Selection of Online Network Traffic Discriminators for on-the-Fly Traffic Classification[*]

Angela María Vargas Arcila[**]
Juan Carlos Corrales Muñoz[***]
Alvaro Rendon Gallon[****]
Araceli Sanchis[*****]

Received: 20/11/2019 • Approved: 23/06/2020
https://doi.org/10.22395/rium.v20n38a4

Abstract

There are several techniques to select a set of traffic features for traffic classification. However, most studies ignore the domain knowledge where traffic analysis or classification is performed and do not consider the always moving information carried in the networks. This paper describes a selection process of online network-traffic discriminators. We obtained 24 traffic features that can be processed on the fly and propose them as a base attribute set for future domain-aware online analysis, processing, or classification. For the selection of a set of traffic discriminators, and to avoid the inconveniences mentioned, we carried out three steps. The first step is a context knowledge-based manual selection of traffic features that meet the condition of being obtained on the fly from the flow. The second step is focused on the quality analysis of previously selected attributes to ensure the relevance of each one when performing a traffic classification. In the third step, the implementation of several incremental learning algorithms verified the usefulness of such attributes in online traffic classification processes.

Keywords: incremental learning; network traffic classification; online classification; traffic feature selection.

[*] This paper presents partial results of the doctoral project *Fault Diagnosis for IP-Based Network with Real-Time Conditions*, which is funded by the Department of Science, Technology, and Innovation –Colciencias (Scholarship program no.° 647) and supported by the Spanish Government under project TRA2016-78886-C3-1-R. The doctoral project is developed at the University of Cauca (Colombia) and the Carlos III University of Madrid (Spain).

[**] MSc in Telematics Engineering from Universidad del Cauca. Ph.D. student in Telematics Engineering at Universidad del Cauca, and Computer Science and Technology at Universidad Carlos III de Madrid. E-mail: amvargas@unicauca.edu.co. Orcid: https://orcid.org/0000-0002-4313-5445

[***] Ph.D. in Computer Science from Université de Versailles Saint-Quentin-en-Yvelines. Full-time Professor and Leader of the Telematics Engineering Group at Universidad del Cauca. E-mail: jcorral@unicauca.edu.co. Orcid: https://orcid.org/0000-0002-5608-9097

[****] Ph.D. in Telecommunications Engineering from Universidad Politécnica de Madrid. Full-time professor and director of the Doctoral Program in Telematics Engineering at Universidad del Cauca. E-mail: arendon@unicauca.edu.co. Orcid: https://orcid.org/0000-0002-2935-7316

[*****] Ph.D. in Computer Science from Universidad Politécnica de Madrid. Ph.D. in Physical Chemistry from Universidad Complutense de Madrid. University Associate Professor of Computer Science at Universidad Carlos III de Madrid. E-mail: masm@inf.uc3m.es. Orcid: https://orcid.org/0000-0002-1429-4092

Selección de discriminadores de tráfico de red para clasificación en tiempo real

Resumen

Existen varias técnicas para seleccionar un conjunto de variables para clasificación del tráfico de red. Sin embargo, muchos estudios ignoran el ámbito del conocimiento en donde el análisis y clasificación del tráfico tiene lugar y no consideran la información, siempre en movimiento, que se transporta en dichas redes. Este artículo describe el proceso de selección de discriminadores tráfico de redes en línea. Se obtuvieron 24 características que pueden procesarse en tiempo real y se proponen como los conjuntos de atributos base para futuros análisis, procesamiento y calificación conscientes del dominio (*domain-aware*). Para la selección de un conjunto de discriminadores de tráfico y con el fin de evitar los inconvenientes mencionados anteriormente, se llevaron a cabo tres etapas. La primera consiste en la selección manual basada en el conocimiento contextual de las características de tráfico de red que tengan las condiciones de obtener en tiempo real a partir del flujo. La segunda etapa se enfoca en la calidad del análisis de los atributos previamente seleccionados para asegurar la relevancia de cada uno a la hora de efectuar la clasificación del tráfico. En la tercera etapa, la implementación de varios algoritmos de aprendizaje incremental verifican la idoneidad de tales atributos en procesos de clasificación de tráfico en línea.

Palabras clave: aprendizaje incremental; clasificación de tráfico de la red; clasificación en línea; selección de características de tráfico.

INTRODUCTION

Operators have frequently used network traffic classification as an instrument because it allows performing traffic analysis to differentiate and prioritize traffic for several purposes. These purposes go from anomaly detection to profiling user resource requirements [1]. Especially, traffic classification is the first step for anomaly detection activities such as intrusion detection by finding attack patterns, fault identification, identify customer use of network resources that in some way infringes the terms of the operator [2-3].

In order to allow traffic classification, the research community has characterized the traffic through different types of discriminators, from well-known port numbers to more sophisticated data such as recognizing statistical patterns in externally observable attributes of the traffic [3]. However, most studies ignore the domain knowledge where traffic classification is performed, for example, if real-time classification is required or if traffic-generating applications wrap their application protocols in others, or a combination of several other conditions.

On the other hand, several algorithms for performing traffic classification have been implemented, mostly based on traditional machine learning without considering that the information carried in the networks is always moving, which is the most relevant feature of networking. As a consequence, new works like [4], focused on incremental learning techniques that can also perform an on-the-fly classification, called online, have emerged. These new approaches allow network operators quickly very quick detection and reaction of network anomalies, security breaches, and, at the same time, planning ahead to adapt their networks to novel usage patterns.

