IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p- ISSN: 2278-8735.Volume 14, Issue 4, Ser. I (Jul.-Aug. 2019), PP 69-77
www.iosrjournals.org

Transfer Function Analysis Using Machine Learning with Data Mining Application

Edson Brito Junior
(Technical Institute/ Federal University of Pará/Brazil)

Abstract: *With the popularization of the global computer network, the Internet, information in the digital format more affordable. Documents, image files, audio and video travel at all times in the network. New applications appear in the network, among them stand out the conferences, downloading video, games etc. Inside from this context the need to achieve higher bit rates than those provided by the traditional connection so that this service is offered with excellent quality to the subscriber, it is of fundamental interest for the operators to know the state of the telephone link of this subscriber. In order to qualify the link of the subscriber, it is proposed to implement a methodology through the extraction of knowledge, obtained from the information stored in a database composed of the measurements of the link transfer function of the subscriber, and analyze the results obtained for some test cases, validate the methodology and application potential and its extensibility to other types of data available.*

Keywords: *Local link, knowledge extraction*

Date of Submission: 28-07-2019 Date of acceptance: 13-08-2019

I. Introduction

With the constant emergence of new applications on the Internet, and with the great evolution of processing computer data, it becomes increasingly difficult to connect to the Internet through the traditional connection dialed. And among the solutions found for access to the Internet at high speed, stands out the DSL technology (Digital Subscriber Line) that uses the existing telephone infrastructure.

An important factor to be observed in DSL technology is that the number of bits in each service is a function of the conditions of the subscriber's link. The better the physical state of the link, and especially without limitations in all its length, more secure and effective will be the service delivered to the end user.

However, in order for this quality in the delivery of services to the subscriber to be achieved, several tests to evaluate how the topology of the link is found and what the decisions should be when requested to the implementation of these services. In addition, when a service is already installed, it is often necessary to maintain and fix some problems encountered in the subscriber's link.

One way to evaluate the current state of the link is to know the basic parameter known as the Transfer with this the operators can arrive at a more precise result on the current condition of the link location of the subscriber. This rating is known as Loop Qualification. Faced with this there are many possibilities of using the information stored in the database to extract knowledge and decision-making regarding the improvement and increase of the bit rate delivered to the subscriber.

Analyzing this growing amount of information is not a trivial task and requires the use of advanced computational techniques to discover hidden and potentially useful patterns between the data stored in the operators' databases. One way to work through all this information is to use the Knowledge Discovery Process through the Data Mining stage for construction of a knowledge model that assists in the tasks of qualification and monitoring of the local link of the subscriber.

II. Fundamentals

2.1 The Local Subscriber Link

The local loop infrastructure of the subscriber, known as the link plan, consists of pair cables (CO - Central Office) to a subscriber (CP - Customer Premise). The telephone link twisted pair is called the local link of the subscriber. The subscriber's link, because it has several variable lengths and often greater than 3 km, consists of one or more sections with different diameters (JUNIOR, 2007). Figure 1.0 shows a representation of a link between the and the subscriber, where the presence of a local link is observed with many sections, with diameters and different lengths.

