IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 20, 2019, accepted September 18, 2019, date of publication September 20, 2019, date of current version October 2, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2942782

Consumption Behavior Analysis of Over the Top Services: Incremental Learning or Traditional Methods?

JUAN SEBASTIÁN ROJAS, ÁLVARO RENDÓN, (Senior Member, IEEE), AND JUAN CARLOS CORRALES

Telematics Engineering Research Group, Universidad del Cauca, Popayán 190003, Colombia

Corresponding author: Juan Sebastián Rojas (jsrojas@unicauca.edu.co)

The work of J.S. Rojas was supported by the Administrative Department of Science, Technology, and Innovation of Colombia (COLCIENCIAS) through a National Doctoral Scholarship provided in the Call for National Doctorates under Grant 727-2015 for a total amount of 59,380 USD according to the current exchange rate.

ABSTRACT Network monitoring and analysis of consumption behavior are important aspects for network operators. The information obtained about consumption trends allows to offer new data plans aimed at specific users and obtain an adequate perspective of the network. Over The Top applications are known by their large consumption of network resources. Service degradation is a common mechanism that applies limits to the amount of information that can be transferred and it is usually applied in a generalized way, affecting the performance of applications consumed by users while leaving aside their behavior and preferences. With this in mind, a proposal of personalizing service degradation policies applied to users has been considered through data mining and traditional machine learning. However, such approach is incapable of considering the swift changes a user can present in their consumption behavior over time. In order to observe which approach is capable of a continuous model adaptation while maintaining their usefulness over time, this paper introduces a performance comparison of traditional and incremental machine learning algorithms applied to information about users' Over The Top consumption behavior. Two datasets are implemented for the tests: the first one is built through a real network experiment holding 1,581 instances, and the second one holds 150,000 instances generated in a synthetic way. After analyzing the obtained results, the best algorithm from the traditional approach was a Support Vector Machine while the best classifier from the incremental approach was an ensemble method composed by Oza Bagging and the K-Nearest Neighbor algorithm.

INDEX TERMS Classification algorithms, dataset, incremental learning, machine learning, OTT applications, supervised learning.

I. INTRODUCTION

Over-the-top (OTT) media and communications services and applications are shifting the Internet consumption by increasing the traffic generation over the different available networks. OTT refers to applications that deliver audio, video, and other media over the Internet by leveraging the infrastructure deployed by network operators but without their involvement in the control or distribution of the content [1]. Furthermore, OTT applications are known by their large consumption of network resources in order to offer their different functionalities, so in the mobile networks scope, where traditional operators offer users data plans with a limited consumption, service degradation is a common mechanism implemented in order to apply limits to the amount of information that can be transferred by the users over a period of time [2]–[4]. This mechanism is usually applied when a user exceeds his/her established consumption limit, in order to save resources and ensure the correct performance of the network. Nevertheless, this degradation is applied in a generalized way, i.e., it affects the performance of all the applications that the user can use. Therefore, the behavior and preferences presented by the user in the consumption of OTT applications are not considered and furthermore it is a

The associate editor coordinating the review of this manuscript and approving it for publication was Simone Bianco.

breach of the service level agreements that the ISP (Internet Service Provider) could have established with certain OTT service providers.

Network monitoring and analysis of consumption behavior represents an important aspect for network operators since it allows to obtain vital information about consumption trends in order to offer new data plans aimed at specific users, obtain an adequate vision perspective of the information exchanged inside the network [5], detect potential threats, maintain the quality of the service, and prevent the collapse of networks, among other functionalities. With this in mind, a proposal of personalizing service degradation policies applied to users once their data plan consumption limit is exceeded has been considered by implementing data mining methodologies and traditional machine learning algorithms [6]. Using a testbed implemented to build a dataset that separates 1,581 user profiles into three groups (Low Consumption, Medium Consumption and High Consumption), a classification model capable of classifying users into one of the three identified groups was built through traditional machine learning algorithms. Once the user is classified in one of the groups, a set of personalized policies are proposed for each group following the PCC (Policy and Charging Control) architecture of a LTE network [7]. However, considering the volatile market and fast changes that OTT applications present, such approach which is built in a static manner using traditional machine learning algorithms, is not capable of considering the swift changes of the Internet and, as a consequence, the changes that a user can present in their consumption behavior over time. Therefore, there is still a need, depending on the data, for dynamically adjusting the classification in order to maintain the usefulness of the classification model over time. One way to tackle such need is through the implementation of an incremental learning approach since this kind of algorithms are capable of a continuous model adaptation when facing new data, maintaining their usefulness over time [8]. The use of incremental machine learning algorithms to deal with changes that users may present in their OTT consumption behavior over time, represents an important advantage for network operators not only on network resource administration but in business models as well, since the knowledge of the OTT consumption trends that users have represent an opportunity to glimpse new market strategies and adapt their service offers to the user's needs.

Considering the previous statements, this paper introduces a comparison in terms of performance of traditional and incremental machine learning algorithms applied to a dataset holding information about users' OTT consumption behavior profiles. Section 2 presents the related works found through a literature review following a systematic mapping methodology [9]; Section 3 presents the current architecture implemented to obtain the dataset and a set of personalized service degradation policies using a traditional supervised learning approach; Section 4 presents a detailed description of the dataset structure implemented for the training and testing of the algorithms; Section 5 introduces the synthetic data generator developed to create the data streams and the proposed test scenarios; Section 6 illustrates and analyzes the obtained results from the test scenarios; and Section 7 shows the conclusions and future works.

II. RELATED WORKS

This proposal is supported by the concept of KDN (Knowledge Defined Networking) defined by Mestres *et al.* [10], where the application of Artificial Intelligence (AI) techniques to control and operate networks is considered due to the continuous technological advances. Another fundamental concept is incremental learning, that refers to machine learning methods with the ability of continuous model adaptation based on a constantly arriving data stream, usually present whenever systems need to act autonomously (e.g., autonomous robotics or driving) [8], [11]. With this in mind, the current state of the art will be presented after implementing a systematic mapping of academic documents, based on the methodology presented in [9], to provide an overview of the research area and determine the amount and type of related works.

The selected topics of interest were: Service degradation, with the objective of identifying how this resource control mechanism is managed in the networks by Internet Service Providers (ISP); Quality of Service (QoS), focusing on identifying the parameters closely related to the service degradation; OTT services, in order to know how this topic has been developed in the research field highlighting the fact that it is a very recent investigation area; Categorization of users in a mobile network, in order to know how operators manage users inside the network; Traffic classification, focusing on identifying which techniques are used for this process and specially in identifying how incremental learning has been implemented within the network traffic classification scope.