This paper proposes a set of attributes for traffic characterization as a base guideline to perform a traffic classification without the need to wait for a complete communication flow to know its nature. This means that the selected traffic attributes in this work are those obtained only on the fly, the reason why it is called online traffic features. In addition to this, the proposed features are a numerous set of attributes on which a re-selection can be performed; any work related to this scenario can use it as a basis for the selection of online attributes for more specific conditions. For example, in a network where traffic generating applications wrap their application protocols in Http, port-related information is not relevant for traffic classification.

The attributes set proposed in work [5]was used as a reference to select those attributes or traffic characteristics, because it provides a wide variety of features to characterize flows, and is the work found by us with the greatest number of traffic discriminators. Consequently, we used the dataset constructed in the same work [5], Cambridge dataset from now on, to perform the research analysis.

A context analysis was carried out on the basis of that set, that is, to determine which traffic characteristics can be obtained before a communication flow ends. Subsequently, a quality analysis was applied to determine which of the online traffic attributes selected by context are irrelevant to obtain the type of traffic. As a result, we obtained a set of 24 attributes.

Finally, to verify the usefulness of the selected traffic characteristics in on-the-fly classification environments, a comparison was performed with the dataset described in [6], whose 11 attributes are also online traffic characteristics and has obtained good performance in online traffic classification processes [6]. In this comparison, ten incremental learning algorithms were used to evaluate the accuracy of the classifications obtained and the performance in time.

The main difference between the attributes considered in these two datasets is that [6] takes into account the online features related only to flow size and physical entities, which correspond with the port numbers involved in communication. Instead, our dataset considers all possible online attributes, therefore, it covers not just size attributes and physical entities, but also time attributes, features related to the flags used at the beginning of a TCP communication (e.g. number of SYN and ACK packages) and flow features (e.g. the number of bytes seen in the initial flight of data).

As a result, this work proposes 24 online features as a base guideline for on-the-fly traffic analysis and online classification processes because of three issues. First, the compared sets of attributes (our 24 attributes set and the 11 attributes set of [6]) showed similar behavior for all the algorithms used. Second, the classification processes show good behavior with different incremental algorithms using the proposed features. Third, these 24 discriminators comprise the widest possible set of relevant online attributes and include the 11 features set with which the comparison was made.

The organization of the paper is as follows. First, we present an overview of related works. Secondly, we describe the dataset with the selection of online features and quality diagnosis processes. Third, we explain the experimental setup and the results and discussion. Finally, we present the conclusions of the work.

1. RELATED WORK

Network traffic classification research can be divided, according to the information used for the classification, into port, payload, statistical measurement and behavioral traffic properties based methods [1]. The common element among the approaches above is the flow defined by a sequence of packets that are part of the same process-to-process communication.

The port-based approach determines the type of traffic according to the source and destination port numbers that IP flow carries because, broadly, the applications have a well-known and registered protocol port number to which other hosts may initiate communication. This port number is determined by Internet Assigned Numbers Authority (IANA). However, nowadays, many applications of the new generation like over-the-top services do not have registered port numbers and use dynamic ports or wrapping in other protocols like Https; therefore, the port-based classification is impossible [2].

The payload-based approach also named deep packet inspection (DPI), analyzes session and application information from the content that carries each packet of the flow. Although this method is very accurate, it cannot process the encrypted content packet and requires computational overhead and additional hardware to achieve efficient processing [7].

Due to the problems encountered in port and payload based approaches, classification based on statistical traffic properties is generally used. This method uses statistical characteristics like average packet length, packet inter-arrival time between server and client and vice versa, distribution of flow duration, flow idle time, and so forth, which are unique for each type of traffic generated by an application [3].

The behavioral traffic properties based approach is lightweight, and it also avoidsaccess to packet payload. This method looks at the whole traffic received by a network element and analyzes traffic patterns such as the number of connected hosts, their transport layer protocols, how many different ports they use, etc., because it assumes that different applications generate different traffic patterns [8].

As noted in [1] and [2-7], there are several works with different methodologies for traffic classification which apply the above approaches. However, at the same time, they use different traffic traces, classification features, and traffic classes, hence comparing them is a difficult task. This is why this work focuses on the dataset described in[5], from now Cambridge dataset, used for years in multiple research with port, payload, and statistical measurement-based classifiers. In our search for a dataset, the Cambridge dataset provided the widest variety of features found to characterize flows, which allows us to make a wide selection of the online traffic discriminators to describe flows.

Although the Cambridge dataset was created in 2005, its attributes are still in force today because the Internet carries the data by the same network protocol stack since the early 1980s. Cambridge features are derived using packet header information, simple statistics about packet length, inter-packet timings and information from the transport protocol; therefore, the current traffic can be described by those attributes. This dataset was collected with the network monitor described in [9] and developed by the University of Cambridge Computer Laboratory. The traffic capture was performed in a full-duplex Gigabit Ethernet link with about 1,000 users connected to a research facility for 24 hours.