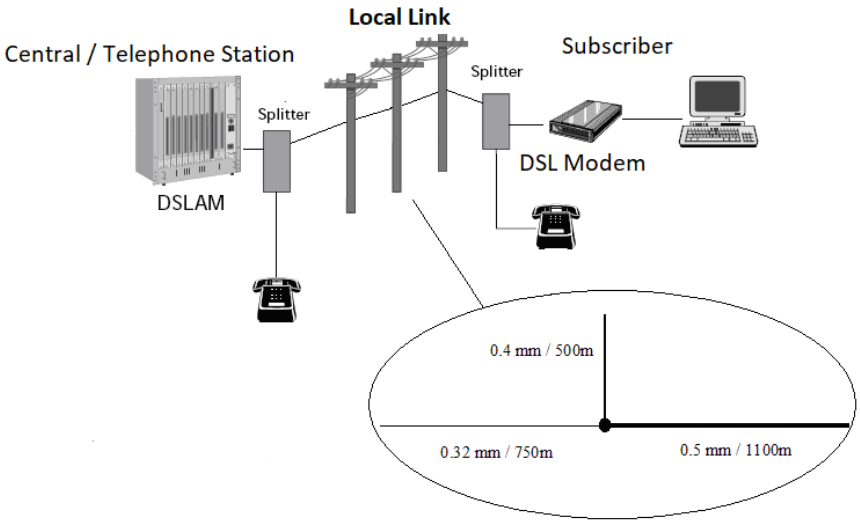

Figure 1.0: Local link between the exchange and the subscriber.

2.2 Transfer Functions

In general, the transfer function of a system can be defined as an expression describing their transfer characteristics, that is to say, the relation between the input and the output of system in terms of transfer characteristics. Transfer function is a dependent quantity of the frequency at which it is attenuated with increase of the same as shown in Figure 2.0. This is a of the great interaction between the signals propagated in the material composing the transmission medium. The function of transfer is an important parameter used to set the maximum local link length of the subscriber.

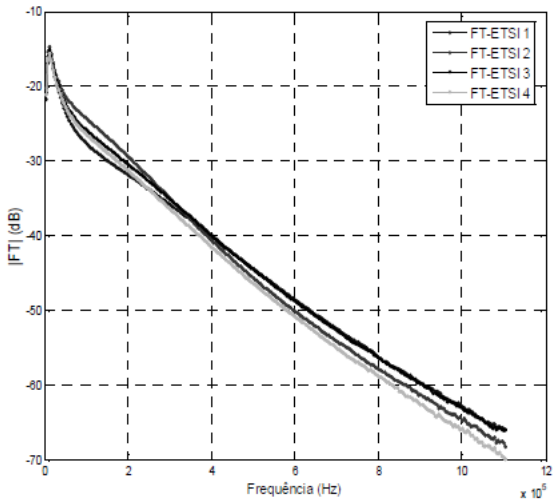

Figure 2.0: Transfer function curve of some subscriber local links.

2.3 Qualification of Local Subscriber Link

When a subscriber requests a certain DSL service from the operator, the first thing they should do is check if the local link of this subscriber can support this service. This is called Link Qualification and it differs for each subscriber since it depends on the topology between the subscriber and the operator (JUNIOR, 2007).

The qualification of the link is a set of techniques, which aim to evaluate the capabilities of a specific link to support and maintain the requirements of a DSL service. Among these techniques, the following stand out: physical layer and simulations by means of mathematical models, that identify a certain link. With the help of the Link Qualification it is possible through the measurements made from the tests made by the operators, the construction of a database, where the stored information serves to comparisons with other databases to provide a tool to monitor and diagnose performance of the DSL service.

2.4 Data Mining

Data Mining is a set of computational techniques for extracting unknown and potentially useful information in large volumes of data through a compact themselves. The term "data mining" is one of the steps in a larger process known as Knowledge Discovery in Databases (KDD) (FAYYAD et al., 2019), which provides the infrastructure necessary for data mining, including the steps required to build a consistent, reduced, and reliable database for the discovery of the desired information. This process is also known as knowledge extraction, data archeology, or information gathering.

2.5 The Knowledge Discovery Process

The data mining process or KDD basically consists of six phases and each phase can interact with the others. In this way, the results produced in a phase can be used to improve the results of the next phases. This scenario indicates that the KDD process is iterative, always seeking to improve the results at each iteration. If the result obtained in the last step is not satisfactory, one must return to step convenient, making the process cyclical. Figure 3.0 illustrates the whole process.