Following the systematic mapping methodology the following related works are highlighted: In [12], Agababov *et al.* presents a proxy service for HTTP connections, that aims at extending the life time of users' data plans in mobile networks, reducing the size of the packages exchanged between the servers and user equipment; Flywheel integrates with the Google Chrome browser and on average reduces by 50% the consumption in the data plans generated by browsing and loading of web pages. In [2] and [3], Chetty *et al.* presented results from a qualitative study of households living with bandwidth caps and the design and implementation of a tool, called ''uCap'', to help home users manage Internet data. In the white paper [4], the US Company Ixia presents an analysis of the QoS policies (dynamic allocation of network resources, priority control, limitation of traffic rates) that are usually implemented in a LTE (Long Term Evolution) mobile network; furthermore, they present the QoS parameters that significantly affect the performance of the different services offered in the network (video, voice, gaming, internet, etc.) and highlight the importance of categorizing users for a mobile operator in order to efficiently manage network resources. In [13], Yang *et al.* present an analysis of the behavior of users in the consumption of mobile internet within a 3G network; through the analysis, they propose three types of characterization for the users: Consumption of the data plan, which focuses on the rate of consumption generated by the user and classifies them into two profiles: normal user and heavy user; Mobility pattern, where they focus on identifying the movement patterns of users within the network; and Application consumption, that focuses on what type of applications are mostly used through the mobile internet link. In [14], G. Sun *et al.* present a preliminary proposal of an incremental Support Vector Machine (SVM) method that is applied to address two issues of current SVM's: the inability to support continuous learning and the fact that this kind of algorithm has high requirements on both memory and CPU; experimental results show that the incremental Support Vector Machine method decreases the training time, while still sustains the high accuracy of traffic classification. In [15], considering that, in order to classify network traffic in today's dynamic environment, data stream mining algorithms have been introduced to overcome the shortcoming of conventional data mining algorithms, H. R. Loo *et al.* presented an online classification method which is aimed for online network traffic classification, by applying an incremental k-means where the classification model can learn from unlabeled and labeled data becoming an incremental semi-supervised learning approach. In [16], having in mind that classification accuracy of supervised approaches is significantly affected if the size of the training set is small, and that a model built using a static training set will not be able to adapt to the non-static nature of Internet traffic, Divakaran et al. developed the concept of "self-learning" to deal with these two challenges; specifically, this paper designs and develops a new classifier called Self-Learning Intelligent Classifier (SLIC); SLIC starts with a small number of training instances, self-learns and rebuilds the classification model dynamically, with the aim of achieving high accuracy in classifying non-static traffic flows.

Considering that this paper is focused on incremental learning and its application to network traffic classification, Figure 1 illustrates the map of 39 papers obtained through the systematic mapping methodology [9] which are distributed through different categories. The vertical axis shows the different research contexts defined for the state of the art investigation, while the horizontal axis presents the research types defined by the systematic mapping methodology. From this map it is possible to remark: there are no works that apply incremental learning in service degradation; most of discarded papers are related to traffic mobility and object and people recognition; there are 8 papers that apply incremental learning to network traffic classification [14]–[21]; there are no papers that describe and share new datasets; and finally there are 2 papers [22], [23] related to service degradation that provide an analysis of the consequences provoked by the application of this resource control mechanism.

After analyzing the previous related works it is important to mention the following conclusions: until now, to the best of our knowledge, there have not been papers proposed by other authors that consider a similar approach to the one presented in this paper; in general, the works that are related to service degradation look for ways that avoid having to implement this mechanism on the user and do not consider a dynamic personalization scheme based on the user's consumption behavior; most of the works related to OTT services focus on the study of business models to favor the ISPs and mobile operators when applying this type of strategy without considering studies on consumption trends and user categorization; although there are works that consider the implementation of both traditional machine learning (supervised and unsupervised learning) and incremental learning in the context of traffic classification, none of those works have considered to leverage such tools to perform an analysis in the user's OTT consumption behavior nor the development of a dynamic classification model for the application of personalized service degradation policies.

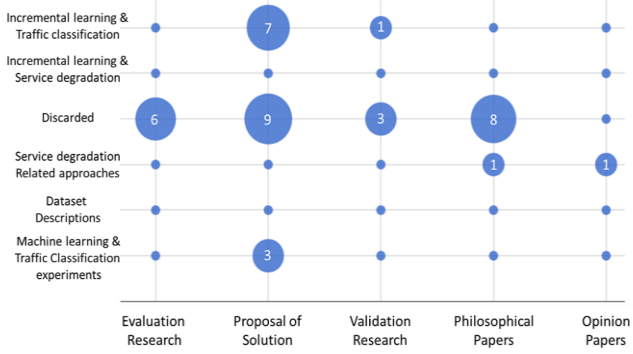

FIGURE 1. Systematic map incremental learning.

III. CURRENT ARCHITECTURE

This section presents the architecture that was implemented in order to gather and preprocess the data that was captured inside the campus of Universidad del Cauca (Unicauca) in order to analyze the OTT consumption behavior presented by the users of the network. It is important to mention that, by following the KDN concept [10], the architecture aims at including a knowledge plane to the network by gathering information relevant to the OTT consumption behavior of users and help the network administrators in the definition process of the personalized service degradation policies. Figure 2 presents the implemented architecture; the previous work performed for the policies definition can be found in [6].

Figure 2 presents the following components:

- Application Plane: comprehends the users that consume the OTT applications within the Unicauca network through different kinds of devices including smartphones, laptops, among others.
- Data & Control Plane: comprehends all the network devices, its topology, the exchanged data and the current functional configuration that is set up by the network administrator.

