A Semi-Stack Approach for Accurate Network Traffic Classification Using Multi-View Stacking

Adil Fahada,*

*Department of Computer Science, College of Computer Science and IT, Al-Baha University, Al-Baha, Saudi Arabia

Email: afalharthi@bu.edu.sa

Abstract. The network traffic classification is used as a basis of network management related works, from security monitoring and Intrusion Detection Systems (IDS) to Quality of Service (QoS). Many works have been carried out on this theme that proposed various approaches. Most of the proposed approaches utilize predefined class label provided by an expert to perform the network traffic classification. Nevertheless, the difficulty to acquire consistent, adequate and up-to-date ground truth for classifying network flows effectively is still a serious problem. This work addresses such an issue via multi-view stacking to fuse the metadata from heterogeneous types of semi-supervised classifiers for accurate classification thru the following steps. Firstly, the original data is represented as multi-view using dimensionality reduction methods, with the aim to have strong discrimination capability. Secondly, integrate different semi-supervised learning algorithms from an ensemble learning perspective, to deliver a better stability and quality classification output. Finally, propose N-fold cross validation on metadata for training the meta-classifier, to produce the final class decision and predict the unlabelled traffic data. Experiment results on four existing network traffic datasets provide the best average accuracy mark is above 96% on all datasets and the best stability performance is up to 99.50%.

1. Introduction

Network classification is the task of classifying a set of network flows into classes based on two categories such as the used protocols (e.g. UDP, TCP) or applications (e.g. Web, Email, etc.). Traffic classification methods have been extensively developed lately as essential tools [1-5, 7] to support network management activities. In turn, the tools assist the Internet Service Provider (ISPs) from providing better QoS, to security control. Promising classification methods using statistical characteristics of Internet Protocol (IP) traffic flows and machine learning algorithms have been proposed and developed by the researchers and the networking industries over the past decade to address the limitations of the traditional network classification methods, including port-based methods and deep packet inspection methods [2, 6, 10, 12, 17]. However, accuracy and efficiency problems are still remain unsolved. The reasons are the proposed methods used force assignment where the classification of the unlabelled traffic flows should be compulsory assigned to fixed predefined classes of protocols or applications, and disregard to learn the appearance of new traffic patterns and emerging applications (e.g. zero-day attacks). Thus, researchers need an automatic procedure for accurate and efficient labelling and building up-to-date ground truth. Such an automatic procedure is an essential to train and test various network classification approaches based on the different machine learning algorithms, and also to replace the manual labelling procedure which is tedious and costly.
To do so, this paper proposes a novel semi-supervised approach, namely Semi-Stack. The proposed approach fuses the metadata from heterogeneous ensemble semi-supervised classifiers as a training set along with N-fold cross validation approach to train the meta-classifier and obtaining the final class label. Experiments using four existing network traffic datasets [1, 9, 15, 16] verify the significant improvements of the proposed approach’s classification capability.

The paper is organized as follows: Section 2 discussed the related work in of network traffic classification. Section 3 introduces our proposed Semi-Stack approach. Section 4 presents the experimental settings, and discussion of the results. Finally, the conclusion and outlines of future work are presented in Section 5.

2. Related Work
Machine Learning-based classification techniques are categorized into supervised and unsupervised learning.

- **Supervised learning**: traffic flow classes are given before the learning [2, 3, 8, 11]. The process of learning technique consists of two main phases: training phase and testing phase. The former phase is about learn and analyse the given training dataset, then generate a predictive classification model. Through the training process, the predictive model builds a capability to generate a prediction for any new instances by examining the characteristic values of unknown flows as based on prior training data. In the testing phase, the obtained model from the training phase is used to classify new unseen traffic flow instances. Supervised learning creates knowledge structures that support the task of classifying new instances into pre-defined classes [13]. Thus, supervised learning concentrates on modelling the relationships of input and output to produce knowledge presented as a decision tree, classification rules, a flowchart, etc. and will be used to classify a new instance.

