A Review of Internet Traffic Classification and Clustering Based on Machine Learning Technique

Nikita Shrivastava
M Tech Scholar
OCT Bhopal, India
E-mail- nikitashrivastava147@gmail.com

Prof. Amit Dubey
HEAD, Department of CSE
OCT Bhopal, India
E-mail- amitdubey@oriental.ac.in

ABSTRACT
Internet traffic classification is very important area of research in the field of traffic engineering. The internet traffic data contains two types of data one is header data and another is payload data. Now a day's used machine learning technique for the purpose of classification and clustering of internet traffic data. In this paper presents the review of clustering and classification method of internet data classification. In this paper discuss the feature based traffic clustering and classification technique. Feature based clustering and classification technique used a feature extraction technique of traffic data.

Keyword: - internet Traffic, Data Mining, Classification.

INTRODUCTION

The classification of network traffic data is very important aspect of traffic engineering. The process of internet traffic is very complex due to this reason the process of classification technique is very complex. The traditional approach of traffic classification has been very successful in the past it depends on mapping applications to well-known port numbers. In order to avoid detection through this approach, P2P applications started disguising themselves by using port numbers for commonly used protocols such as FTP and HTTP and started using dynamic port numbers. Many recent studies confirm that port-based identification of network traffic is ineffective [8]. Various technique of internet traffic classification based on the process of data payload and process of protocol. Some classification technique based on the port traffic analysis. The payload approach, packet are analyzed to determine whether they contain characteristic signatures of known applications. The classification method based on machine learning is very efficient. Instead of that some market based packet shaping tools have started using these techniques. On the other hand, P2P applications such as Bit Torrent are beginning to escape this technique by using obfuscation methods such as plain-text ciphers, variable-length padding, and encoding of data. In addition, there is some other limitation [5]. First, these techniques only identify traffic for which signatures are available and are unable to classify any other traffic. The common way is to extract statistical features to represent the traffic flows and then apply machine learning (ML) techniques for classification. The ML algorithms used in this area can be generally divided into two categories, i.e. classification (or supervised learning) [6] and clustering (or unsupervised learning) [7]. In classification, there is a known, fixed set of classes, and pre-labeled training data is taken to induce a classifying model. On the other hand, clustering algorithms partition an unlabeled data set into groups of similar items, typically by optimizing an objective function. Compared with supervised learning, clustering has some key advantages such as the elimination of requirements for fully labeled training data sets and the ability to discover hidden classes that may represent previously unknown applications. In section II discuss related work in internet classification. In section III discuss clustering and classification technique. In section IV feature extraction technique and finally discuss conclusion and future work in section V.

II. RELATED WORK

In this section discuss the related work of internet traffic classification using different machine learning algorithm. The machine learning algorithm provides a Varity of algorithm for classification and classification. The process of clustering used different types of clustering algorithm. Instead of that discuss classification algorithm also.

[1] In this paper author presents a scheme of constrained clustering that makes decisions to the observed traffic statistics with consideration of few background information. Specifically, we make use of equivalence set constraints indicating that particular sets of flows are using the same application layer protocols, which can be efficiently inferred from packet headers according to the background knowledge of TCP/IP networking. We model the observed data and constraints using Gaussian mixture density and adapt an approximate algorithm for the maximum likelihood estimation of model parameters. Moreover, we study the effects of unsupervised feature discretization on traffic clustering by using a fundamental binning method. A number of real-world Internet traffic traces have been used in our evaluation, and the results show that the proposed approach not only improves the quality of traffic clusters in terms of overall accuracy and per-class metrics, but also speeds up the convergence.

[2] In this paper author propose a scheme of novel traffic classification to improve classification performance when few training data are available. In the proposed scheme, traffic flows are described using the discredited statistical features and flow correlation information is modeled by bag-of-flow (BOF). We solve the BOF-based traffic classification in a classifier combination framework and theoretically analyze the performance benefit. Furthermore, a new BOF-based traffic classification method is proposed to aggregate the naive Bays (NB) predictions of the correlated flows.