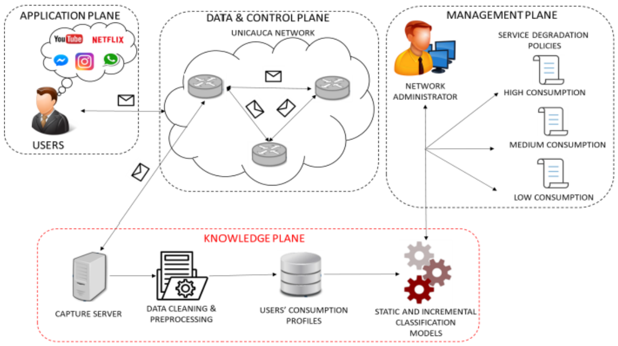

FIGURE 2. **Current architecture.**

- Knowledge Plane: represents the elements that deliver valuable knowledge and insight about the network to the network administrator, aiming at helping in the definition of personalized service degradation policies. The main elements are: Capture Server, that was set up within one of the network cores of the Unicauca network in order to capture the traffic containing the information related to the OTT consumption behavior presented by the users; such information was captured using Wireshark, during 6 days in 2017, performing capture sessions of 30 minutes. Data Cleaning & Preprocessing, where the raw data, captured in pcap files with Wireshark, are processed, analyzed and cleaned following the CRISP-DM methodology [24] and the conceptual framework for data quality proposed by Corrales *et al.* [25], in order to obtain a summary and insight about the users' OTT consumption behavior; specifically, the software tools implemented to carry out the data preprocessing were CICFlowmeter [26], ntopng [27], R programming language and Weka [28]; a further explanation of this process can be consulted in [6]. Users' Consumption Profiles dataset, that holds the OTT consumption behavior of 1,581 users within the network; a deeper explanation of this dataset will be given in the next section. Finally, Static and dynamic classification models, which are machine learning algorithms used for the classification of users into their corresponding consumption profile; this element allows to identify which users present a specific consumption profile so the network administrator can define the personalized service degradation policies for each consumption group; it is important to mention that this paper has as its main objective the performance comparison of both traditional and incremental approaches when classifying the users' consumption behavior.
- Management Plane: comprehends all the decisions that the network administrator can make in order to define the service degradation policies. As can be observed, after analyzing the users' behavior, three consumption profiles (high, medium and low consumption) were identified. Therefore, the network administrator can define different service degradation policies while being supported by the knowledge plane in the decision making process.

IV. DATASET DESCRIPTION

The dataset presented in this paper is a subsequent result (third version) of the preprocessing performed on its first two versions that are described in [6] and are available for download in [29] and [30]. This dataset contains 1,581 instances and 131 attributes on a single file. Each instance represents a user's consumption profile which holds summarized information about the consumption behavior of the user, related to the 29 OTT applications identified in the different IP flows captured in order to create the dataset.

The OTT applications that the users interacted with during the capture experiment and were stored in the dataset are: Amazon, Apple store, Apple Icloud, Apple Itunes, Deezer, Dropbox, EasyTaxi, Ebay, Facebook, Gmail, Google suite, Google Maps, Browsing (HTTP, HTTP_Connect, HTTP_Download, HTTP_Proxy), Instagram, LastFM, Microsoft One Drive (MS_One_Drive), Facebook Messenger (MSN), Netflix, Skype, Spotify, Teamspeak, Teamviewer, Twitch, Twitter, Waze, Whatsapp, Wikipedia, Yahoo and Youtube. Each application has 4 different types of attributes (quantity of generated flows, mean duration of the flows, average size of the packets exchanged on the flows and the mean bytes per second on the flows). These attributes summarize the interaction that the user had with the respective OTT application in terms of consumption. Furthermore, the dataset contains the user's IP address in network and decimal format which are used as user identifiers. Finally, the User Group attribute represents the objective class in which a user is classified considering his/her OTT consumption behavior; there are 643 users for the high consumption profile, 475 users for the medium consumption profile and 463 users for the low consumption profile [6]; when analyzing the classes distribution it can be observed that although mostly all the users access the same applications, they vary in the intensity of their consumption; the high consumption users consume the higher number of applications (focused specially on 14 applications): Amazon, Apple, Browsing, Dropbox, Facebook, Gmail, Google, MSN, Skype, Twitter, Whatsapp, Wikipedia, Yahoo and YouTube; the medium consumption users mostly consume the following 13 applications: Amazon, Apple, Browsing, Dropbox, Ebay, Facebook, Gmail, Google, MSN, Skype, Twitter, Yahoo, YouTube; finally the low consumption users exhibit the behavior with the least consumption intensity in time and quantity of applications, consuming mostly 5 applications: Amazon, Browsing, Facebook, Google and YouTube. All of this information gives a total of 131 attributes. Table 1 describes each attribute in detail.

V. EXPERIMENT DESIGN

With the aim of comparing traditional and incremental machine learning algorithms applied to the dataset previously described, which holds information about the user's

TABLE 1. Attributes description.

Attribute name	Attribute description
Source.Decimal	This attribute holds the user's IP address in decimal format and it is mainly used as a user identifier.
Source.IP	This attribute holds the user's IP address in network format (e.g., 192.168.14.35) and as in the previous case its main function is to work as a user identifier.
Application-Name.Flows	This group of attributes hold the information about the quantity of IP flows that a user generated toward an OTT application. As was mentioned before each application has a group of 4 attributes that describe the interaction of the user with a specific OTT application (an example for this case would be Netflix.Flows or Facebook.Flows).
Application-Name.Flow.Duration.Mean	This group of attributes hold the information related to the mean duration (time) of the flows generated by the user towards a specific OTT application, measured in seconds. Examples of how this attributes are stored in the dataset are: Amazon.Flow.Duration.Mean or Instagram.Flow.Duration.Mean
Application-Name.AVG.Packet.Size	This group of attributes hold the average size of the IP packets that were exchanged in all the flows generated by the user towards a specific OTT application, measured in bytes. It is important to notice that this size is focused on the packet's header only. Examples of how this attribute are presented on the dataset are: Google_Maps.AVG.Packet.Size or Spotify.AVG.Packet.Size
Application-Name.Flow.Bytes.Per.Sec	This group of attributes hold the mean number of bytes per second that were exchanged in the flows generated by the user towards a specific OTT application. Examples of this kind of attributes in the dataset are: Deezer.Flow.Bytes.Per.Sec or Skype.Flow.Bytes.Per.Sec.
User.Group	This group of attribute represents the objective class of the dataset i.e., the different groups that the users are classified in according to their OTT consumption behavior. Those groups are: High consumption (643 instances), Medium consumption (463 instances) and Low consumption (475 instances).

OTT consumption behavior, two steps were done, explained in the following subsections: first, a synthetic data generator was developed for generating multiple streams of data needed to test and train the incremental learning algorithms, and second, three test scenarios where defined for comparing in terms of performance 3 traditional machine learning algorithms and 9 incremental learning algorithms.

A. SYNTHETIC DATA GENERATOR

This subsection presents the synthetic data generator that was developed in order to generate the multiple data streams needed for the tests. This considers that incremental learning refers to online learning strategies and such approach works differently from the traditional batch mode for the inference of the classification model, where all the instances are stored up to a specific time step in memory; rather, incremental learning has to rely on a compact representation of the already observed signals, such as an efficient statistics of the data, an alternative compact memory model, or an implicit data representation in terms of the model parameters itself. At the same time, it has to provide accurate results for all relevant settings, despite its limited memory resources [8].