We review the most significant works that cover the period from 2005 to 2018 related to the Cambridge dataset. Two types of work were found. The first one focused on feature selection techniques approaches [10–12] for searching how to choose the right features, which can bring more precise results in the traffic classification task. On the second one, focused on traffic classification techniques approaches. These works, according to their machine learning techniques used, maybe categorized as supervised, unsupervised, and incremental learning approaches as listed in table 1. For each work, the number of classes, features, attribute selection approaches and information used for the classification are indicated.

Table 1. Cambridge Dataset Related Works.

Work	#Classes	#Features	Attribute selection approach
Supervised approaches			
[13]	12	All	--
[14]	12	All	--
[15]	9	5*(P,S)*	A
[16]	10	All	--
[17]	10	12 *(P,S)*	A
Unsupervised approaches			
[18]	2 *(P2P, NOT-P2P)*	5(P,S)	A
Incremental learning approaches			
[6]	10	11 *(P,S)*	C
[19]	10	All	--
[20]	2 *(ATTACK, NOT-ATTACK)*	All	--
[21]	12	All	--

Source: own elaboration.

The works using 9 to 12 classes (being 12 all classes in Cambridge dataset), usually merge bulk traffic type into a single class and eliminate games, interactive and multimedia classes less frequent in the dataset. Nevertheless, at this point, it is crucial to clarify the current circumstances about games, interactive, and multimedia, which have not been reflected in the used dataset. In recent years, gaming has become the most extensive virtual leisure activity [22]. Also, games are used for other purposes that are intended to create pleasant experiences in virtual non-gaming environments. Then, those traffic types become present in our daily interaction with the Internet without even recognizing it [23].

On the other hand, those works which use few classes are focused on the classification of a single traffic type, generally P2P and Attack, and merge the rest of the traffic type in another single class. There is also one work [15] focused on single traffic but using all existing classes.

Column #Features indicates the number of features used by the work. For works using fewer features than the original dataset, this column also describes which type of information is used for traffic classification through two letters: P and S. “P” if the features are based on port, “S” if the features are statistical traffic measures. Related works that use all the dataset features perform the classification with port (P), statistical (S), and payload information.

Finally, if a work has a value in the last column, it means that it performs a feature selection process by the algorithm (A) or based on the domain knowledge for which traffic classification is performed (C).

This paper focuses on the features that can be calculated on the fly because they enable streaming solutions for the analysis of network traffic and allow an online classification. Therefore, Loo research [6] will be the reference for our experiments because it only uses online attributes and performs an incremental approach.

2. DATA UNDERSTANDING AND PREPARATION

Cambridge is a real TCP traffic dataset described in [5] and captured in one day, split into ten blocks of 28 minutes each. It contains 397,152 instances, which each one represents a flow. This dataset is intended to provide a wide variety of features to characterize flows; then, each flow is described by 248 discriminators or features. IP traffic is caused by 12 application types which are: WWW, Mail, FTP-Control, FTP-Pasv, Attack, P2P, Database, FTP-Data, Multimedia, Services, Interactive, and Games. Feature selection and quality diagnosis are described below.

2.1 Selection of online features

The features selection task is essential to significantly improve the accuracy and computational performance of traffic classification [24]. More specifically, our focus is on those features that allow a classification when real-time or near real-time traffic identification is required, for example, a scenario where the traffic identification is one of the first steps to timely detect anomalies in a network and must be executed on the fly. In other words, our purpose is to select a set of features as a guideline for online traffic classification.

This selection task was done combining two approaches, first through domain and context knowledge, and second through an algorithm, choosing a subset of features that can identify a class as effectively as many works presented in the related work section.

For the first approach, it is important to keep in mind that an online traffic feature is one that can be calculated on the fly before a flow is completed. So, through an analysis of the 248 traffic discriminators of the Cambridge dataset, 32 online attributes were found.

In this analysis, we consider those features that can be extracted from any package belonging to the flow (e.g., port numbers). The first quartile features which are the first statistical characteristics of the flow (e.g., first quartile inter-arrival time, first quartile of total bytes in IP packet). The features that can be extracted during the opening of the connection or 3-way handshake (e.g., number of packets with the SYN bits set, the round-trip time value calculated from the TCP 3-way handshake). Moreover, the features that can be obtained in the first TCP window (e.g., the total number of bytes sent in the initial window, the total number of packets sent in the initial window).

These attributes were categorized by physical entities, time, flags, size and flow attributes in accordance with [24]. The results of this selection are shown in table 2.

Table 2. Cambridge online features.

Type	ID as in Cambridge dataset [5]	Feature Name	Description
Phy. entities	1	Server_Port	Source port number
	2	Client_Port	Destination port number
Time	4	q1_IAT	First quartile inter-arrival time
	196	q1_IAT_ab	First quartile of packet inter-arrival time (uplink)
	203	q1_IAT_ba	First quartile of packet inter-arrival time (downlink)