[3] In this paper author 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. They find that C4.5 performs the best under any circumstances, as well as the reason why; be-cause the algorithm discreteness input features during classification operations. They also find that the entropy-based Minimum Description Length discretization on ports and packet size features substantially improve the classification accuracy of every machine learning algorithm tested (by as much as 59.8%!) and make all of them achieve >93% accuracy on average without any algorithm septic tuning processes.

[4] In this work author present a novel semi-supervised learning method using constrained clustering algorithms. The motivation is that in network domain a lot of background information is available in addition to the data instances themselves. For example, we might know that flow f1 and f2 are using the same application protocol because they are visiting the same host address at the same port simultaneously. In this case, f1 and f2 shall be grouped into the same cluster ideally. Therefore, we describe these correlations in the form of pair-wise must-link constraints and incorporate them in the process of clustering. We have applied three constrained variants of the K-Means algorithm, which perform hard or soft constraint satisfaction and metric learning from constraints.

[5] In this paper author provide context and motivation for the application of ML techniques to IP traffic classification, and review 18 significant works that cover the dominant period from 2004 to early 2007. These works are categorized and reviewed according to their choice of ML strategies and primary contributions to the literature. We also discuss a number of key requirements for the employment of ML-based traffic classifiers in operational IP networks, and qualitatively critique the extent to which the reviewed works meet these requirements. Open issues and challenges in the field are also discussed.

[6] In this paper author presents a statistical analysis of the amount of information that the features of traffic flows observed at the packet-level carry, with respect to the protocol that gen-erased them. We show that the amount of information of the majority of such features remain constant irrespective of the point of observation (Internet core vs. Internet edge) and to the capture time. The results of the analysis show that the information carried by the main packet-level features of Internet traffic flows tends to remain rather constant both in space and in time. A few exceptions also emerge from our analysis, and are discussed in the rest of this work. Finally, we present a brief comparative analysis of how several classification mechanisms fare in relation to the information carried by the features evaluated earlier.

[7] In this paper author presents work considers two unsupervised clustering algorithms, namely K-Means and DBSCAN that have previously not been used for network traffic classification. We evaluate these two algorithms and compare them to the previously used Auto-Class algorithm, using empirical Internet traces. The experimental results show that both K-Means and DBSCAN work very well and much more quickly than Auto-Class. Our results indicate that although DBSCAN has lower accuracy compared to K-Means and Auto-Class, DBSCAN produces better clusters.

[10] In this paper to categorize traffic by application, author apply a naive bays estimator. Uniquely, our work capitalizes on hand-classier network data, using it as input to a supervised naive bays estimator. In this paper we illustrate with the naive bays estimator the high level of accuracy can be achieved. We further illustrate the improved accuracy of reined variants of this estimator. their results indicate that with the simplest of naive bays estimator we are able to achieve about 65% accuracy on per-flow classification and with two powerful refinement we can improve this value to better than 95%; this is a vast improvement over traditional techniques that achieve 50–70%. While our technique uses training data, with categories derived from packet-content, all of our training and testing was done using header-derived discriminators. We emphasize this as a powerful aspect of our approach: using samples of well-known traffic to allow the categorization of traffic using commonly-available information alone.

[11] In this paper, for Internet traffic identification author apply an unsupervised machine learning approach and compare the results with that of a previously applied supervised machine learning approach. Supervised approach uses the naive bays classifier and unsupervised approach uses an Expectation Maximization (EM) based clustering algorithm . We find the unsupervised clustering technique has accuracy up to 91% and outperform the supervised technique by up to 9%. We also come to know that in order to discover traffic from previously unknown applications unsupervised technique can be used and for exploring Internet traffic it has the potential to become an excellent tool.