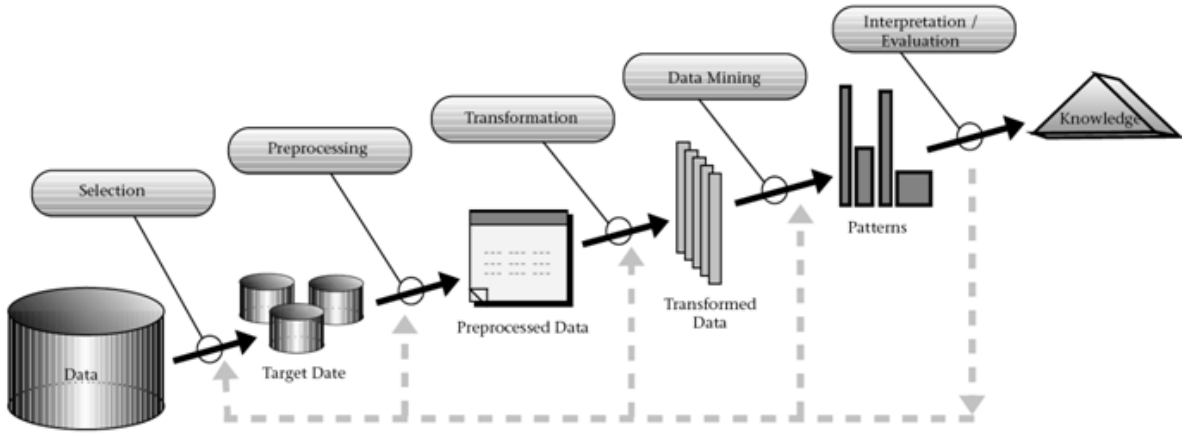

Figure 3.0: Stages of the knowledge discovery cycle (FAYYAD et al., 2019).

2.6 Data Mining Functionalities

The types of patterns that can be discovered depend on the functionalities (or tasks) employed in data mining. There are two main types of features or goals in data mining: descriptive data mining, describing the existing characteristics of the data, and data mining predictive model, which attempts to predict attribute values based on the inference of available data. Classification is a of the key features of data mining. Also known as supervised classification, uses a particular labeled class to sort the objects in a data collection. Usually uses a training set where all objects are associated with known classes. The algorithm of Classification learns from the training set and builds a model. The template is used to sort new objects.

2.7 Classification Model

One of the most important and most popular KDD tasks is the task of sorting. Informally, as shown in Figure 4.0, this task can be understood as searching for a function that allows correctly associate each Xi record of a database with a single categorical label, Yj, called class (GOLDSCHMIDT et al., 2015). Once identified, this function can be applied to new form to predict the class in which such records fit. In order to formalize the classification task, we consider an ordered pair of the form (x, f (x)), where x is an n-dimensional input vector and f (x) is the output of a function f. The function h is called a hypothesis or f.

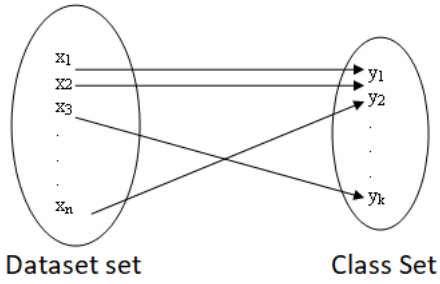

Figure 4.0: Associations between data records and classes (GOLDSCHMIDT et al., 2005).

In cases where the image of f is formed by class labels, the task of inductive inference is called classification and the whole hypothesis h is called the classifier (GOLDSCHMIDT, 2015). The identification of the function h consists of a search process in the space of hypotheses H, by the function that most closely approximates the original function f. This process is called learning (RUSSELL et al, 1995). Any algorithm that can be used in the execution of the learning process is called the learning algorithm. The set of all the hypotheses that can be obtained from a learning algorithm L is represented by HL. Each hypothesis belonging to HL is represented by hL.

Once a hypothesis (classifier) is induced this may be very specific for the training set used. If this set is not sufficiently representative, the classifier can perform well in the training set, but not in the test set. It is said, in this case, that the classifier has to the training set, occurring a phenomenon called overfitting (GOLDSCHMIDT et al., 2015). On the other hand, when the classifier adjusts very little to the training set, it is said to occur an underfitting (GOLDSCHMIDT et al., 2015). This phenomenon usually occurs due to parametrizations learning algorithm.