Type	ID as in Cambridge dataset [5]	Feature Name	Description
Flags	61	SYN_pkts_sent_ab	The count of all the packets seen with the SYN bits set in the TCP header respectively (ab: uplink, ba: downlink)
	63	SYN_pkts_sent_ba	
	69	adv_wind_scale_ab	The window scaling factor used. (ab: uplink, ba: downlink)
	70	adv_wind_scale_ba	
	71	req_sack_ab	If the end-point sent a SACK permitted option in the SYN packet opening the connection, a 'Y' is printed; otherwise 'N' is printed. (ab: uplink, ba: downlink)
	72	req_sack_ba	
	79	mss_requested_ab	The Maximum Segment Size (MSS) requested as a
	80	mss_requested_ba	TCP option in the SYN packet opening the connection. (ab: uplink, ba: downlink)
	123	RTT_from_3WHS_ab	The RTT value calculated from the TCP 3-Way Hand-Shake (connection opening), assuming that the SYN packets of the connection were captured. (ab: uplink, ba: downlink)
	124	RTT_from_3WHS_ba	
Size	11	q1_data_wire	First quartile of total bytes in Ethernet packet
	18	q1_data_ip	First quartile of total bytes in IP packet
	25	q1_data_control	First quartile of total of control bytes in packet
	154	q1_data_wire_ab	First quartile of total bytes in Ethernet packet (uplink)
	161	q1_data_ip_ab	First quartile of total bytes in IP packet (uplink)
	168	q1_data_control_ab	First quartile of total of control bytes in packet (uplink)
	175	q1_data_wire_ba	First quartile of total bytes in Ethernet packet (downlink)
	182	q1_data_ip_ba	First quartile of total bytes in IP packet (downlink)
	189	q1_data_control_ba	First quartile of total of control bytes in packet (downlink)
Flow	65	req_1323_ws_ab	If the endpoint requested Window Scaling (ws)/Time Stamp (ts) options, a 'Y' is printed on the respective field. If the option was not requested, an 'N' is printed. (ab: uplink, ba: downlink)
	66	req_1323_ts_ab	
	67	req_1323_ws_ba	
	68	req_1323_ts_ba	
	95	initial_window-bytes_ab	The total number of bytes/packets sent in the initial window, that is the number of bytes/packets seen in the initial flight of data before receiving the first ack packet from the other endpoint. Note that the ack packet from the other endpoint is the first ack acknowledging some data (the ACKs part of the 3-way handshake do not count). (ab: ab: uplink, ba: downlink)
	96	initial_window-bytes_ba	
	97	initial_window-packets_ab	
	98	initial_window-packets_ba	

Source: Adapted from [5]

With this obtained set of online features, the next step is to verify which have less influence on traffic classification in order to remove them from the selected online attributes. This process represents the second approach of the feature selection task and was performed in the dataset quality diagnosis presented below. As you will see, we removed eight features from the Cambridge dataset. Four of them describe if the endpoint requested to increase the window size allowed in TCP protocol (req_1323_ws_ab, req_1323_ts_ab, req_1323_ws_ba, req_1323_ts_ba). Two of them describe the TCP windows scaling factor (adv_wind_scale_ab, adv_wind_scale_ba). And the remaining describes if the TCP connection uses the Selective Acknowledge SACK (req_sack_ab, req_sack_ba).

2.2 Quality diagnosis

The dataset used for quality diagnosis is the Cambridge dataset with the 32 features selected, Cambridge32 dataset from now on. The data of the Games and Interactive classes are not used following [6] decision, because there are not sufficient instances for the training and testing process.

Framework for Data Quality in Knowledge Discovery Tasks (FDQ-KDT) [25] has been chosen to perform quality diagnosis and clean the dataset. Because Cambridge32 comes from a data set already built, only three of the five phases of the framework were executed: data quality diagnosis, select data and clean data.

As a result, incompleteness was detected in eight features. Table 3 indicates these features with the corresponding number of missing values. In some cases, instances with missing values for a feature are the same instances with missing values for another feature, so they have the same missing values.

Table 3. Missing values in the selected dataset.

Feature	# missing values
req_1323_ws_ab	145094
req_1323_ts_ab	
req_1323_ws_ba	
req_1323_ts_ba	
adv_wind_scale_ab	145562
adv_wind_scale_ba	
req_sack_ab	59955
req_sack_ba	

Source: own elaboration.

On the other hand, the dataset may contain irrelevant and redundant features, therefore five algorithms based on feature ranking were executed to obtain features with the lowest rank. The selected algorithms are based on the gain ratio, chi-square, correlation, information gain, and symmetrical uncertainty [26]. In accordance with the common results, *server_port*, *initial_window-bytes_ab*, *initial_window-bytes_ba,*and*q1_data_ip_ab*valuesare more related with the class, while *req_1323_ws_ab* and *adv_wind_scale_ba* are irrelevant for classification.

According to previous results, we decided to delete attributes*req_1323_ws_ab* and *adv_wind_scale_ba* from the Cambrige32 dataset. The other attributes of table 3 were deleted too because the number of missing values oscillates between 15 and 37 percent of the total number of instances, which will negatively affect the classification results according to [27]. Furthermore, it is important to highlight that the high number of missing values mean problems during their collection; it is, therefore, inappropriate to propose them as a guideline for online traffic classification.

In summary, the dataset obtained, from the now Cambridge24 dataset, has 10 classes, 24 attributes, and 397,030 instances.

2.3 EXPERIMENT SETUP

In the previous section, the feature selection procedure was carried out considering the knowledge of the domain and the context of traffic classification. This section aims to compare the performance of several incremental learning algorithms using the dataset composed of the selected attributes. Incremental learning algorithms can learn over-the-fly and perform traffic classification each time a new traffic flow arrives. So, this comparison will verify that our proposed feature set allows for good performance when performing online traffic classification tasks. Furthermore, to confirm that our proposal will serve as a basis for future works, those results will also be compared with the results using a dataset composed of a subset of the selected discriminators.