- **Unsupervised learning**: is used to make inferences from datasets without labelled traffic flows. Cluster analysis is the most usual unsupervised learning technique used for investigative data analysis to discover unseen patterns. McGregor et al. in [8] and Rotsos et al. [14] are ones of the earliest works to employ unsupervised machine learning technique to model the underlying structure and distribution of network traffic flows. The authors use Expectation Maximization (EM) method and consider Transport Layer attributes as traffic flow features. Zander et al. [19], extend the works by introducing different EM algorithm named as Auto Class. The authors investigate the best set of attributes to use. Then, using Bayesian clustering technique combined with standard EM algorithm will guarantee the convergence of the combined method to a local maximum. Next, the Auto Class performs the EM searches iteratively, starting from pseudo-random points in parameter space to search the global maximum. Therefore, it works considerably superior to the original EM method. The proposed idea in [3] has shown that the use of Transport Layer attributes only, enables cluster analysis to classify Internet traffic flows. Furthermore, the use DBSCAN and K-Mean clustering methods to evaluate the predicting performance has been presented by Erman et al. [5]. The authors verified that both DBSCAN and K-Mean provide better accuracy and faster processing time compared to the Auto Class clustering method presented in [19]. Overall, the unsupervised learning techniques are not as good as supervised techniques. Therefore, this work attempts to explore the benefits of both learning techniques to improve essentially stability and accuracy of traffic flow classifiers.

3. The Proposed Semi-Stack Approach
The proposed Semi-Stack approach considers the advantages combination of semi-supervised and meta-learning algorithms as a basis in its development.

3.1. The proposed Semi-Stack approach overview
A huge labelled traffic dataset is necessary for developing accurate network traffic classifiers based on machine learning (ML) methods. However, current ground truth creation for network classifiers requires a fantastic cost and workload. This work proposes a novel semi-supervised (Semi-Stack) approach for proper training and evaluation of network classifiers that reduces tiresome manual processes of labelling and at the same time reaches the goal at realistic cost. Figure 1 depicts the grand architecture of the proposed Semi-Stack approach. The Semi-Stack approach is structured by three
layers: (1) the projection of the data to achieve diversity among traffic data, (2) ensemble Semi-Stack to increase the class prediction accuracy and the learning process efficiency thru unlabelled data exploitation, and (3) meta-learning to robust the automatic labelling process by generating the final prediction for the unlabelled traffic data.

3.2. Multi-view layer
This layer increases the performance of the labelling process by giving multiple representations of the initial data.
This work assumes that not all of the interesting patterns can be take out from a single view. On the other hand, a representation may have different data structures.

Figure 1. General Architecture of the Semi-Stack approach.

This layer process the network traffic data to reduce dimensionality and acts as representation methods. Three common famous methods are selected [20], they are:
- Random Projections (RP): projects the data into lower dimensional spaces using random projection matrices.
- Isomap: extends the multidimensional scaling metric and incorporates a weighted graph to estimates the geodesic manifold of intrinsic geometry of the nonlinear data.
- Kernel Principle Component Analysis (KPCA): generalizes the PCA. The PCA is a key statistical technique for extracting features and data modelling.

3.3. Semi-ensemble Layer
The proposed Semi-Ensemble Layer learns from different representations of the traffic flows generated previously by the former step (the multi-view layer), and then generates a pool of the predication values of the unlabelled data, referred to as meta-level data. Figure 2 shows a visualization of Semi-Ensemble Layer process.
3.4. Meta-learning layer

In this work, the author uses a homogeneous base learner rather than a single base learning algorithm to decrease variance, bias and improve predictions in classifier model. A widely adopted approach to this objective is to apply meta-learning layer to benefit from its repetitive process and to compute descriptive models where the learning instances are not represented by original descriptive features, e.g. port number, IP address features, but by the output features (the meta-level data) generated previously from semi-ensemble layer. In particular, meta-learning layer has two steps. The first step is trained using the predictions of the semi-supervised classifiers (meta-level data) and the true class of the original traffic data. The second step generates the final prediction results by testing the meta-classifier model. Figure 3 shows the process of the meta-learning layer. Firstly, split the meta-level data into two sets: training set and testing set. The former set is used as initial labelled data to train the meta-classifier. The author uses the Decision Tree [3], an inherently interpretable algorithm. Once the training process is accomplished, the latter set, which comprises part of labelled data with unlabelled one, is applied on the Meta classifier model to obtain the final predictions. 10-fold cross validation is used to improve the quality of meta-classifier model by generating more meta-instances (traffic flows).

3.5. Experimental procedure

Figure 4 shows the algorithm of experimental procedure of this work.
4. Experimental Evaluation

4.1. Traffic datasets

Four publicly available traffic data sets are used for experimental evaluation of the proposed Semi-Stack approach, including Internet Traffic Data (ITD, defined in [11]); DARPA datasets [15]; wide dataset [1]; and ISP dataset [16].