The completeness of a classifier refers to its ability to classify (present a response) to all the examples in the database. Consistency, on the other hand, indicates the ability of the classifier to correctly sort the available samples in the database.

2.8 Methods of Data Mining

Data Mining methods require different pre-processing needs (MORIK, 2000). These needs vary according to the extensional aspect of the database in which the will be used. Due to the great diversity of data preprocessing methods, many possible combinations of methods. The choice of these alternatives may influence the quality of the results of the KDD process (MORIK, 2000). Although there are many methods used in Data Mining, only the Instances and Induction Decision Trees are described below due to data type and preprocessing performed in the treatment phase of the data used in the work developed.

2.8.1 Instance-Based Method

The expression "instance-based method" indicates that the method, when processing a new record, takes into account existing instances or records in the database. One of the main methods of Mining is called K-NN (K-Nearest Neighbors Next). The K-NN method is widely used in applications involving classification tasks. It is a a method that is easy to understand and implement and does not require prior training to be applied (GOLDSCHMIDT, 2015).

2.8.2 Methods Based on Induction of Decision Trees

Some of the main methods of Data Mining are based on the construction of decision trees from the databases. In general the construction of a decision tree is performed according to some recursive approach to partitioning the database (GOLDSCHMIDT, 2015). A decision tree is a model of knowledge in which each internal node of the tree represents a decision on an attribute that determines how the data is partitioned by its child nodes. Initially, the root of the tree contains the entire database with mixed examples of various classes. A predicate, called the point of separation, is chosen as the condition that best separates or discriminates classes. Such a predicate involves exactly one of the attributes of the problem and divides the database into two or more sets, which are each associated a child node. Each new node therefore comprises a database partition that is recursively that set associated to each leaf node consist entirely or predominantly of records of the same (GOLDSCHMIDT, 2015).

III. Methodology for Qualification of Local Subscriber Link Based on Mining of Data

The development of a methodology for the construction of classification models based on data mining is useful and necessary for link qualification activities. The implementation of the used in this work consists of almost all stages of the knowledge process, as shown in the diagram in Figure 5.0.

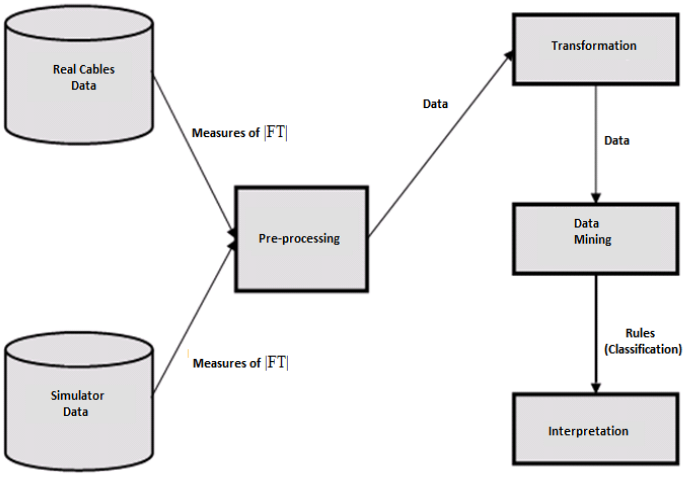

Figure 5.0: Implementation Diagram of the Methodology.

3.1 Collection and Selection of Data

The collected data were obtained through measurements of the transfer function of a set of local links of the subscriber as shown in Figure 6.0, composed of real cables and line simulators. At the Measurement campaign was aimed at adopting a procedure for all measurements taken, and In addition, the same experiment was repeated three times for the same measurement, thereby ensuring that the data be analyzed and compared in order to arrive at a reliability of the method used.