The experiment uses the Cambridge24 dataset described earlier, and the Cambridge dataset used by [6], which consists of a subset of 11 online features of the Cambridge dataset, Cambridge11 dataset from now on. Cambridge11 dataset has been used in online traffic classification processes where a good performance has been obtained [6]; therefore, our goal is to take it as acceptable threshold when comparing the performance in time of different incremental algorithms for both data sets, while expecting a very similar precision and behavior; otherwise our set of 24 features could not be a guideline for the online traffic classification.

A data stream is an unbounded sequence of data that arrive continuously [28]. The network traffic fits this definition, as well as, data around network analysis such as logs, traffic, monitoring requests, and so on. As clearly stated by [29], data streams are typically divided into two types: static and evolving. Static data streams are those relating to historical data or with a regular bulk arrival. Evolving data streams refer to real-time data, so it updates continuously.

Consequently, ten incremental learning algorithms available in MOA (Massive Online Analysis) [30] and a framework for data stream mining were executed for each dataset to obtain their performance in time [31]. Seven of them are classifiers for static streams:

- Incremental Naive Bayes
- Hoeffding tree
- Hoeffding tree adaptative
- Hoeffding option tree
- OzaBag
- OzaBoost
- OCBoost

And three of them are classifiers for evolving streams:

- OzaBagAdwin
- Single classifier drift
- Ada Hoeffding option tree

The work in [6], we do not use pre-trained algorithms; hence, we assume that there are not pre-collected flow instances and therefore there is not model initialization. This decision was taken because the network traffic is continuous, massive, and its heterogeneity is growing due to the nature of the real-world networks. Hence, it is a challenge to obtain a full model of the network and its traffic. Consequently, it is necessary that the model adapts to the new type of traffic and quickly adjust to the constant network changes. In addition, this allows us to use lighter computational power and minimize computational costs.

As well as in [6], performance measurements were performed after a number of instances or chunks have been received. The indicators measured are immediate accuracy (*Acc$_i$*), cumulative accuracy (*Acc$_c$*), and average accuracy (*Acc$_{avg}$*), with the same formula and parameters (chunk size N$_{\text{chunk}}$ is equal to 1,000) predefined by those authors. The mathematical formulas (1) and (2) correspond to Acc_i and Acc_c; Acc_{avg}i s Acc_c for the last chunk. Table 4 describes each indicator.

$$Acc_i\left(f\right)=\frac{n_{chunk}}{N_{chunk}}\text{x}100\% \tag{1}$$

$$Acc_c\left(f\right)=\frac{\sum_{i=1}^{f}Acc_i}{f} \tag{2}$$

Table 4. Indicators description.

Indicator measured	Description	Formula
Immediate accuracy	Total of instances (traffic flows) correctly classified in a chunk	(1)
Cumulative accuracy	Average of the immediate precisions at the fth chunk	(2)
Average accuracy	Cumulative accuracy for the last chunk	(2); where f variable is the last chunk

Source: own elaboration.

Accuracy is the most common metric used to evaluate the performance of a classification model because it represents the percentage of correct classifications, so the most accurate algorithm makes fewest mistakes. The classification process described here is on the fly, so it is necessary to measure the immediate accuracy that indicates the total of instances correctly classified in an f chunk, and the cumulative accuracy that means the average of the immediate precisions at the fth chunk.

3. RESULTS

Incremental learning algorithms were evaluated through the interleaved test-then-train method, also called the prequential technique. This technique means that each instance can be used to test the model before it is used for training; thus, the accuracy can be incrementally updated [30]. Figure 1 shows the average accuracy obtained by each algorithm for each dataset.