Besides, by analyzing the dataset previously presented, it is clear that the number of instances stored (1,581 instances) are not sufficient to perform incremental learning tests. Additionally, the process needed to be carried out in order to generate more instances from raw data of IP flows would require an important effort in terms of time, aiming at capturing enough data. Therefore, the development of a synthetic data generator of users' OTT consumption profiles based on the statistical distribution obtained on the current dataset is the most adequate option. With this in mind, by assuming that all the attributes in the dataset were mutually independent and by leveraging the features of the R programming language, specifically the "fitdist" function [31] that enables to fit univariate theoretical distributions (normal distribution, beta distribution, gamma distribution, etc.) to non-censored data using different estimation methods, it was possible to infer a statistical distribution for each attribute in order to generate new data that presented the same statistical behavior observed on the original dataset.

The first step in order to perform the statistical estimation was to separate all the instances from each objective class (high, medium and low consumption) since the statistical estimation had to analyzed the distribution of all the attributes on each type of consumption profile individually. After separating the instances of each class it was decided that the estimation should be performed on 130 attributes (removing the objective class). Each attribute distribution was analyzed through a Cullen and Frey graph where the kurtosis and square of the skewness [32] of the data distribution are compared to the same measures of different known theoretical distributions (normal distribution, gamma distribution, etc.) in order to determine which distribution fits best to the behavior of the data. As an example of the process, Figure 3 illustrates the Cullen and Frey graph obtained for the attribute Twitter.Flows of the instances classified as high consumption users; however it is important to mention that this process was carried out for all the 29 OTT applications

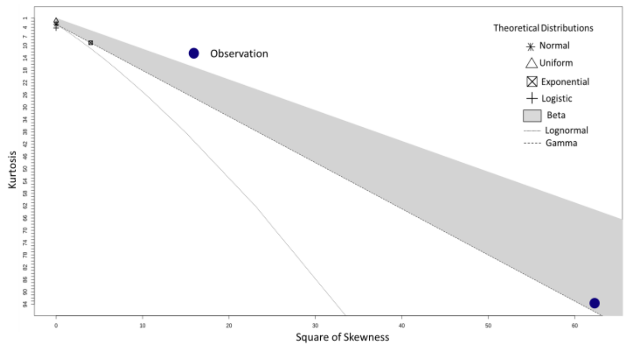

FIGURE 3. Cullen and Frey Graph - Twitter.Flows attribute - High consumption profile.

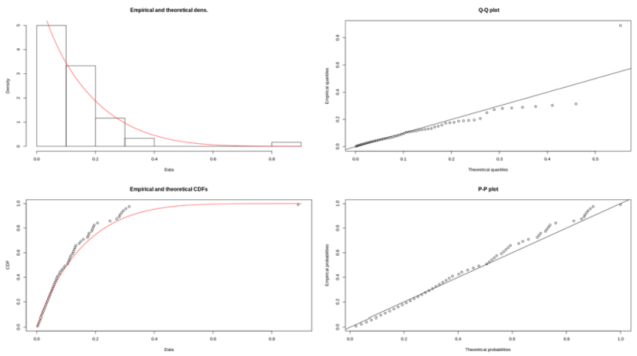

FIGURE 4. CDF comparison - Twitter.Flows attribute - High consumption profile.

stored on the dataset. The blue dot illustrates the intersection of the square of the skewness and the kurtosis obtained from the attribute data, and the different lines and areas represent a possible statistical distribution that fits the data. In this specific case the data can be fit into a beta or gamma distribution. Therefore, in order to decide which of these two distributions is the better fit, a one-sample Kolmogorov-Smirnov test [33], [34] is performed obtaining two values: The first one is the maximum distance between the empirical CDF (Cumulative Distribution Function) obtained from the data and the theoretical CDF from the beta and gamma distributions in this specific case. The second one is the p-value, that allows to accept the null hypothesis (the data shows a distribution similar to a specific theoretical distribution) if it is higher than 0.05. As can be observed on Table 2, by comparing the results obtained with the Kolmogorov-Smirnov test, the distribution that fits best to the data from the Twitter.Flows attribute is the beta distribution since the maximum distance between the CDF's is closer to zero and the p-value is higher than 0.05 and is higher than the p-value from the gamma distribution.

TABLE 2. Kolmogorov - Smirnov test - Twitter.Flows attribute - High consumption profile.

Theoretical Distributions	Maximum distance between CDF's	P-value
Beta Distribution	0.0039431	0.4185
Gamma Distribution	0.0051021	0.148

Once the best theoretical distribution is identified, a comparison between the empirical and theoretical CDF's is performed, fitting the theoretical statistical distribution to the attribute data distribution using maximum likelihood estimation.

Such comparison is illustrated in Figure 4 through 4 plots: the empirical and theoretical density, the Quantile-Quantile plot (Q-Q plot), the empirical and theoretical CDF's and the Probability-Probability plot (P-P plot). The red line represents the behavior of the CDF from the beta distribution. It can be observed that this distribution is a good fit since the data exhibits a similar behavior. Afterwards, the variables that describe the distribution (α and β for this case since it is a beta distribution) are calculated and the synthetic data are generated for the attribute using the R programming language. This process is repeated for all the other attributes from the dataset creating 3 data generators, one per consumption profile (high, medium and low consumption).

Once the three data generators were finished, a synthetic dataset holding 150.000 instances (50.000 for each class – high, medium and low consumption) was created and the instances were shuffled in order to proceed with the test scenarios described in the following subsection.

It is important to mention that the data generation process was satisfactory since it allows to obtain several data streams for further experimentation based on statistical distributions from the consumption behavior of real users inside a network; however, the data generators can be improved if the generation is performed without the assumption of statistical independence between attributes, since such assumption could provoke inconsistencies among the generated data.

B. TEST SCENARIOS

This subsection describes the three test scenarios that were defined in order to compare the performance of traditional machine learning algorithms with incremental learning algorithms. This considers that users can change their consumption behavior over time, that the market that involves the OTT applications is highly volatile and that, although the KDN paradigm [10] proposed the introduction of AI techniques such as machine learning to network management, it does not consider the need of machine learning models that maintain their usefulness over time. Furthermore, it is important to mention that, from all the algorithms presented in [6], the best 3 traditional algorithms in terms of performance were selected for these tests. On the other hand, a set of 9 incremental learning algorithms were also selected for the test, trying to find a suitable classification model from this approach.