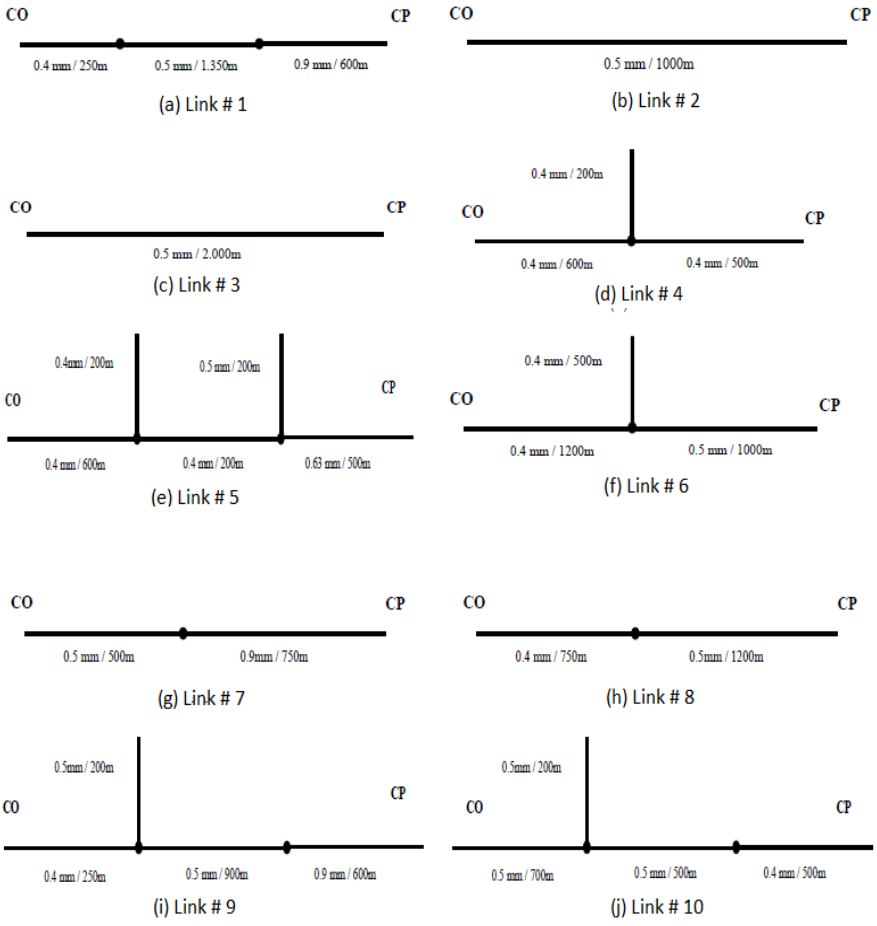

Figure 6.0: Subscriber Local Link Sets.

3.2 - Pre-Processing

Another point taken into consideration was the pre-processing of the data. The pre-processing steps made during the implementation of the methodology were:

3.2.1 - Data Statistics Analysis

Most measurements performed in this work, however optimized their is not exempt from causing errors when in use. Therefore, it is necessary to statistical treatment of the values of the transfer function of the local subscriber links measured, with the systematic and / or random errors in measurement processes.

3.2.2 Outliers Detection Test

In order for the collected data to be stored it is necessary to carry out a statistical analysis to accuracy. During measurements, situations may occur that some samples and this may influence the performance of the classifier model and present results that do not correspond to reality. The data (samples) that do not have characteristics similar to the rest of the are called outliers (JUNIOR, 2008). There are some tests that help the analysis to detect samples that diverged from the population. However, in this study, the Dixon test (JUNIOR, 2008) was used, due to this presents a good accuracy and speed in the detection of outliers values. The choice of this test was due to its ease and because it is a simple and effective method. In addition, the determining factor in your choice is: possibility of being applied in a considerable quantity of samples. Another advantage is that it is not the a priori knowledge of the estimation of these values is necessary.

3.3 - Transformation

The original data obtained through the measurements of the transfer function were worked out, that they can easily be used in the construction of classifier models. Table 1 shows the database of the transfer function of each subscriber's local link, composed of the frequency tones which characterize a given link of the subscriber according to Figure 6.0. The obtained database consists of 730 instances and 257 attributes (including class) of numeric type.

Table 1: Transfer Function Measurement Database.