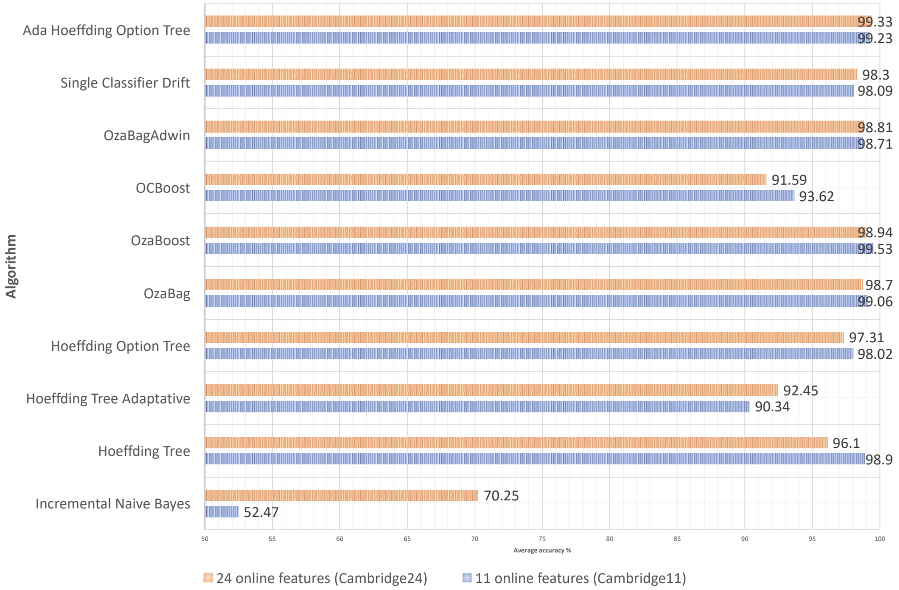
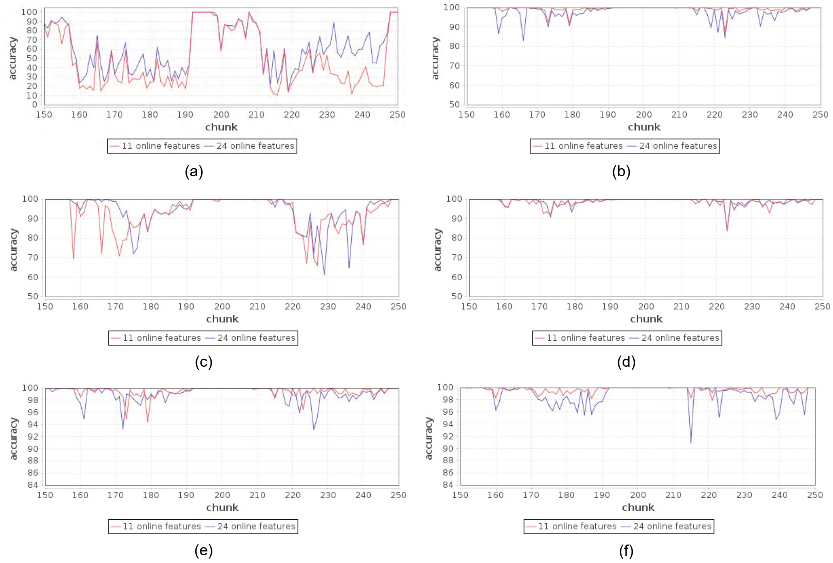


Figure 1. Average accuracy for incremental learning approaches

Source: own elaboration.

The similarity of the results obtained with the two datasets is evident for almost all algorithms. Therefore, it is necessary to analyze the behavior of each algorithm through their accuracy in time series. Figures 2 and 3 show immediate accuracy comparison in time for each of the algorithms. The graphs only show from chunk 150 to chunk 250 to illustrate the difference of accuracy between datasets and to analyze the range of values selected by [6].



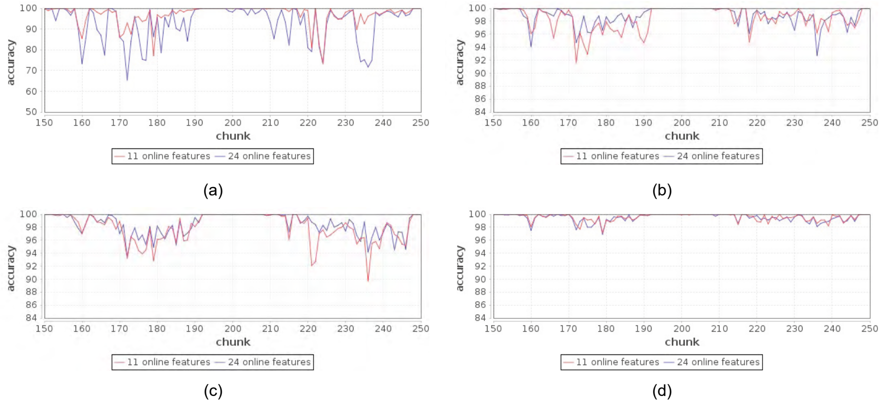
(a) (b)

(c) (d)

(e) (f)

Figure 2. Immediate accuracy comparison a) Naive Bayes b) Hoeffding Tree c) HoeffdingTree Adaptative d) Hoeffding Option Tree e) OzaBag f) OzaBoost

Source: ownelaboration.



(a) (b)

(c) (d)

Figure 3. Immediate accuracy comparison a) OCBoost b) OzaBagAdwin c) Single Classifier Drift d) Ada Hoeffding Option Tree

Source: own elaboration.

To demonstrate that the 24 discriminators of Cambridge24 dataset are at least as useful for traffic classification processes as the Cambridge11 discriminators, which has been used for online traffic classification through various methods obtaining high precision and good performance as described in [6], it is important to verify that the ten incremental learning algorithms described in section 4 keep similar time behavior for both datasets. The above is demonstrated in figures 2 and 3, where the changing points of traffic are evident with up and downs at the same instant but with different depth according to each dataset. Thus, the behavior is similar for the two datasets for each algorithm considering the time.

Incremental Naive Bayes has less accuracy over time and is more susceptible to changes in traffic; however, this algorithm improves accuracy considerably with the Cambridge24 dataset. On the other hand, it is important to highlight that Ada Hoeffding Option Tree has the best accuracy over time and the better results with the Cambridge24 dataset than the Cambridge11 dataset. These two results reflect that the 11 attributes of the Cambridge11 dataset are not adequate to achieve a good accuracy using these two algorithms, while the Cambridge24 dataset contains features that allow algorithms a better performance. On the other hand, if it is not possible to obtain some of the 11 discriminators used in the Cambridge11 dataset in a specific context, the response of these two algorithms would be worse; however, a subset of the Cambridge24 online attributes could surely be found as an alternative, and possibly would give a better response with these algorithms.

