Negative: The number of features of other classes correctly classified as not belonging to class Y.
- False Negative: The number of features of other class Y incorrectly classified as not belonging to class Y.
- Accuracy: It is defined as the fraction of the number of correctly classified flows or bytes over the size of the data.
- Classification Error: It is defined as the raw count of flows which were correctly classified divided by the total number of flows. This metric is used to find classifier’s accuracy for the whole system.
- F-measure: *2 * Precision * recall /(precision +recall).*This metric is used to rank and compare the per-application performance of ML algorithms.

ML uses the following metrics:

- Recall: The numbers of features of class Y correctly classified as belonging to class Y.
- Precision: The numbers of those instances that truly have class Y, among all those classified as class Y.

Confusion matrix is used to show the relationship between performance metric terms

Table I Confusion Matrix

	Predicate classes		
Actual classes		Yes	No
	Yes	TP	FN
	No	FP	TN

III. MACHINE LEARNING APPROACHES

Machine learning can be defined as "*the study of making machines acquire new knowledge, new skills, and reorganise existing knowledge*" [1]. ML approaches for traffic classification are supervised, unsupervised and semi-supervised. The input for ML is a dataset containing network traces and the output for ML is a classification of the network trace.

The supervised learning approaches are based on pre-defined or pre-labelled knowledge. It provides the support for classifying new instances of traffic into predefined classes which are trained by previously gained knowledge structure.The learning machine is provided with a collection of sample instances, pre-classified into classes. Output of the learning process is a classification model that is constructed by examining and generalizing from the provided instances. There are two major phases (steps) in supervised learning: (i) Training: The learning phase that examines the provided data (called the training dataset) and constructs (builds) a classification model. (ii) Testing (also known as classifying): The model that has been built in the training phase is used to classify new unseen instances.The Unsupervised learning approaches provide a learning technique which doesn't require any pre-determined or defined knowledge (unlabeled) instead they discover natural groups (clusters) in the data using internalized heuristics. It focuses on finding patterns in the input data. It clusters instances with similar properties (defined by a specific distance measuring approach, such as Euclidean space) into groups. They may be hierarchical, where there is a division of instances into groups at the top level, and then each of these groups is refined further down to the level of individual instances.

The semi-supervised or hybrid approaches enable us to classify traffic using flow statistics from both labeled and unlabeled flows. Advantage of this approach is that it provides accurate and fast classifiers by considering both labeled and unlabeled flows. It also handles unseen applications and changed application behavior [3].

IV. REVIEW OF TRAFFIC CLASSIFICATION TECHNIQUES USING ML

Some of the recent traffic classification techniques using ML are discussed below

A. Unsupervised traffic classification:

In [4] a novel unsupervised approach is proposed which takes full data trace collected over network as an input. In the training phase, it extracts features from the dataset collected and gives it as an input for clustering purposes. The most effective and simplest K-means clustering algorithm is used. It produces clusters as an output. The clusters which are produced in the above step are given as an input to build a classifier model called cluster aggregation. The output of this model is few aggregated clusters, which have higher similarity within themselves. In the testing phase, a few aggregated clusters are used to check whether the classifier achieves its intended performance. The novel unsupervised approach classifies traffic based on flows and payload characteristics. It clusters traffic traces collected based on flow statistical features using K-Means and provides high purity clusters when K is large. Features considered are Packets (unidirectional), Bytes (unidirectional), Packet size (min, mean, max, stdDev) and Inter-Packet time (min, mean, max, stdDev), totally 20 flow statistical features are taken. Of the large number of clusters formed those having similar characteristics are aggregated based on payload content. For cluster aggregation, bag of words model (BoW) is used to find code words which represent applications, and also uses tf-idf (term frequency-inverse document frequency) for cluster representation.