Attributes					Class
Tones					Link
Ton_1	Ton_2	Ton_3	...	Ton_256	1L
Ton_1	Ton_2	Ton_3	...	Ton_256	2L
Ton_1	Ton_2	Ton_3	...	Ton_256	3L
...	...	...	...	...	...
Ton_1	Ton_2	Ton_3	...	Ton_256	10L

3.4 - Data Mining and Interpretation

The Data Mining step consists of the construction of the classifier. In order to train, validate and test a classifier was used the data mining package Waikato Environment for Knowledge Analysis - WEKA (WEKA, 2019). WEKA is an open package written in Java and easy to manipulate by the user that integrates various Artificial Intelligence algorithms applicable to data mining. For the construction of the classifier, the following sequence:

1 - The database is divided into two sets: training sets (training data with 634 instances) and test set (validation data or sample validation 96 instances).

2 - Training (or learning): construction of a model (classifier) analyzing the samples of the set training.

3 - Validation of the model: application of the model on the validation set. The percentage of accuracy of the classes predicted by the model in relation to the expected (or unknown) classes. This percentage is called the precision of the model for the validation set in question. Tables 2 and 3 show the percentage of classes correctly classified by the model using the decision tree algorithms J.48 and K-NN as much for the training phase and validation of the model.

Table 2: Results of using the training set (634 instances) using WEKA learning algorithms.

Classifier	Tree J.48	K-NN
Correctly Classified Intents	97.0032%	93.5135%
Incorrectly Classified Intents	2.9968%	6.4865%

Table 3: Results of using the validation set (96 instances) using learning algorithms WEKA.

Classifier	Tree J.48	K-NN
Correctly Classified Intents	96.875%	90%
Incorrectly Classified Intents	3.125%	10%

For the test phase of the classifier, a set of data with 90 instances was used. of the transfer function of the scenarios of Figure 6.0 with values independent of the training set and validation and test results using J.48 and K-NN are shown in Table 4.

Table 4: Results of using the test set (90 instances) using WEKA learning algorithms.

Classifier	Tree J.48	K-NN
Correctly Classified Intents	95.7143%	94.2857%
Incorrectly Classified Intents	4.2857%	5.7143%

According to the results obtained, the decision tree J.48 showed to have a better performance in relation to the K-NN algorithm. However, the results obtained, in general, compared to those of the literature, were the expected ones. To evaluate the performance of the J.48 classifier model when predicting an unknown link, we used a set of data considering that the class of each scenario is not known a priori. For this, the "?" symbol was used instead of the class attribute as shown in Table 5.

Table 5: Database of Unknown Scenarios.

Attributes					Class
Tones					Link
Ton_1	Ton_2	Ton_3	...	Ton_256	?
Ton_1	Ton_2	Ton_3	...	Ton_256	?
Ton_1	Ton_2	Ton_3	...	Ton_256	?
...	...	...	...	...	...
Ton_1	Ton_2	Ton_3	...	Ton_256	?

3.5 - Results for the Classifier Test

The following are the results obtained with the classifier test for the links unknown. Figures 7 to 10 show the graphs of the transfer function curves between bond unknown and predicted (known) by the classifier along with the errors between such curves.

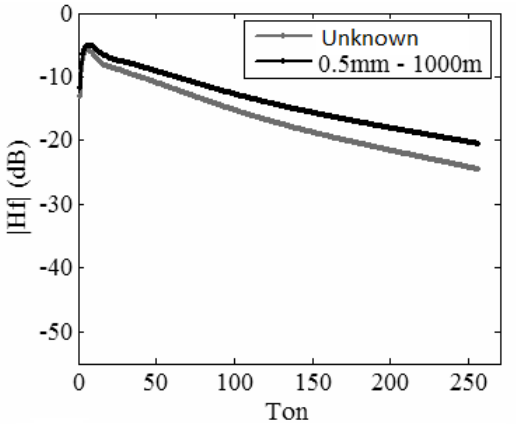

Figure 7: (a) Graph of the Transfer Function Curve of the Unknown link versus known link.