III CLUSTERING TECHNIQUE

Partition clustering technique is very important clustering technique in data mining. This clustering technique gives various clustering algorithm such as k-means, FCM k-maids and optics. Basically such type of clustering technique used in small data range. The size of data are increase used density

based clustering technique [4]. In this clustering technique two important algorithm are used such as DBSCAN and IDBSCAN. Both the algorithm is very efficient but faced a problem of cluster id selection. The hierarchal clustering technique used in depth of data analysis. In this clustering technique used two clustering algorithm one is Agglomerative clustering and another one is dividend clustering technique [2]. The agglomerative cluster process small cluster into big cluster for merging a pattern. Instead of that divined clustering technique used for break big pattern into small pattern. The design of the internal indices is based on three elements: the data set, the point level partitions, and centroids [1]. Mean square error (MSE) is a conventional criterion for evaluating clustering, which is calculated by these three elements. External indices, however, use only partitions by comparing the given clustering against the ground truth. The ground truth is usually built by using human assessors or the output of another clustering algorithm. External indices count the pairs of points of agreement or disagreement of the two partitions. A criterion such as MSE uses quantities and features inherent in the dataset, which gives a global level of evaluation. Since it relates to points and clusters, its time complexity is at least O (MN) [11]. The partition-based criteria are based on point wise evaluation of two partitions, which usually gives a time complexity of O (N2).

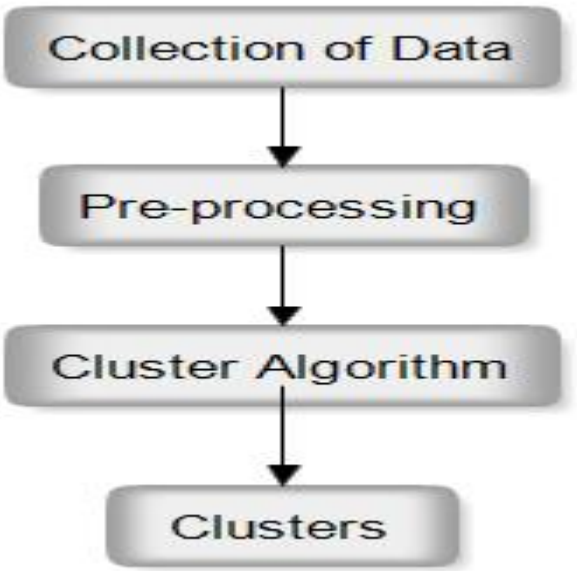

Figure 1: The stages of the process of clustering.

For the purpose of internet traffic data clustering various clustering algorithm are applied, such as clustering, weighted clustering, and regression. Two of the most critical and well generalized problems of traffic data are its new evolved feature and concept-drift. Since a traffic data is a fast and continuous event, it is assumed to have infinite length. Therefore, it is difficult to store and use all the historical data for training. The most discover alternative is an incremental learning technique. Several incremental learners have been proposed to address this problem. In addition, concept-drift occurs in the traffic when the underlying concepts of the traffic change over time. A variety of techniques have also been proposed in the literature for addressing concept-drift [11], in data traffic clustering. However, there are two other significant characteristics of data multi-categories, such as concept evolution and feature evolution that are ignored by most of the existing techniques. Concept-evolution occurs when new classes evolve in the data. On the category process we found some important problem in cluster oriented traffic data clustering. These problems are given below.

Traffic data clustering suffered from multiple feature evaluation,

- Selection of number of cluster for multi-level [1,3,6].
- Diversity of feature selection process [10].
- Boundary value of cluster
- Outlier data treat as noise
- Number of iteration
- Increase value of MSR

IV FEATURE EXTRACTION

Traffic data can either have single variable approach or a multi-variable approach to classier internet depending on the algorithm used. In the single variable approach a single variable of the system is analyzed. This can be, for example, port number, CPU usage of a local machine etc. In multi-variable approach a combination of several features and their inter-correlations are analyzed. [4] In addition based on the method the way in which features are chosen for the ITA can be divided into two groups; into feature selection and feature reduction.