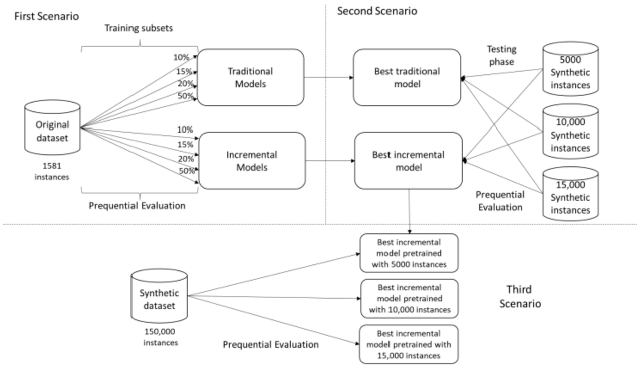

FIGURE 5. **Test scenarios.**

Figure 5 illustrates the different test scenarios. In order to understand the experimental process, it is important to mention the following remarks:

- The traditional models were trained and tested using a percentage split [35] configuration on the first scenario. Therefore different subsets of the original dataset were used to train the models and the remaining subsets were used to test their performance (the training and testing datasets are different).
- From the second scenario onwards, only the performance of the best traditional model obtained from the first scenario was tested (i.e., no additional training phases were performed). Hence, the model received datasets apart from the one used in the training phase (the synthetic dataset), in order to be tested.
- The incremental models were tested using a prequential evaluation [36] or interleaved test-then-train configuration. Such evaluation is an alternative to the traditional holdout evaluation, inherited from batch setting problems. This method consists of using each sample to test the model, which means to make predictions, and then the same sample is used to train the model. This way the model is always tested on samples that it hasn't seen yet.

The tests scenarios will be described as follows:

1) FIRST TEST SCENARIO

On this scenario, the aim was to identify if the incremental learning models were capable of obtaining a similar or better performance with a fewer quantity of instances on the training phase, by comparing them with the traditional models using the original dataset with 1,581 instances. The training phase of the models was carried out using different sizes of the dataset. Specifically, 10%, 15%, 20% and 50%. The traditional learning algorithms implemented were: Adaboost with J48 decision tree as base classifier, KNN with 30 neighbors (such number of neighbors was identified using a cross validation approach) and the SMO (Sequential Minimal Optimization) algorithm. All the tests performed for the traditional algorithms were performed using Weka. On the other hand, the incremental learning algorithms that were trained and tested are: Hoeffding tree, Naive Bayes, Adaptive Random Forest (ARF), Leverage Bagging using KNN with 8 neighbors as base classifier, Leverage Bagging using ARF as base classifier, Perceptron mask (neural network), Oza Bagging using KNN with 8 neighbors as base classifier, Oza bagging using a Hoeffding tree as base classifier, and KNN with 5 neighbors. The tests using the incremental learning algorithms were performed using Scikit Multiflow, a framework for learning from data streams and multi-output learning in Python [37].

2) SECOND TEST SCENARIO

On this scenario, the aim was to compare the performance behavior of the best models of each approach, obtained from the first test scenario, with a subset of instances taken from the synthetic dataset. Such analysis was carried out by observing two aspects: first of all, if the best incremental model was able to maintain a good performance while adapting to the new data. Second, if the best traditional model was able to exhibit a good performance against the new data using only the knowledge obtained from the training of the first scenario. To perform these tests three subsets of 5,000, 10,000 and 15,000 instances were taken from the synthetic dataset and used individually to evaluate both models. Figure 5 illustrates the setting established for the second scenario.

3) THIRD TEST SCENARIO

On this last scenario, the aim was to observe how the best incremental learning model, obtained from the first scenario, would behave in terms of performance when receiving all of the instances from the synthetic dataset (150,000 instances) having a ''warm-up'' or pretrain with the different subsets from the second scenario. All of this in order to observe if the incremental model reaches a maximum point in terms of performance or when receiving a high number of instances its performance begins to be affected in a negative way. Figure 5 illustrates the proposed configuration for this scenario.

VI. RESULTS ANALYSIS

This section presents the results and analysis obtained on each of the test scenarios focusing on comparing the precision, recall, kappa statistic and confusion matrix (in certain cases) in order to determine which model achieves better performance.

A. RESULTS – FIRST SCENARIO

As was mentioned before, on this scenario the training was performed with different subsets taken from the original dataset, and a total of 12 algorithms were tested (3 traditional algorithms and 9 incremental algorithms).

1) TRAINING WITH 10% AND 15%

Figure 6 present the obtained results for the traditional and incremental models respectively, with a training phase using 10% and 15% of the original dataset. In general, all three traditional algorithms exhibit a good performance, with SMO being the best by a small difference.

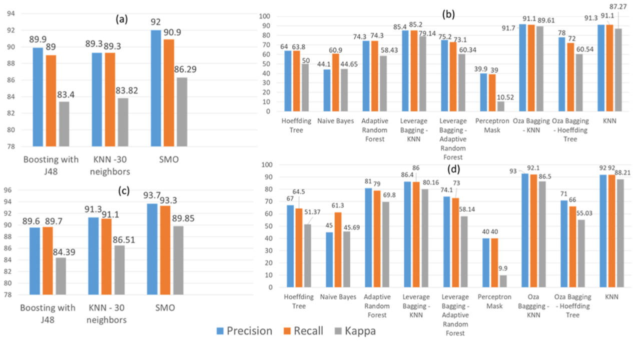

FIGURE 6. Results first scenario: (a) Traditional models 10% - (b) Incremental models 10% (c) Traditional models 15% - (d) Incremental models 15%.

2) TRAINING WITH 20% AND 50%

Figure 7 present the obtained results for the traditional and incremental models respectively, with a training phase using 20% and 50% of the original dataset.

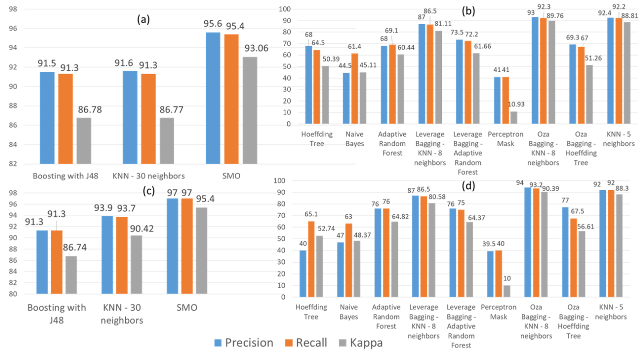

FIGURE 7. Results first scenario: (a) Traditional models 20% - (b) Incremental models 20% (c) Traditional models 50% - (d) Incremental models 50%.

After analyzing all the results obtained from the first scenario it can be observed that all the traditional algorithms (Boosting with J48, KNN with 30 neighbors and SMO) exhibit a good overall performance, highlighting the fact that SMO is the best with all the training subsets. On the other hand, the incremental algorithms present a more volatile behavior, where, unexpectedly, the Perceptron Mask shows the worst behavior on all cases and the KNN algorithm and the composition of Oza Bagging with KNN are the best incremental classifiers on all cases. Furthermore, it is important to mention that for this scenario, where the training and testing is done using the same dataset, as expected, the traditional models keep improving their performance when the training set has more instances, with the SMO algorithm having the best overall performance when the training set holds 50% of the instances. However, the Oza Bagging with KNN and the KNN algorithm also exhibit a really similar performance to the SMO algorithm on all the cases. Having this in mind, the model selected as the best classifier from the incremental learning approach on this scenario is the composition of Oza Bagging with KNN trained with the subset holding 10% instances from the original dataset, since the difference with the other cases is not significant in terms of performance and this case required less computational resources in order to be tested.