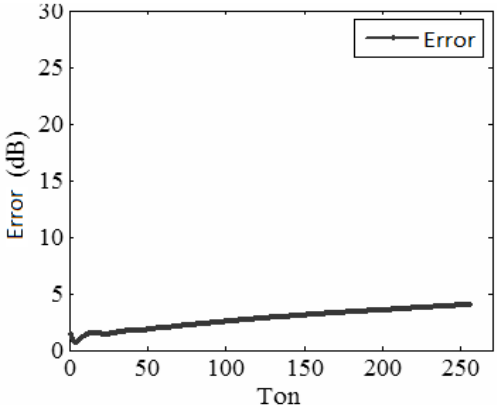

Figure 7: (b) Error graph found between curves.

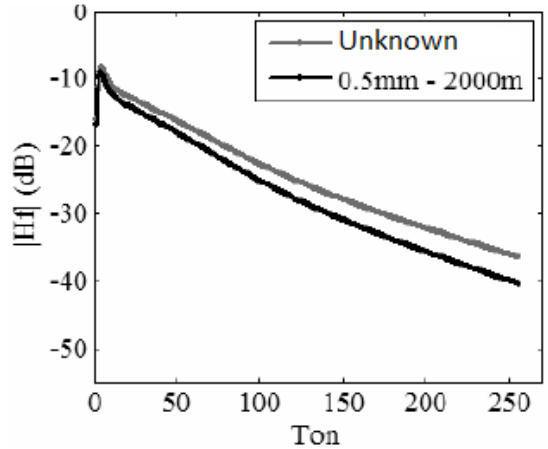

Figure 8: (a) Graph of the Transfer Function Curve

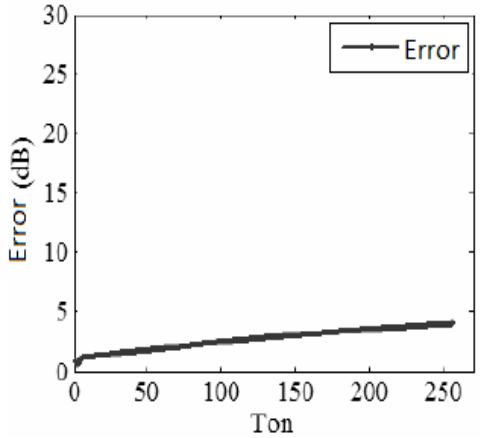

Figure 8: (b) Error Graph found between curves.

of the Unknown link versus known link.

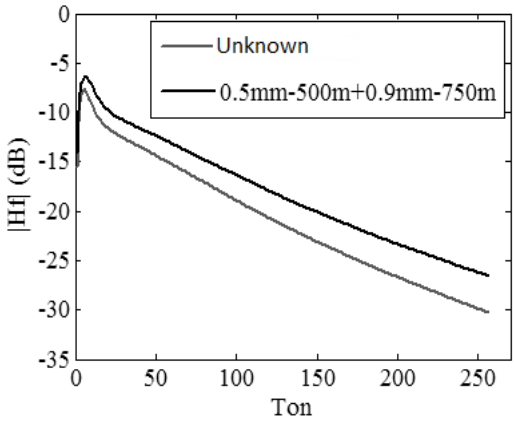

Figure 9: (a) Graph of the Transfer Function Curve of the Unknown link versus known link

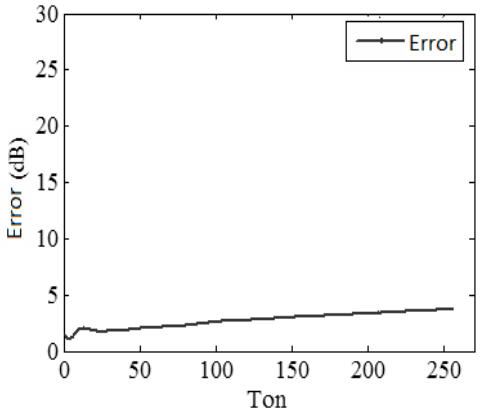

Figure 9: (b) Error graph found between curves.