In the feature selection method the features are either picked manually from the data monitored or by using a specific feature selection tool. The most suitable features are selected by handpicking from the feature spectrum based on the prior knowledge about the environment that the IDS are monitoring. For example features that can distinguish certain type of traffic from the traffic flows are picked for the network traffic model training. The idea behind the feature selection tools is to reduce the amount of features into a feasible subset of features that do not correlate with each other. Examples of feature selection tools are Bayesian networks (BN) and classification and regression tree (CART). Bayesian network is a probabilistic graphical model that represents the probabilistic relationships between features [9]. CART is a technique that uses tree-building algorithms to construct a tree-like if-then prediction patterns that can be used to determine different classes from the dataset. [4] Feature selection process is illustrated in Figure 2 On the left there are the features (F0…FN) that are available from the data monitored, which is, for example, from network traffic. On the right side is the output (F0...FM) of the selection tool. The number of features in the output varies based on the selection tool used and the inter-correlation of features in the input. Following the basic principles of feature analysis the number of features in the output (M in Figure2) is in most of the cases less than the number of features in the input (N in Figure 2). However, it is possible that the output is equal to the input.

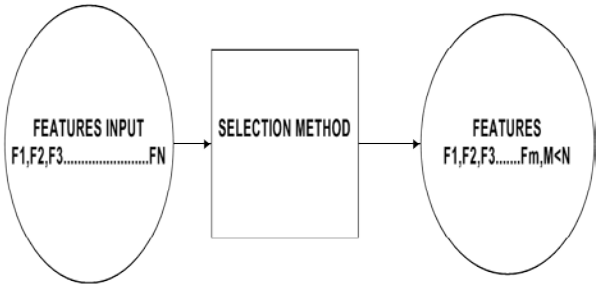

Figure 2: Feature selection process in feature variable.

V CONCLUSION AND FUTURE WORK

In this paper we have present the internet traffic clustering of supervised Machine-Learning to classify network traffic by application. We further discuss the technique using test and training sets separated by traffic data. The technique we have described uses our ability to train the classier with known data. We have demonstrated that a classification model constructed using this technique is then able to be applied when far less information is available about the traffic. Critically, we have illustrated a classification technique that may be used retrospectively on data-traces that have previously not been examined in any detail due to the lack of complete information.

REFERENCES

[1]Hue and Lie "Internet traffic classification Using Constrained Clustering" IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, 2013. Pp 1-11.

[2] Jun Zhang, Chao Chen, Yang Xiang,WanleiZhou, Yong Xiang "Internet Traffic Classification by Aggregating Correlated Naive Bayes Predictions" IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL-8, 2013. Pp 5-16.

[3] Yeon-sup Lim, Hyun-chul Kim, JiwoongJeong, Chong-kwon Kim "Internet Traffic Classification Demystified: On the Sources of the Discriminative Power" ACM, 2010. Pp 1-12.

[4] Yu Wang, Yang Xiang, Jun Zhang, Shunzheng Yu "A Novel Semi-Supervised Approach for Network Traffic Clustering" IEEE, 2011. Pp 169-174.

[5] Thuy T.T. Nguyen, Grenville Armitage "A Survey of Techniques for Internet Traffic Classification using Machine Learning" IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL-10. 2008. Pp 56-76.

[6] Alice Este, Francesco Gringoli, Luca Salgarelli "On the Stability of the Information Carried by Traffic Flow Features at the Packet Level" ACM, 2009. Pp 13-19.

[7] Jeffrey Erman, Martin Arlitt, AnirbanMahanti "Traffic Classification Using Clustering Algorithms" SIGCOMM'06 Workshops, 2006. Pp 281-286.

[8] MarcinPietrzyk, Jean-Laurent, Guillaume Urvoy-Keller, Taoufik "Challenging Statistical Classification for Operational Usage: the ADSL Case" ACM, 2009. Pp 1-14.

[9] Jeffrey Erman ,AnirbanMahanti, Martin Arlitt, Ira Cohen, Carey Williamson "Semi-Supervised Network Traffic Classification" ACM, 2010. Pp 1-2.

[10] Andrew W. Moore, Denis Zuev "Internet Traffic Classification Using Bayesian Analysis Techniques" ACM, 2005. Pp 50-61.

[11] Jeffrey Erman, AnirbanMahanti, Martin Arlitt "Internet Traffic Identification using Machine Learning" 2007. Pp 1-6.