Towards Evaluating Quality of Datasets for Network Traffic Domain

Dominik Soukup
CTU in Prague
Thakurova 9, Prague, Czech Republic
soukudom@fit.cvut.cz

Peter Tisovčík
Brno University of Technology
Božetěchova 2, Brno, Czech Republic
itisovcik@fit.vutbr.cz

Karel Hynek Tomáš Čejka
CESNET
Zikova 4 Prague, Czech Republic
hynekkar@cesnet.cz cejkat@cesnet.cz

***Abstract*—This paper deals with the quality of network traffic datasets created to train and validate machine learning classification and detection methods. Naturally, there is a long epoch of research targeted at data quality; however, it is focused mainly on data consistency, validity, precision, and other metrics, which are insufficient for network traffic use-cases. The rise of Machine learning usage in network monitoring applications requires a new methodology for evaluation datasets. There is a need to evaluate and compare traffic samples captured at different conditions and decide the usability of the already captured and annotated data. This paper aims to explain a use case of dataset creation, propose definitions regarding the quality of the network traffic datasets, and finally, describe a framework for datasets analysis.**

***Index Terms*—Dataset, Data Quality, Evaluation, Network traffic analysis.**

I. Introduction

Machine learning (ML) techniques have been known for a long time, and they are used in many areas of the real world. Recently, machine learning becomes very popular in network traffic monitoring and analysis as well. With the rise of privacy-preserving and secured technologies, ML seems to be the only feasible way to deal with such traffic.

However, the performance of the ML models highly depends on the provided *high-quality* datasets. It is not rare that an ML-based classifier works well in a laboratory environment; however, when deployed to production, its accuracy drastically decreases due to the incomplete training dataset that does not contain all behavior patterns. Also, network traffic constantly evolves, and new applications and communication protocols change the traffic structure in time. Thus, the datasets become obsolete and are no longer usable for current deployment. Therefore, we recognize an urgent need to create a methodology for evaluating the quality of datasets.

There are many papers regarding the quality of data and the evaluation of classification or detection models. Well-known related works are listed in Sec. II, however, none of them completely fits our needs for *the quality of datasets evaluation*.

To introduce our domain, network infrastructures are usually monitored by one or multiple monitoring probes. In large-scale infrastructures, it is impossible to deploy highly resource-consuming analytical methods on the monitoring probes directly due to performance limits. Thus an aggregated form of data is being used, usually known as IP flow data (NetFlow or IPFIX). Additionally, such flow data can be extended by easily computable information about the traffic and packets. Altogether, such data proves to be feasible as input for accurate machine learning models.

Even though the probes cannot comprehensively analyze the whole traffic at the packet level (especially at speeds above 100 Gb/s), there is enough capacity to capture packets selectively besides the computation of IP flow data (as it was described in [1]). This way allows for automatic capture of packet samples that can be the base of automatically created datasets of IP flow data. Moreover, this dataset can be annotated by deep packet inspection at the packet level.

As a result, it is possible to create annotated datasets of theoretically limitless size automatically. The main challenge is to evaluate the feasibility and quality of the created datasets for training some ML models, e.g., for some classification problems. Furthermore, large network datasets usually contain a lot of redundant data, which cost much storage capacity and computational power during the training without significant benefits in the model's performance. Thus, a method for dataset reduction is needed to speed up the training process and save additional costs for data storage.

A similar challenge is an evaluation of existing external datasets. There are already many publicly available datasets; however, it is not clear how to decide whether they can be used for training ML models or not. The only way is usually to go through the complete process of training, testing, and validation of the machine learning model.

This paper summarizes the state of the art of data quality, formalizes the dataset quality definition, and introduces the required terminology necessary for the quality evaluation. Additionally, we propose a novel framework as the first attempt to evaluate the data quality. Finally, we identify open research challenges that should be addressed in the future.

A. Research Challenges

This paper is an introduction to our research covering the state of the art study and proposing preliminary definitions of essential terms, and a framework for evaluation of the quality of datasets. The research aims to explore possibilities of data science and evaluation methods to deal with the following open research questions:

1) Having the automatic continuous dataset capturing and annotation, can we decide that the dataset is ready

978-3-903176-36-2 ©2021 IFIP

enough, or should the capturing continue?
2) Assuming we have two datasets of the same network traffic captured at different times, are we able to decide if they are equivalent?
3) If not, is it possible to compare them and decide which one is better (or if they should be merged together)?
4) If one dataset is better, is it valuable to use it to retrain the ML model?

An additional challenge focuses on the methodology of a new dataset creation when model retraining is necessary. Possibly, there are two options: (1) We can create an entirely new dataset and retrain the ML model once we have sufficient results, (2) We can combine the previous dataset with the novel data. The second option is a very challenging task, requiring thorough data analysis, but the benefits of the older data preservation ensure robust and continuously evolving machine learning models. The ideas about dataset dissimilarities have already been published in Mokbel et al. [2]; however, this research does not provide a clear data combination technique. Ideally, we would like to keep the smallest possible dataset with only relevant data. To our best knowledge, there is no methodology on how this can proceed properly.