The cluster provided by previous step is optimized using LSA (latent semantic analysis). Finally, the clusters produced by LSA are merged in the concept space. Some stop criteria like similarity threshold are used to stop aggregation. The similarity based aggregated clusters is given as an input for testing, which evaluates the flow based classifier performance in terms of accuracy. Thirteen applications which are xmpp, bitorrent, smartfox, smtp, imap, msn, yahoomsg, dns, ftp, ssh, pop3, rtmp and http have been considered in [4].

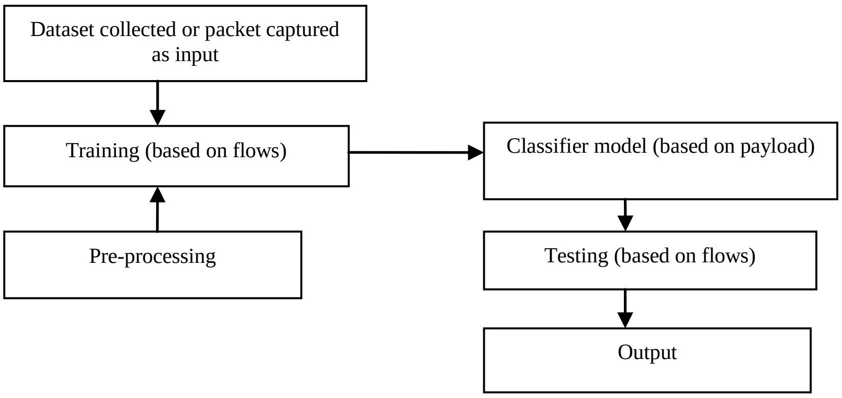
Dataset collected or packet captured as input
Training (based on flows)
Classifier model (based on payload)
Pre-processing
Testing (based on flows)
Output

Fig. 1 Novel Unsupervised Approach

The novel unsupervised approach proposed in [4] outperforms the competing supervised classification methods (C4.5, K-NN, SVM, Naive Bayes, Neural Network and Bayes Net) which used 1000 training samples, without any supervised data. The accuracy of the proposed aggregation method is 89 %.

B. *Support Vector Machines (SVM) based traffic classifier*

In [5], an approach to traffic classification is proposed which is based on SVM, that approach is used to solve the mutli-class problem that arises in SVM, classify network traffic and apply optimization algorithm to make classifier perform properly even with a small training set for hundreds of samples.

SVM is one of the most promising ML tools, a binary classifier suitable for solving high dimensional feature space and small training set size problems. The proposed approach uses flow representation that describes the statistical characteristics of application protocol through monitoring node whose duty is to assign flows to the concerned application classes it was trained with or with unknown class. It considers bidirectional flows only, which follows proper TCP three way handshakes and proper termination. After packets are captured, each flow is mapped to feature values which are based on packet's length to determine which application the packet belongs to. The main feature of this proposed classifier is based on packet size as it is being captured on the application layer.