According to the results, we can propose these 24 online traffic features to be used as a reference or starting point for any on-the-fly traffic classification work. The Cambridge24 dataset has a similar behavior than Cambridge11 for the learning of incremental algorithms but includes the widest possible set of relevant online features.

4. CONCLUSIONS AND FUTURE WORK

The networking field has greeted machine learning for many purposes. Traffic prediction and classification, resource management, network performance, and intrusion detection are but a few examples where machine learning helps to improve network comprehension and decision making through analysis of the various and large amounts of network data.

In the emerging field of machine learning for networking, the traffic is characterized by different types of discriminators, from well-known port numbers to statistical patterns of the traffic. However, most studies that cover some type of traffic processing ignore the domain knowledge for which traffic processing is performed, for example, if real-time traffic classification is required, or if traffic generating applications wrap their application protocols in others, or other conditions. Consequently, they use traffic discriminators that, in each context, cannot be used, which highlights the importance of an appropriate traffic feature selection according to the context.

We are aware that the selection of traffic features for a specific context of network traffic classification must be hand in hand with domain knowledge. It is also known that the network traffic is massive and always moving; therefore, it should be processed on-the-fly, as far as possible. Considering these two issues, the context for which our work makes the selection of features includes any process of traffic analysis, processing, or classification that requires obtaining results on the fly without the need to wait for the communication flows to end. For this reason, we selected a set of 24 traffic discriminators (attributes of time, flags, size and physical entities) thought for such contexts, and we propose them as a base guideline for future works of online analysis, processing, or classification.

A dataset with these 24 traffic discriminators was subjected to classification processes using several incremental learning algorithms. In these classifications, the algorithms were not pre-trained because of the premise that network traffic is continuous; therefore, these algorithms must adjust quickly to constant changes in the network. The results showed high precision and good behavior over time of the immediate precision measurements.

However, it is important to clarify that the set of discriminators that was taken as the basis for making our feature selection [5] covers the widest possible number of attributes of connection-oriented traffic. So, the selection of online traffic discriminators for non-connection-oriented traffic is a pending issue.

The selected online network traffic discriminators that are the result of this work will save the time of analysis to new investigations with the same approach because it will be known in advance which attributes describe the traffic and can be obtained on the fly. Since it comprises the widest possible set of online attributes, it is expected that any subset of attributes defined from the constraints of specific network traffic offers a good behavior with incremental algorithms. The subset definition can be performed based either on the context of the work, on feature selection algorithms, or according to various parameters established by the project that uses them. They can be used, for example, in real-time traffic analysis processes, because they will only have to focus on getting statistical data from the first quartile of each flow. It can also be adapted to the particular conditions of other processes. For example, for traffic classification projects in Virtual Private Networks surely the parameters related to physical entities (ports) could introduce classification errors, because the same port number will be used for multiple types of traffic, so in advance you can discard the use of 2 of the 24 attributes and only perform the collection of the remaining 22. There may be projects in which the flags are not significant because they will work with non-connection-oriented traffic, so different parameters than the previous example will be discarded. Alternatively, there may also be projects in which their context allows them to use the 24 discriminators and only require a selection of attributes according to the dataset obtained.

In summary, the collection of online traffic features can be performed depending on the tools, protocols, and capabilities of the network elements. For this reason, this paper selected a broad set of optional parameters that could be collected by the online way for on-the-fly traffic classification. The proposed feature set represents a range of options that can be reduced according to the conditions of each network, that is, according to each online classification context.

The attributes of the datasets used in the experimentation were extracted after collecting all the packets from all the traffic, so research should be done on how to collect these proposed features on-the-fly. Furthermore, it would be important to perform a new traffic collection for several types of networks based on these features.

ACKNOWLEDGMENT

This work has been developed thanks to the support of Telematics Engineering Group (GIT in Spanish) of the University of Cauca and Systems Control, Learning and Optimization group (CAOS in Spanish) of the Carlos III University of Madrid, Spain. In addition, the authors are grateful to the Administrative Department of Science, Technology, and Innovation –Colciencias– for funding the Ph. D. program in which this work was developed (Scholarship program No. 647). This work has been also supported by the Spanish Government under project TRA2016-78886-C3-1-R.

REFERENCES

[1] T. Bakhshi and B. Ghita, "On Internet Traffic Classification: A Two-Phased Machine Learning Approach," *J. Comput. Netw. Commun.*, vol. 2016, pp. 21, 2016.

[2] N. Namdev, S. Agrawal, and S. Silkari, "Recent Advancement in Machine Learning Based Internet Traffic Classification," *Procedia Comput. Sci.*, vol. 60, pp. 784-791, Jan. 2015.

[3] T. T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *IEEE Commun. Surv. Tutor.*, vol. 10, 4, pp. 56-76, 2008.

[4] A. Baer *et al.*, "DBStream: A holistic approach to large-scale network traffic monitoring and analysis," *Comput. Netw.*, vol. 107, pp. 5-19, Oct. 2016.

[5] A. Moore, M. Crogan, and D. Zuev, "Discriminators for use in flow-based classification (Technical report No. RR-05-13)," University of London, Department of Computer Science, Queen Mary, 2005.

[6] H. R. Loo and M. N. Marsono, "Online network traffic classification with incremental learning," *Evol. Syst.*, vol. 7, 2, pp. 129-143, Jun. 2016.

[7] F. Ertam and E. Avcı, "A new approach for internet traffic classification: GA-WK-ELM," *Measurement*, vol. 95, pp. 135-142, Jan. 2017.