The remainder of this article is structured as follows. Sec. II provides an overview of state of the art in dataset and information quality evaluation. Sec. III lists definitions that we needed to describe our goals and scope of interest. Sec. IV proposes a framework that is being designed for the evaluation of the quality of datasets and was already applied on several sample use cases. Sec. V discusses the possible next steps within this research topic. Finally, Sec. VI concludes this article and outlines our future work plans.

II. State of the Art

The most comprehensive papers are done around the traditional area of data or information quality. Generally, understanding what are good data and how data quality can be measured or improved is a difficult task due to various definitions. There are also different contexts of data applicability that increase the complexity. Laranjeiro et al. [3] prepared a comprehensive summary in data quality. The quality of data is analyzed using dimensions. The number and definitions of dimensions usually differ based on the specific use case. As a model for data quality assessment, we can consider ISO/IEC 25012 standard [4] that describes 15 dimensions. These dimensions are organized into two categories: *inherent*, which has intrinsic potential to satisfy required needs under specific conditions, and *system dependent*, which is focused on how data is stored and used in computer systems under particular conditions. The number of categories also varies among different authors who use more granular segmentation. For example, Lee et al. [5] use four main categories: *intrinsic*, *accessibility*, *contextual*, and *representational*.

Another approach for data quality assessment proposed by Byabazaire et al. [6] is focused on trust in IoT systems. The discovered framework is very relevant to the IoT environment with heterogeneous sensors. These sensors can provide the same data but with different quality and build various datasets. In the trust framework, the authors define three stages (*starting*, *investigation*, *results*) that are mapped to data processing workflow in IoT systems and allow to have end-to-end trust score. In order to evaluate a complete trust score, some steps need to be validated manually, and ground truth is required. It is essential to verify data quality over time, especially for heterogeneous sources with different reliability [7].

In [8], the author defines requirements and definitions to interpret Random Forest models. Defined inputs are used in the developed CHRISP framework to explain Random Forest models. Our vision is similar to this paper. However, instead of ML model level, we are focused on dataset level. With the dataset approach, it is also possible to improve the overall performance of ML models on top.

It is usually assumed that the dataset used for the prototype in a testing environment will be the same as in a real environment. However, this assumption is wrong because of the network traffic dynamic. The testing dataset is often constructed with bias and can reflect our classes with different imbalanced ratios and data distribution. This is a well-known challenge that is described by Brabec et al. [9].

Analysis of dynamic data also requires updates of the training dataset. Updates can add new data discovered in our environment or a new dataset column that brings information and improves the results of an ML algorithm. This challenge is known as the data drift concept [10]. The drift concept is still in the early stages for proper detection of data change, but it is essential since it can automatically detect when it is beneficial to retrain the ML algorithm.

To sum up the state of the art, most of the papers are focused on data dimensions that describe the quality of data. However, a method for the evaluation of datasets quality is missing. Beyond the classical data quality concept, we need to consider, for example, class imbalance and the amount of data for each class with respect to the intended domain. In this paper, we leverage the current state of the art in quality of data and extend it about quality of datasets. We introduce relevant definitions and the framework that address opened research questions from a high-level perspective.

III. Defining Quality of Dataset

Based on the current state of the art and definition used for data quality, we identified essential terms for the quality of datasets that need to be specified. The scope of our definitions is general and can be applied to different datasets, ML models, and domains. More detailed application of datasets quality evaluation is a subject of future work. However, it is important to consolidate used terms and definitions to cooperate with different research teams. During our experiments and cooperation with several research teams, we identified the following terms.

A. Dataset

Input data is a mandatory prerequisite for all ML models. The collection of input data is called a *dataset*. Each dataset is

structured in data records that define the format.

Based on the specific domain, the format of the data record can be different. The general form of the data record is a pair of data and respective annotations. The requirement of the annotation part is set by the type of ML model. There are not any strict requirements on datasets format. More specifically, datasets are described by characterizing features grouping, content, relatedness, and purpose as mentioned in [11].

To evaluate ML models, we split our dataset into training and testing groups. This is a critical point for quality testing to know that both parts are relevant for the final model score.

As described in [12] we have different ML algorithms that leverage different types of learning. Based on the selected learning method, the transformation and extraction of relevant data is done. The output of this transformation is called a *feature dataset* or *featureset*.

B. Good Dataset

All terms above are common among ML researchers and natively used. As mentioned in [13], there are no consistent metrics for datasets. The quality of published datasets is usually based on the authors' reputations. Moreover, due to the class imbalance factor and dynamic changes in some domains (e.g., networking), it can be hard to reuse public datasets in a different environment. Therefore, we need to verify if the intended dataset is good enough for our use case.