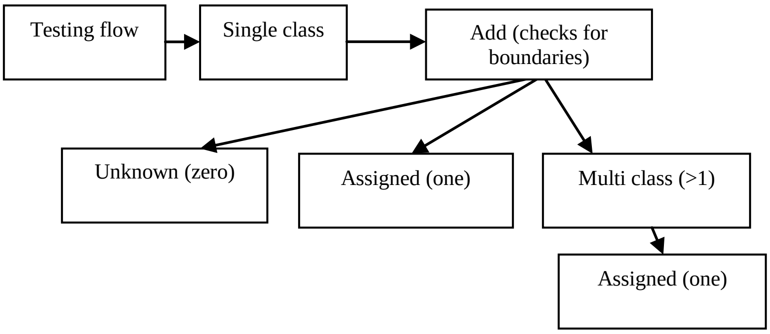

Fig. 2 SVM based Traffic Classifier

In the training phase, the dataset is classified correctly into appropriate application classes according to the pre-labelled or pre-classified data. Payload based pattern matching technique is used when the dataset contains application payload data. When the payload based pattern matching technique is not possible, the use of server transport layer port to group flows of traffic is a possibility. The proposed classifier contains two stages, single class stage and multi class stage. In single-class model, each application protocol is assigned to a separate surface in an n-dimensional space. This surface should include several elements of a class that belongs to a particular application protocol. A single-class model is based on surface boundaries, a value of 1 means that the element is inside the surface of the ith class, a value of 0 means it is outside the surface of the ith class. Multi-class stage is used to limit the complexity of optimization of M binary SVM for each one-against-all case. For that some relation expression are set in [5]. The number 0 means Unknown, 1 means Assigned, >1 means Multi-class stage.

The proposed classifier uses three different datasets named CAIDA, LBNL and UNIBS. In all three datasets, the classifier accuracy is very good with True Positive of 90%.

C. Traffic Classification using Correlation information

In general, the Nearest-Neighbor (NN) based approaches provide better classifier performance, but this is not possible when the size of training data is small. In [6], a novel non-parametric approach is proposed that combines correlation of flows to improve classifier performance for small training samples.

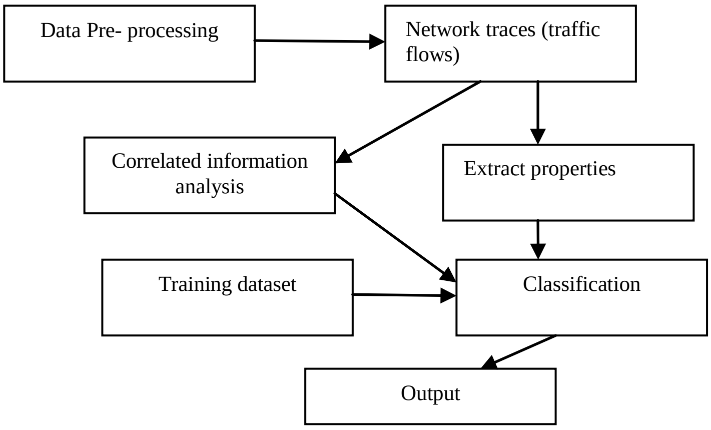

Fig. 3 A Novel Non-Parametric Approach

In the pre-processing step, the ip packets are collected from the network and flows of traffic are constructed. Flow is composed of 5-tuples: {src_ip, dst_ip, src_port, dst_port, proto} and with this flow, features are extracted for classification purposes. The correlation information is analysed in the traffic flows and it is given as input to the classifier. Finally the robust classifier classifies traffic flows into application based classes based on flow statistical features and correlated information of traffic flows. Bag of Flows (BoF) are used to model correlation information in traffic flows. The conventional classification algorithms like SVM, NN, and Neural Network severely affect the classifier performance when only few training samples are considered. The NN provides 60% classifier accuracy when considering only a few training samples on two datasets (wide and isp), which is very low.

Conventional supervised methods classify traffic based on individual and independent instances, but the proposed approach takes the correlation information among traffic flows and incorporates them into classification process. The feature extraction or selection using Correlation Based Filter (CFS) is done to optimize learning process and computational complexity. Unidirectional statistical features are extracted. Three classification methods are used AVG-NN, MIN-NN and MVT-NN. The classifier performance is measured by two performance metrics, overall accuracy and F-measure.

The proposed approach requires very little prior knowledge and doesn’t involve any learning process like conventional approaches (SVM, Neural Network), doesn’t have any mapping problem and can handle encrypted applications based on flow features. AVG-NN provides better performance by considering per-experiment, overall and per-class performances. It is used for automatic recognition of unknown applications.

D. KISS

The goal of KISS [7] is to discover application protocol header format. The approach KISS is used because of the rise of UDP traffic and it proposes a classification framework which is based on statistical properties of payload. Statistical signatures are derived by Chi-square like tests to extract the protocol format and not extract synchronization and semantic rules. The signature derived will be the input for decision process. The decision process is based on either Euclidean distance or SVM and is tested based on ground truth. Testbed is created for artificial traffic to know the ground truth. In pre-processing step, features are extracted and decision process occurs which assigns an observed sample to which application or feature belongs.