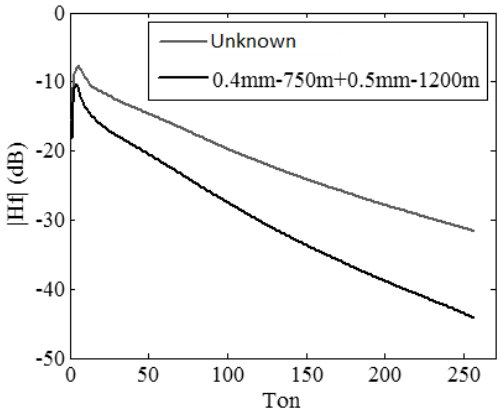

Figure 10: (a) Graph of the Transfer Function Curve of unknown link versus known link.

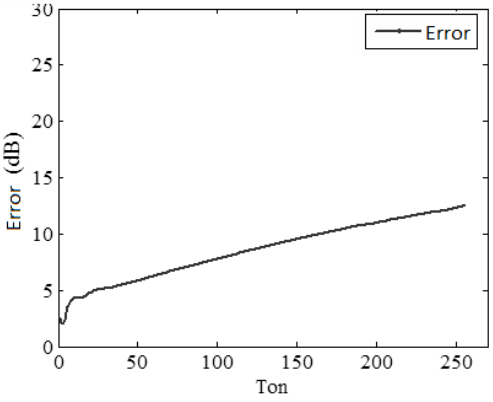

Figure 10: (b) Error Graph found between curves

3.6 - Results Analysis for the Classifier Test

For simple links (1 and 2 sections) the model (classifier) performed well when estimating links of the same gauge and approximate length for the set of tests presented. For links with approximately equal diameters or lengths, the error between the link curves predicted and unknown is less than 5 dB considering all 256 tones (frequency range). However, in the Figure 10 shows that the error is greater than 5 dB. predict an unknown class. To further validate the applicability of data mining, in the qualification of the local link of the subscriber, it is proposed to perform tests with more complex links: with three or more sections and with presence of derivations.

IV. Final Considerations

In this work the data mining was used as a method that assists in the qualification of the link location of the subscriber. In this context, it is necessary to know the infrastructure of the local that it can carry a quality service to the subscriber. In relation to the qualification of the it is crucial to measure the transfer function in which it allows a diagnosis of the operational state of the network. The applicability of data mining was verified through the construction of classifier models, which when used properly, provide reliable results for decision making by operators of DSL services.

As a suggestion of future work, it is intended to make a methodology for derivation detection present on the subscriber's link. Since the measurements of the physical parameters of the subscriber's link accurate in their results they can be used for the detection of the binary classification since the information contained in the database gives no indication a priori about a possible topology. Since one of the important information for the operators is to know if the link location of the subscriber has or does not have leads.

References

[1]. JUNIOR, E. B. Methodology for the Measurement of Parameters Relating to Digital Link Qualification of the Subscriber. Master's Dissertation, 2007.
[2]. FAYYAD, U. M .; PIATETSKY-SHAPIRO, G .; SMYTH, P. From Data Mining to Knowledge Discovery: An Overview. Knowledge Discovery and Data Mining, Menlo Park: AAAI Press, 2019.
[3]. WEKA (visited in June, 2019). http://www.cs.waikato.ac.nz/ml/weka.
[4]. GOLDSCHMIDT, R .; E. STEPS, E. Data Mining a Practical Guide. Rio de Janeiro, Elsevier, 2015.
[5]. RUSSEL, S .; NORVIG, P. Artificial Intelligence: A Modern Approach. New Jersey: Prentice-Hall, 1995.
[6]. MORIK, K. The Representation Race - Preprocessing for Handling Time Phenomena. Proceedings of the European Conference on Machine Learning 2000, Lecture Notes in Artificial Intelligence 1810. Berlin: Springer Verlag, 2000.

Edson Brito Junior. " Transfer Function Analysis Using Machine Learning with Data Mining Application" IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) 14.4 (2019): 69-77.