It is important to mention that, for the time being, with the results observed on this first scenario it can be concluded that, while using the same dataset for training and testing, there is no major difference in terms of performance between some algorithms from both approaches. However, the incremental models hold an important advantage for a network operator considering that their training and testing is performed in a more efficient way than the traditional models (requiring less time and computational resources) [8].

B. RESULTS – SECOND SCENARIO

As was presented in a previous section, this scenario aims at comparing the performance behavior of the best algorithms from each learning approach obtained on the first scenario, using subsets from the generated synthetic dataset. Figure 8 illustrates the obtained results and Table 3 and Table 4 shows the confusion matrix for both algorithms in the case tested with 15,000 instances. The following conclusions are stated:

- The incremental model (Oza Bagging with KNN) not only was able to maintain a good performance with new (synthetic) data, but shows a better performance with an important difference in all cases (more than 40% in all of the performance metrics). This is because the model is capable of analyzing the incoming data streams, keep learning on new data and adapt itself.

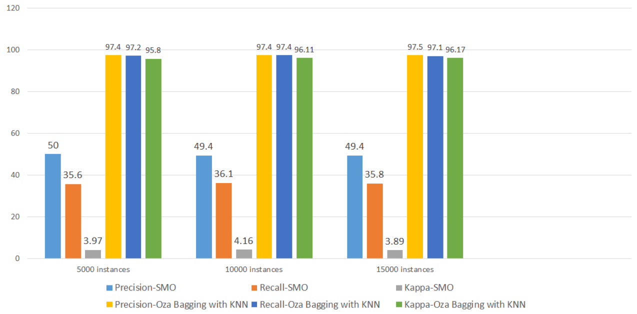

FIGURE 8. Performance results - second scenario.

TABLE 3. Confusion matrix - SMO trained with 50% - 15000 instances.

SMO 50% - 15000 inst			
-	High Consumption	Medium Consumption	Low Consumption
High Consumption	4932	2	0
Medium Consumption	5037	24	0
Low Consumption	260	4338	407

TABLE 4. **Confusion matrix - Oza Bagging with KNN trained with 10% - 15000 instances.**

Oza Bagging - KNN - 8 neighbors - 15000 inst			
-	High Consumption	Medium Consumption	Low Consumption
High Consumption	4672	261	1
Medium Consumption	36	4944	81
Low Consumption	3	1	5001

- The traditional model (SMO) is able to identify only two classes out of the three that exists on the dataset. Such behavior can be happening considering that the differences between the users classified on the medium and high consumption profiles are very subtle and the knowledge acquired by the traditional model in the training phase is insufficient for the classification task. Therefore, since it is not learning anymore, it is incapable of separating these instances and ends up gathering them on a single group. This can be observed in the confusion matrix of the traditional model where most of the medium consumption instances are being classified as high consumption and most of the instances from the low consumption class are being classified as medium consumption.
- The traditional model (SMO) was not able to exhibit a good overall performance when facing the new (synthetic) data. This is because the new data has meaningful changes and the knowledge obtained by the traditional model in the training phases is not sufficient to properly process them. The synthetic data was generated using the same statistical distribution as in the real data used in the training phase, assuming statistical independence; however, this independence is not entirely true, and therefore this assumption introduced changes in the dataset that the traditional model was not able to handle with its current knowledge. It is worth mentioning that in order to obtain results with real data, it is still required to implement a dataset reflecting changes in the user's OTT consumption behavior over time.

The previous statements allow us to conclude that, when we consider the volatility of the Internet and OTT applications, the incremental learning approach is a suitable option when dealing with possible changes that users may present in their OTT consumption behavior over time i.e., since the attributes that define the user consumption profile (Number of generated flows, time of consumption and required network resources) can present multiple changes on a time period, the incremental learning approach represents an important advantage for network administrators, since it overcomes the weakness that a traditional model presents about their incapability of adapting to new data without a new training process.

Furthermore, by having knowledge of the OTT consumption trends that users have, network operators can glimpse new strategies to offer a better and personalized quality of service.

C. RESULTS – THIRD SCENARIO

As previously stated, the aim for this scenario was to observe if the best incremental model reaches a maximum point in terms of performance or is affected in a negative manner, when receiving a high number of instances obtained from the synthetic dataset (150,000 instances). This is considered since an usual inconvenience of the adaptability of incremental learning models is that they might "forget" important information about the data [8] when adapting to incoming data streams, decreasing their performance in the classification. Figure 9 and Figure 10 illustrates the obtained results for this scenario.

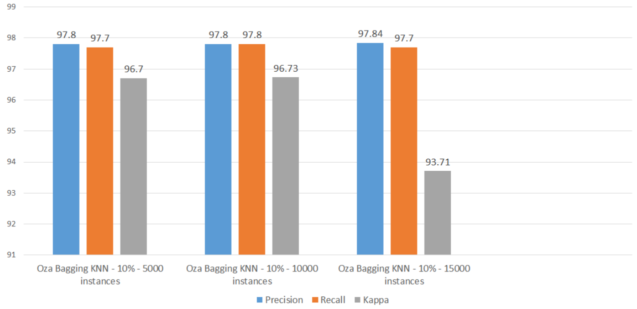

FIGURE 9. **Performance results - third scenario.**

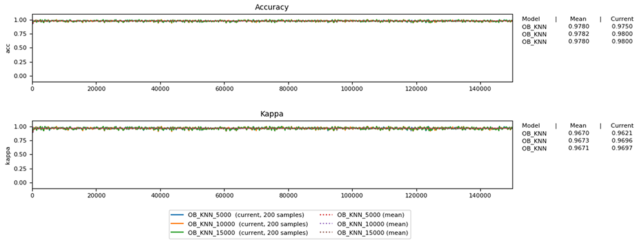

FIGURE 10. **Performance results - third scenario - learning evolution.**

By analyzing the obtained results two conclusions are clear: first, the number of instances used as "warm-up" for the model are irrelevant since the performance is almost identical in all cases. Second, the incremental model maintains a good overall performance after being tested with 150,000 synthetic instances without showing any negative decline in the accuracy or kappa statistic.