KISS learning process takes the input as traffic which is given to chunker. The chunkers are used to derive KISS signatures and it is sampled by a sampler to create training sets. The training set is given to a learning process and it produces the KISS model. Two Machine learning processes are used. Euclidean distance is used to find a point that doesn’t fall into any group or sphere that point will be considered as an unknown application. SVM [5] classifies traffic based on prior knowledge, the traffic is labelled.

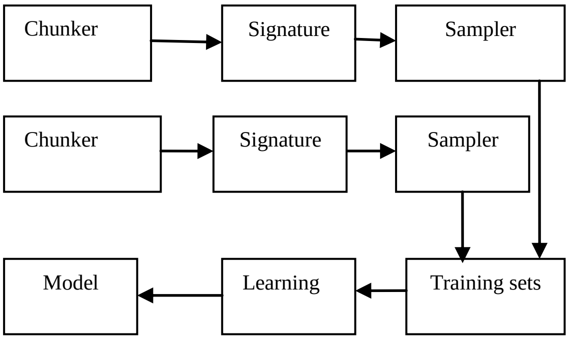

Fig. 4 KISS Learning Process

If there is a new region or traffic that doesn’t belong to any pre-defined data, it should be identified by SVM. For that purpose, it uses two types of signatures (i) one that represents traffic generated by the application to classify (ii) another that represents the remaining application traffic. SVM methods are robust for all dataset sizes, limited computation and memory requirements, it also achieves higher accuracy. Classification entities used are flow and endpoint.

Data traces used are Real-Trace-I, Real-Trace-II, P2P trace and Skype. The common Performance metrics used are TN, TP, FP, FN and it is represented by a confusion matrix. The average TP percentage is 99.6 % and less than 1 % FP percentage. KISS is therefore very robust.

E. Constrained Clustering

In general, Clustering provides good classification performance. In [8] a novel semi-supervised method is proposed based on flow statistics and IP packet content with the use of constrained clustering algorithm. In this approach, if the flows belong to same port and same host address then it is inferred that they are based on the same application protocol and this information is used as pair-wise multi-constraint and is given as an input to K-Means Clustering algorithm. The flows considered here are {dst_ip, dst_port, proto}. The flows are paired according to the constraints, multi-link constraint and cannot link constraint.

The K-Means algorithm uses Euclidean distance as a measure to find similar traffic flows. To include constraints into K-Means, two methods are used, the first method directly satisfies constraints and the other method learns a distance metric. For this purpose, linear time constrained vector quantization error algorithm (LCVQE), MPCK-Means algorithm, COP-Means algorithms are used. These algorithms will incorporate the constraints into the unsupervised k-means algorithm. However it is not possible to directly include constraints into algorithms, so constraints are compressed using transitive properties. It is done in the pre-processing stage itself. Performance metric used is overall accuracy. This semi-supervised method using constrained clustering algorithm provides high purity clusters and improves overall accuracy.

F. ACAS

In general, the traffic flows are classified based on flow statistics, port number, payload, signature format, application protocol format. But in [9] a novel approach is proposed which derives the signature from unknown traffic automatically. Traffic flows belonging to one application are clustered based on flow-statistics and signatures are constructed based on payload content. For automatic application identification, two processes are done. First, online classification captures traffic flows which is given as input to signature classifiers, if the signature is already there, it finds its application if not found it goes for offline training. In offline training, unknown flows are clustered, signatures are constructed and new signatures are inserted into the signature classifier which is a part of the online classification. Two supervised algorithms Naive Bayes and C4.5 decision tree are used for signature construction.

Performance measures used are overall accuracy, precision and recall. The precision rate is 80 % except for PoP3, the recall rate is 95% except for http and ftp. Overall accuracy rates of Naive Bayes and C4.5 decision tree are 93.96 % and 92.72% respectively.