[8] S. Valenti, D. Rossi, A. Dainotti, A. Pescapè, A. Finamore, and M. Mellia, "Reviewing Traffic Classification," in *Data Traffic Monitoring and Analysis: From Measurement, Classification, and Anomaly Detection to Quality of Experience*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 123-147.

[9] A. Moore, J. Hall, C. Kreibich, E. Harris, and I. Pratt, "Architecture of a Network Monitor," in *Passive & Active Measurement Workshop 2003 (PAM2003)*, 2003.

[10] D. Lei, Y. Xiaochun, and X. Jun, "Optimizing Traffic Classification Using Hybrid Feature Selection," in *2008 The Ninth International Conference on Web-Age Information Management*, Zhangjiajie Hunan, China, 2008, pp. 520-525.

[11] D. Lei, C. You, and Y. Xiaochun, "Optimizing IP Flow Classification Using Feature Selection," in *Eighth International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT 2007)*, Adelaide, SA, Australia, 2007, pp. 39-45.

[12] H. Zhang, G. Lu, M. T. Qassrawi, Y. Zhang, and X. Yu, "Feature selection for optimizing traffic classification," *Comput. Commun.*, vol. 35, 12, pp. 1457-1471, Jul. 2012.

[13] D. Zuev and A. W. Moore, "Traffic Classification Using a Statistical Approach," in *Passive and Active Network Measurement*, 2005, pp. 321-324.

[14] A. W. Moore and D. Zuev, "Internet Traffic Classification Using Bayesian Analysis Techniques," in *Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, New York, NY, USA, 2005, pp. 50–60.

[15] G. P. S. Junior, J. E. B. Maia, R. Holanda, and J. N. de Sousa, "P2P Traffic Identification using Cluster Analysis," in *2007 First International Global Information Infrastructure Symposium*, 2007, pp. 128–133.

[16] T. Auld, A. W. Moore, and S. F. Gull, "Bayesian Neural Networks for Internet Traffic Classification," *IEEE Trans. Neural Netw.*, vol. 18, 1, pp. 223-239, Jan. 2007.

[17] N. Jing, M. Yang, S. Cheng, Q. Dong, and H. Xiong, "An efficient SVM-based method for multi-class network traffic classification," in *30th IEEE International Performance Computing and Communications Conference*, Orlando, FL, 2011, pp. 1-8.

[18] R. Holanda Filho, M. F. Fontenelle do Carmo, J. E. B. Maia, and G. Paulino Siqueira, "An Internet traffic classification methodology based on statistical discriminators," in *NOMS 2008 - 2008 IEEE Network Operations and Management Symposium*, Salvador, Bahia, Brazil, 2008, pp. 907-910.

[19] Y. Liu, H. Liu, H. Zhang, and X. Luan, "The Internet Traffic Classification an Online SVM Approach," in *2008 International Conference on Information Networking*, Busan, South Korea, 2008, pp. 1-5.

[20] F. Noorbehbahani, A. Fanian, R. Mousavi, and H. Hasannejad, "An incremental intrusion detection system using a new semi-supervised stream classification method," *Int. J. Commun. Syst.*, vol. 30, 4, p. e3002, Mar. 2017.

[21] G. Sun, T. Chen, Y. Su, and C. Li, "Internet Traffic Classification Based on Incremental Support Vector Machines," *Mob. Netw. Appl.*, vol. 23, 4, pp. 789-796, Aug. 2018.

[22] G. Baptista and T. Oliveira, "Gamification and serious games: A literature meta-analysis and integrative model," *Computers in Human Behavior*, vol. 92, pp. 306–315, Mar. 2019, doi: 10.1016/j.chb.2018.11.030.

[23] J. Hamari and L. Keronen, "Why do people play games? A meta-analysis," *International Journal of Information Management*, vol. 37, 3, pp. 125–141, Jun. 2017, doi: 10.1016/j.ijinfomgt.2017.01.006.[24] H. A. Jamil, A. Mohammed, A. Hamza, S. M. Nor, and M. N. Marsono, "Selection of On-line Features for Peer-to-Peer Network Traffic Classification," in *Recent Advances in Intelligent Informatics*, 2014, pp. 379-390.

[25] D. C. Corrales, A. Ledezma, and J. C. Corrales, "A Conceptual Framework for Data Quality in Knowledge Discovery Tasks (FDQ-KDT): A Proposal," *J. Comput.*, vol. 10, 6, pp. 396-405, Nov. 2015.

[26] M. Bramer, Principles of Data Mining. Springer, 2016.

[27] M. Juhola and J. Laurikkala, "Missing values: how many can they be to preserve classification reliability?," *Artif. Intell. Rev.*, vol. 40, 3, pp. 231-245, Oct. 2013.

[28] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Springer, 2018.

[29] M. M. Patil, "Handling Concept Drift in Data Streams by Using Drift Detection Methods," in *Data Management, Analytics and Innovation*, Singapore, 2019, pp. 155-166.

[30] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "MOA: Massive Online Analysis," *J. Mach. Learn. Res.*, vol. 11, pp. 1601–1604, 2010.

[31] L. Rutkowski, M. Jaworski, and P. Duda, Stream Data Mining: Algorithms and Their Probabilistic Properties. Springer, 2019.