Therefore, it is possible to conclude that the composition of Oza Bagging with KNN may be a suitable incremental learning approach when dealing with the possible changes that users may present in their OTT consumption behavior over time, without being affected by the quantity of data streams it can receive. This represents an important advantage to network administrators since the classification model maintains its utility over time, saving efforts and resources on its renovation process. Furthermore, the incremental approach surpasses the overall performance of traditional models when the training and testing datasets are different and shows a similar performance when the training and testing sets are the same.

VII. CONCLUSION AND FUTURE WORK

Having in mind the volatile market and fast changes that the Internet and specially the OTT applications exhibit, this paper presented a comparison in terms of performance of traditional and incremental machine learning algorithms with the aim of personalizing, in a dynamic way, the service degradation policies that network operators apply to users once their data plan consumption limit is exceeded. Such comparison was considered since a traditional classification model is not capable of considering the swift changes of the Internet and, as a consequence, the changes that a user can present in their OTT consumption behavior over time. In order to perform the comparison, a subsequent version of the dataset presented in [6] holding information of the OTT consumption behavior from different users inside the campus of Universidad del Cauca was obtained. However, since the number of instances stored on the dataset were insufficient, a synthetic data generator, using R programming language and statistical analysis tools, was developed with the aim of generating numerous data streams needed for the testing of the incremental learning algorithms.

In order to compare both learning approaches three test scenarios were defined. From the first scenario it can be concluded that, while using the same dataset for training and testing, there is no major difference in terms of performance between some algorithms from both approaches. However, the incremental models hold an important advantage for a network operator considering that their training and testing is performed in a more efficient way than the traditional models (requiring less time and computational resources).

From the second scenario it can be concluded that the incremental learning models exhibit a better performance due to their capacity of adapting to the new incoming data, while the traditional models identified only two classes out of the three that exists on the dataset. This can be happening considering that the differences between the users classified on the medium and high consumption profiles are very subtle, since such differences lie especially on the quantity of flows generated per OTT application, and the knowledge acquired by the traditional model in the training phase is insufficient for the classification task. Hence it ends up gathering them on a single group, demonstrating that a traditional model is incapable of adapting to new data without a new training phase.

From the third scenario it can be concluded that the number of instances used as ''warm-up'' for the model are irrelevant since the performance is almost identical in all cases, and that the incremental model maintains a good overall performance without showing any negative decline. This represents an important advantage to network administrators since the classification model maintains its utility over time, saving efforts and resources on its renovation process.

The previous statements allows us to conclude that, when we consider the volatility of the Internet and OTT applications, the incremental learning approach may be a suitable option when dealing with the possible changes that users may present in their OTT consumption behavior over time, and this represents an important advantage for network administrators since such approach overcomes the weakness that a traditional model presents about their incapability of adapting to new data without a new training process. Furthermore, by having knowledge of the OTT consumption trends that users have, network operators can glimpse new strategies to offer a better and personalized quality of service.

As future works it is proposed to perform a comparison of both learning approaches using a new real dataset captured from users inside the campus of Universidad del Cauca. Also, the development of an application that enables the use of the best classification model to classify users according to their consumption behavior, and to perform the study and implementation of a mechanism that enables the enforcement of personalized service degradation policies inside the architecture of a real network. Finally, an exploration of developing a synthetic data generator that does not assume statistical independence between attributes, in order to obtain data that are closer to a real network scenario.

ACKNOWLEDGMENT

The authors would like to thank Universidad del Cauca for supporting this research.

REFERENCES

[1] T. Sudtasan and H. Mitomo, ''Effects of OTT services on consumer's willingness to pay for optical fiber broadband connection in Thailand,'' in *Proc. 27th Eur. Regional Conf. Int. Telecommun. Soc.*, 2016, pp. 1–11.

[2] M. Chetty, R. Banks, A. J. Brush, J. Donner, and R. Grinter, ''You're capped: Understanding the effects of bandwidth caps on broadband use in the home,'' in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2012, pp. 3021–3030.

[3] M. Chetty, H. Kim, S. Sundaresan, S. Burnett, N. Feamster, and W. K. Edwards, ''uCap: An Internet data management tool for the home,'' in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.*, 2015, pp. 3093–3102.

[4] *Quality of Service (QoS) and Policy Management in Mobile Data Networks | Ixia*. Accessed: Dec. 1, 2016. [Online]. Available: https://www.ixiacom.com/resources/quality-service-qos-and-policy-management-mobile-data-networks

[5] V. Carela-Español, ''Network traffic classification?: From theory to practice,'' Ph.D. dissertation, Dept. d'Arquitectura Computadors, Univ. Politècnica Catalunya, Barcelona, Spain, 2014.

[6] J. S. Rojas, Á. R. Gallón, and J. C. Corrales, ''Personalized service degradation policies on OTT applications based on the consumption behavior of users,'' in *Computational Science and Its Applications—ICCSA*. Cham, Switzerland: Springer, 2018, pp. 543–557.

[7] ITU. *ETSI TS 23.203: Policy and Charging Control Architecture*. Accessed: Dec. 7, 2017. [Online]. Available: http://www.itu.int/itu-t/workprog/wp_a5_out.aspx?isn=6084

[8] A. Gepperth and B. Hammer, ''Incremental learning algorithms and applications,'' presented at the Eur. Symp. Artif. Neural Netw. (ESANN), 2016.

[9] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, ''Systematic mapping studies in software engineering,'' in *Proc. 12th Int. Conf. Eval. Assessment Softw. Eng.*, Swindon, U.K., 2008, pp. 68–77.

[10] A. Mestres, A. Rodriguez-Natal, J. Carner, P. Barlet-Ros, E. Alarcón, M. Solé, V. Muntés-Mulero, D. Meyer, S. Barkai, M. J. Hibbett, G. Estrada, K. Ma'ruf, F. Coras, V. Ermagan, H. Latapie, C. Cassar, J. Evans, F. Maino, J. Walrand, and A. Cabellos, "Knowledge-defined networking," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 47, no. 3, pp. 2–10, Sep. 2017.

[11] D. L. Silver, "Machine lifelong learning: Challenges and benefits for artificial general intelligence," in *Artificial General Intelligence*. Berlin, Germany: Springer, 2011, pp. 370–375.

[12] V. Agababov, M. Buettner, V. Chudnovsky, M. Cogan, B. Greenstein, S. McDaniel, M. Piatek, C. Scott, M. Welsh, and B. Yin, "Flywheel: Google's data compression proxy for the mobile Web," in *Proc. 12th USENIX Symp. Netw. Syst. Design Implement. (NSDI)*, 2015, pp. 367–380.

[13] J. Yang, Y. Qiao, X. Zhang, H. He, F. Liu, and G. Cheng, "Characterizing user behavior in mobile internet," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 1, pp. 95–106, Mar. 2015.