Table II Summary of Reviewed Papers

Work	ML algorithms	Features	Data Traces	Traffic considered
Jun Zhang et al. [4]	K-Means	Packets, Bytes, Packet size, Inter-Packet time	Full-packet trace collected Melbourne, Australia	xmpp, bitorrent, smartfox, smtp, imap, msn, yahoomsg, dns, ftp, ssh, pop3, rtmp and http.
Alice Este et al. [5]	SVM (support vector machine)	Packet size, Bytes, Packets	CAIDA ,UNIBS, LBNL	http, ftp, bitttorrent, edonkey, smtp, msn, unknown, pop3, edonkey
Jun zhang et al. [6]	NN (nearest Neighbor)	Packets, Bytes, Packet size ,Inter packet time	Wide ,isp ,sigcomm, lbml, Keio	dns, mail,p2p,ftp, chat, http, imap, msn, ssl, xmpp, pop3, smtp, ssh
Alessandro Finamore et al. [7]	Euclidean distance , SVM	Packets, Bytes, Endpoints ,Flows	p2ptrace, skype, realtrace-I and II	Skype, p2p,dns, backg, emule, joost, ppLive Sopcast
Y.Wang et al. [8]	K-Means (COP, LCQVE, MPCK)	Duration, Packet, Bytes, Packet size, Inter packet time	wide,isp, keio, sigcomm, lbnl	http,pop3smtp,bittorrent, edonkey,guntella,imap, msn,ssh, others, unknown, ssl
Y. Wang et al. [9]	Naive Bayes,K-Means,C4.5	Packets, Bytes	full packet trace which is Collected at an educational site on the Internet.	Bittorrent,edonkey ftp http,pop3,smtp,dns

V. CONCLUSION

This paper surveys some of the recent work in the field of Machine learning based IP traffic classification in the period 2009 to 2013. The ML algorithms for new applications like Skype, Video streaming, Voice over IP and peer to peer file sharing are also reviewed to a certain extent.

REFERENCES

[1] T.T. Nguyen, G. Armitage, *A survey of techniques for Internet traffic classification using machine learning*, IEEE Commun. Surveys Tutor. 10 (4) (2008) 56–76.

[2] https://www.cs.waikato.ac.nz/ml/weka/**book**.html

[3] J. Erman, A. Mahanti, M. Arlitt, I. Cohen, C. Williamson, *Offline/realtime traffic classification using semi-supervised learning*, Performance Evaluation 64 (9–12) (2007) 1194–1213.

[4] Jun Zhang, Yang Xiang, Wanlei Zhou, Yu Wang, *Unsupervised traffic classification using flow statistical properties and IP packet payload*, Journal of Computer and System Sciences 79 (2013) 573–585.

[5] A. Este, F. Gringoli, L. Salgarelli, *Support vector machines for tcp traffic classification*, Computer Networks 53 (14) (2009) 2476–2490.

[6] J. Zhang, Y. Xiang, Y. Wang, W. Zhou, Y. Xiang, Y. Guan, *Network traffic classification using correlation information*, IEEE Trans. Parallel Distrib. Syst. (2012)1–15.

[7] A. Finamore, M. Mellia, M. Meo, D. Rossi, *KISS: Stochastic packet inspection classifier for UDP traffic*, IEEE/ACM Trans. Netw. 18 (5) (2010) 1505–1515.

[8] Y. Wang, Y. Xiang, J. Zhang, S.-Z. Yu, *A novel semi-supervised approach for network traffic clustering*, in: International Conference on Network and System Security, Milan, Italy, September 2011.

[9] Y. Wang, Y. Xiang, S.-Z. Yu, *An automatic application signature construction system for unknown traffic*, Concurrency Computat. Pract. Exper. 22 (2010) 1927–194.

[10] https://code.google.com/p/netmate-flowcalc/wiki/Features