This work focuses on TCP traffic flows. The advantage is it has the clear start-end information. Table 1 recapitulates attributes of each dataset, i.e.: the number of classes, the number of features, and the data allocation for instances in training and testing phases.

| Dataset | # of Classes | # of Features | Instances# (Training) | Instances# (Testing) |
|---------|--------------|---------------|-----------------------|----------------------|
| ITD     | 12           | 149           | 15750                 | 5250                 |
| DARPA   | 2            | 41            | 7500                  | 2500                 |
| wide    | 6            | 20            | 10030                 | 3100                 |
| ISP     | 12           | 149           | 15750                 | 5250                 |

4.2. Experimental set up

The experiment uses the 10-fold cross-validation strategy to obtain robust results. The strategy is repeated 5 times to avoid bias toward the data ordering, which may degrade the performance of the Semi-Stack approach and the four comparison methods. All experiments were conducted on a computer with processor of 4-duo and Core(i7), 3.30 GHz Intel CPU, 8-GB RAM and 64-bit Windows system. The Semi-Stack approach is implemented using Java Platform and combined with Weka software [18].

4.3. Experimental results

Experiments are conducted to measure the performance criteria: overall accuracy, runtime, and stability.
1) Accuracy Performance: Table 2 indicates that all four baseline (comparison) methods achieve significant overall accuracy marks on all datasets (normal or attack). Specifically, Semi-Stack approach reaches the best average accuracy marks, by performing accuracy above 96% on all datasets. The rational is because the Semi-Stack approach is able to associate results of multiple labelling processes to gain accurate class label and remove unconfirmed results.

**Table 2. Overall accuracy for each model and each dataset.**

| Dataset | Semi-Stack | SemTra [20] | PGM | BGCM | ORTSC |
|---------|------------|-------------|-----|------|-------|
|         | Mean St.D  | Mean St.D   | Mean St.D | Mean St.D | Mean St.D |
| ITD     | 96.78 2.0  | 95.3 0.88   | 93.1 3.46 | 93.65 2.77 | 92.52 4.01 |
| DARPA   | 97.08 0.34 | 95.8 0.25   | 93.24 3.37 | 93.13 3.61 | 93.88 2.55 |
| wide    | 96.74 1.42 | 94.59 1.92  | 94.59 3.6  | 93.52 2.21 | 92.32 3.24 |
| ISP     | 98.48 1.94 | 94.64 1.33  | 94.71 2.3  | 93.72 2.78 | 92.54 4.42 |

2) Runtime performance: The runtime measurement was repeated ten times for each method/approaches to give the average processing time with high confidence level. Table 3 shows the runtime measurement results. The ORTSC, PGM and BGCM are much faster than the Semi-Stack approach. The Semi-Stack is slightly faster than SemTra. Figure 5 shows that ORTSC and BGCM methods tend to scale better than PGM and Semi-Stack.

**Table 3. Runtime performance in labelling process.**

| Dataset | Semi-Stack | SemTra | PGM | BGCM | ORTSC |
|---------|------------|--------|-----|------|-------|
|         | Mean St.D  | Mean St.D | Mean St.D | Mean St.D | Mean St.D |
| ITD     | 64.85 79.75 | 1.24 1.87 | 0.39 |
| DARPA   | 54.79 67.38 | 0.67 1.26 | 0.21 |
| wide    | 30.48 37.49 | 0.86 1.77 | 0.28 |
| ISP     | 64.11 78.85 | 1.25 1.90 | 0.39 |

**Figure 5. Experimental results on scalability.**

3) Stability: The Semi-Stack approach also aims to increase the stability in producing accurate class labels for unlabeled flows for all possible runs. The number of trials was set to 100 in order to obtain higher confidence level of the results. The experiments use 30% of data in each class for testing and the remaining data was randomly selected to form testing. Table 4 exhibits the Semi-Stack approach achieved the best stability performance compared to the baseline methods with margin between 2.38% and 12.3%. The rational is that the Semi-Stack approach combines the advantages of several machine learning methods, which results in robustness of assigning class labels with low variance and bias of
the predictive model. Notably, the lowest stability belongs to the ORTSC method on all traffic data, this due to the sensitive and random initialization of K-means. The BGCM method was the second worse stability result, especially on ITD and DARPA datasets.

Table 4. Stability performance comparison.

| Dataset | Semi-Stack | SemTra | PGM | BGCM | ORTSC |
|---------|------------|--------|-----|------|-------|
| ITD     | 97.44      | 97.35  | 94.08 | 90.98 | 90.82 |
| DARPA   | 99.50      | 98.98  | 93.77 | 91.48 | 91.55 |
| Wide    | 98.11      | 98.71  | 94.90 | 92.73 | 92.93 |
| ISP     | 98.44      | 97.91  | 94.97 | 91.55 | 90.82 |

5. Conclusion

This paper proposed a novel semi-supervised traffic classification approach, namely Semi-Stack. The proposed Semi-Stack approach combines the advantages of several machine learning methods, including: feature projection, ensemble learning and stacking, to enhance and improve the process of network traffic classification with low variance and bias of the predictive classifier model. Experiment results on four existing network traffic datasets verified the significant improvements of the proposed approach’s classification capability. Even the Semi-Stack approach has the second worse runtime, however it reached the best average accuracy marks, by performing accuracy above 96% on all datasets. Besides, the Semi-Stack approach achieved the best stability performance up to 99.50% on DARPA dataset compared to the baseline methods with margin between 2.38% and 12.3%.