[14] G. Sun, S. Li, T. Chen, Y. Su, and F. Lang, "Traffic classification based on incremental learning method," in *Advanced Hybrid Information Processing*. Cham, Switzerland: Springer, 2018, pp. 341–348.

[15] H. R. Loo and M. N. Marsono, "Online network traffic classification with incremental learning," *Evolving Syst.*, vol. 7, no. 2, pp. 129–143, Jun. 2016.

[16] D. M. Divakaran, L. Su, Y. S. Liau, and V. L. L. Thing, "SLIC: Self-learning intelligent classifier for network traffic," *Comput. Netw.*, vol. 91, pp. 283–297, Nov. 2015.

[17] G. Sun, T. Chen, Y. Su, and C. Li, "Internet traffic classification based on incremental support vector machines," *Mobile Netw. Appl.*, vol. 23, no. 4, pp. 789–796, Aug. 2018.

[18] C. Swartz and A. Joshi, "Identification in encrypted wireless networks using supervised learning," in *Proc. IEEE Mil. Commun. Conf.*, Oct. 2014, pp. 210–215.

[19] P. Li, Y. Wang, and X. Tao, "A semi-supervised network traffic classification method based on incremental learning," in *Proc. Int. Conf. Inf. Technol.*, 2013, pp. 955–964.

[20] S. Lal, P. Kulkarni, U. Singh, and A. Singh, "An efficient approach for network traffic classification," in *Proc. IEEE Int. Conf. Comput. Intell. Comput. Res.*, Dec. 2013, pp. 1–5.

[21] S. Chen, K. Zeng, and P. Mohapatra, "Efficient data capturing for network forensics in cognitive radio networks," *IEEE/ACM Trans. Netw.*, vol. 22, no. 6, pp. 1988–2000, Dec. 2014.

[22] D. Karamshuk, N. Sastry, A. Secker, and J. Chandaria, "On factors affecting the usage and adoption of a nation-wide TV streaming service," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr./May 2015, pp. 837–845.

[23] *T-Mobile's Binge on Violates Key Net Neutrality Principles*. Accessed: May 20, 2019. [Online]. Available: http://cyberlaw.stanford.edu/blog/2016/01/t-mobiles-binge-violates-key-net-neutrality-principles

[24] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0: Step-by-step data mining guide," NCR Syst. Eng. Copenhagen, Copenhagen, Denmark, DaimlerChrysler AG, Stuttgart, Germany, SPSS Inc., Chicago, IL, USA, OHRA Verzekeringen en Bank Groep B.V, Arnhem, The Netherlands, Tech. Rep. Volume 16, 2000.

[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, no. 6, pp. 396–405, Nov. 2015.

[26] *NetFlowMeter*. Accessed: May 20, 2019. [Online]. Available: http://www.netflowmeter.ca/

[27] *Ntopng Source Code Repository*. Ntop, Pisa, Italy, 2018.

[28] *Weka 3—Data Mining With Open Source Machine Learning Software in Java*. Accessed: Apr. 3, 2018. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/

[29] *IP Network Traffic Flows Labeled With 75 Apps*. Accessed: Apr. 15, 2019. [Online]. Available: https://kaggle.com/jsrojas/ip-network-traffic-flows-labeled-with-87-apps

[30] *OTT Consumption Profile—Unicauca Dataset*. Accessed: Apr. 15, 2019. [Online]. Available: https://kaggle.com/jsrojas/ott-consumption-profile-dataset

[31] *Fitdist Function | R Documentation*. Accessed: Apr. 15, 2019. [Online]. Available: https://www.rdocumentation.org/packages/fitdistrplus/versions/0.2-1/topics/fitdist

[32] *Measures of Shape: Skewness and Kurtosis*. Accessed: Apr. 15, 2019. [Online]. Available: https://brownmath.com/stat/shape.htm

[33] ResearchGate. *Evaluating Kolmogorov's Distribution*. Accessed: Apr. 15, 2019. [Online]. Available: https://www.researchgate.net/publication/5142829_Evaluating_Kolmogorov%27s_Distribution

[34] W. W. Daniel, *Applied Nonparametric Statistics*. Boston, MA, USA: PWS-Kent Publishing, 1990.

[35] *Data Splitting. Z. Reitermanov. Introduction. Cross-Validation Techniques*. Accessed: Jul. 2, 2019. [Online]. Available: https://docplayer.net/26609777-Data-splitting-z-reitermanova-introduction-cross-validation-techniques.html.

[36] A. Bifet, G. de Francisci Morales, J. Read, G. Holmes, and B. Pfahringer, "Efficient online evaluation of big data stream classifiers," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, 2015, pp. 59–68.

[37] J. Montiel, J. Read, A. Bifet, and T. Abdessalem, "Scikit-multiflow: A multi-output streaming framework," *J. Mach. Learn. Res.*, vol. 19, no. 72, pp. 1–5, 2018.

JUAN SEBASTIÁN ROJAS received the bachelor's degree in electronics and telecommunications engineering and the M.Sc. degree in telematics engineering from Universidad Del Cauca, Colombia, in 2015 and 2018, respectively, where he is currently pursuing the Ph.D. degree in telematics engineering. Furthermore, he has been an Assistant Professor in several courses related to converged services, data mining, and machine learning. He is also an Active Researcher with the Telematics Engineering Group, Universidad Del Cauca, Colombia. His main research topic involves the study of users' Over The Top consumption behavior aimed at the personalization of service degradation through incremental learning supported by the Knowledge Defined Networking Paradigm. In 2016, he received the Ph.D. Scholarship by Colciencias, the Leader Entity of the Science, Technology and Innovation System in Colombia.

ÁLVARO RENDÓN (M'81–SM'11) received the B.S. degree in electronics engineering and the M.S. degree in telematics engineering from the University of Cauca, Colombia, in 1979 and 1989, respectively, and the Ph.D. degree in telecommunications engineering from the Technical University of Madrid, in 1997.

He is currently a Full Professor with the Department of Telematics, University of Cauca, where he is also the Director of the Doctoral Program in telematics engineering. His research interests include telecommunications services, e-health, and S&T management.

Dr. Rendón was a recipient of the Outstanding Paper Award in the category of Tools Track at the First IEEE International Conference on Engineering of Complex Computer Systems, in 1995.

JUAN CARLOS CORRALES received the bachelor's and master's degrees in telematics engineering from the University of Cauca, Colombia, in 1999 and 2004, respectively, and the Ph.D. degree in sciences, specialty in computer science, from the University of Versailles Saint-Quentin-en-Yvelines, France, in 2008. He is currently a full time Professor with the University of Cauca, where he leads the Telematics Engineering Group. His research interests include service composition and data analysis.