The author considers the use of parallel computing, i.e. Graphics Processing Units (GPUs) or multi-core processing (CPU) to minimize the execution time as one of future works.

References

[1] Mawi working group traffic archive. Available: http://mawi.wide.ad.jp/mawi/.
[2] Bernaille L, Teixeira R, Akodkenou I, Soule A, Salamatian K 2006 Traffic classification on the fly, ACM SIGCOMM Comput. Communication Rev. 36(2) 23-6.
[3] Erman J, Mahanti A, Arlitt M 2006 Qrp05-4: Internet traffic identification using machine learning. IEEE Global Telecommunications Conf. 1-6.
[4] Erman J, Arlitt M, Mahanti A 2006 Traffic classification using clustering algorithms, SIGCOMM workshop on Mining network data, 281-6.
[5] Erman J, Mahanti A, Arlitt M, Cohen I, Williamson C 2007 Offline/real-time traffic classification using semi-supervised learning. Perform. Evaluat. 64(9) 1194-213.
[6] Karagiannis A, Broido M, Faloutsos, et al. 2004 Transport layer identification of p2p traffic, 4th ACM-SIGCOMM conf. Internet measure. 121-34.
[7] Karagiannis T, Papagiannaki K, Faloutsos M 2005 Blinc: multilevel traffic classification in the dark. ACM-SIGCOMM Comput. Communication Rev. 35 229-40.
[8] McGregor A, Hall M, Lorier P, Brunskill J 2004 Flow clustering using machine learning techniques. Passive Act. Network Measure., Springer, 205-14.
[9] Moore A, Hall J, Kreibich C, Harris E, Pratt I 2003 Architecture of a network monitor. Passive & Active Measurement Workshop2003 (PAM2003). Citeseer.
[10] Moore A.W, Papagiannaki K 2005 Toward the accurate identification of network applications. Passive Act. Network Measure. Springer, 41-54.
[11] Moore A.W, Zuev D 2005 Internet traffic classification using Bayesian analysis techniques. ACM-SIGMETRICS Perform. Evaluation Rev. 33 50-60.
[12] Nguyen T T, Armitage G A 2008 survey of techniques for internet traffic classification using machine learning. IEEE Communications Surveys Tutor. 10(4) 56-76.
[13] Reich Y, Fenves S J 1991 The formation and use of abstract concepts in design concept formation: knowledge and experience in unsupervised learning. Citeseer.
[14] Rotsos C, Van Gael J, Moore A W, Ghahramani Z. 2010 Probabilistic graphical models for semi-supervised traffic classification. *Proceedings of the 6th International Wireless Communications and Mobile Computing Conference*, 752-7.

[15] Stolfo S J, Fan W, Lee W, Prodomidis A, Chan P K 2000 Cost-based modelling for fraud and intrusion detection: Results from the jam project. *IEEE-DARPA Information Survivability Conference and Exposition, (DISCEX’00)*, 2 130-44.

[16] Wang Y, Xiang Y, Zhang J, Yu S 2011 A novel semi-supervised approach for network traffic clustering, *5th IEEE Int. Conf. Network Syst. Security (NSS)*, 169-75.

[17] Wang Y, Xiang Y, Zhang J, Zhou W, Wei G, Yang L T 2013 Internet traffic classification using constrained clustering. *IEEE Transact. Parallel Distributed Syst.* p.1.

[18] Witten I H, Frank E, Trigg L E, Hall M A, Holmes G, Cunningham S J 1999 Weka: Practical machine learning tools and techniques with java implementations.

[19] Zander S, Nguyen T, Armitage G 2005 Automated traffic classification and application identification using machine learning. *30th Anniversary IEEE Conf. Local Comput. Networks*, 250-7.

[20] Fahad A, Almalawi A, Tari Z, Alharthi K, Al Qahtani F S, Cheriete M 2019 SemTra: A semi-supervised approach for traffic flow labeling with minimal human efforts, *Pattern Recognition, Elsevier J.* 91 1-12.