Good Dataset: Dataset that fulfills the following conditions: 1) It contains annotations for each row that represents the required output; 2) It contains a reasonable amount of data to train and test ML algorithm; 3) It is compliant with inherent data quality (*accuracy, completeness, consistency, credibility, currentness*) items from [4] with respect to the ML model type.

Completeness and Reliability Measures: If one or more of these conditions is not satisfied, the dataset cannot be considered a *Good dataset*. However, the inherent data quality measures are vague, imprecise and their fulfillment would need to be evaluated subjectively. Therefore, we have formulated a more precise probabilistic definition.

Assuming that we have a dataset D and domain of ML deployment (for example, one particular network), and we observe the domain with monitoring apparatus that can extract correct features and labels (ensures the *Credibility*). We can formulate dataset quality by calculating the probability vectors of completeness measure C_k, and reliability measure R_k:

$$C_k = P(x \in D | x = L_k); \forall k \text{ in } \{0, 1, \ldots, \dim(L)\} \quad (1)$$

$$R_k = P(x = L_k \wedge d(x) = L_k | x \in D); \quad \forall k \text{ in } \{0, 1, \ldots, \dim(L)\} \quad (2)$$

where L is a vector of correct labels, $\dim(L)$ is the number of labels, x is an observed value in the domain of deployment, and $d(x)$ is the label of x present in the dataset D.

The C shows the probability that the data record included in the dataset D occurred in the domain of ML deployment. Due to the imbalanced problems, the probability is calculated for each label separately. Similarly, R represents the occurrence probability of falsely labeled data samples for each label separately. The vector C covers *Completeness* and *Currentness*. Similarly, the vector R covers *Accuracy* and *Consistency*.

Data Quality Quantification: Finally, the C and R are then combined into a single number by calculating the harmonic mean between vectors' averages:

$$D_q = \frac{2 \cdot \text{mean}(C) \cdot \text{mean}(R)}{\text{mean}(C) + \text{mean}(R)} \quad (3)$$

Even though the statistical definition is much more precise and allows us to quantify the data quality D_q, it is very hard to obtain such a number. It would require constant monitoring of the deployment environment and techniques that can reliably label data. This is generally not possible; therefore, the probabilistic definition can only be approximated. For example, we can use a smaller amount of offline labeled data or tailored data synthesis.

C. Better Dataset

For any ML problem, there is not only one definite dataset that would work perfectly. We can find many alternative datasets that can differ for every problem, e.g., by environment or time. Thus, we need to compare our datasets to know which one is more suitable or if their partial or complete merge is more relevant. To compare datasets, we need to have a method to validate their feasibility for the final application.

Thus, to decide if a specific dataset is better, we need to have the following inputs: 1) Dataset; 2) Domain; 3) ML model.

This input uniquely identifies the usage domain and allows to validate the quality of the dataset. In other words, we can say that dataset A is better than dataset B if the following conditions are satisfied: 1) Dataset A is a Good dataset; 2) Quality score of dataset A is higher than dataset B for the domain and ML model.

If any condition is not fulfilled, dataset A is not better than dataset B.

D. Minimal Dataset

Each dataset consists of a minimal dataset that can expand with additional data records. The added data records are used to extend the end states by other examples. However, the dataset is minimal, if and only if there is no duplication in it, and by removing any record, we will lose *Completeness* (a unique description of any state). This term is not new, and it is also used, for example, in [14]. A promising way to effectively reduce datasets seems to be clustering methods.

IV. Our Approach: Method Overview

Existing dimensions and definitions covered in previous sections perfectly describe the ideal state of datasets and their relationships. This is important to understand possible challenges and improve the overall results of an ML model. However, none of these definitions is providing an objective way to validate our datasets. To address this gap, we designed a

framework to evaluate the quality of datasets which is depicted in Figure 1.

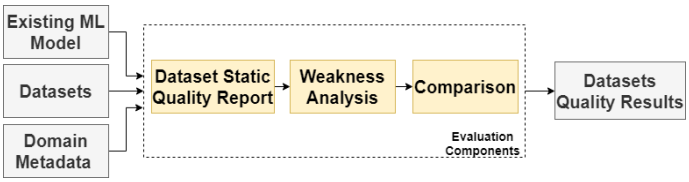

Fig. 1. High-level design of the framework for quality of datasets evaluation.

The aim of this framework is to find weaknesses and evaluate the quality of datasets with respect to the particular environment and domain as described in Sec. III. Therefore, as input, we require datasets and an ML model with its metadata describing the domain. At the first stage, input datasets are statically tested to do an initial check of dataset format and *Good dataset* conditions. The second stage (*Weakness analysis*) is doing a dynamic evaluation of the dataset. Together with input metadata about the domain and ML model, we generate different datasets to see their bias. In the last stage, we compare different versions of generated datasets and optionally input datasets with each other to identify *Better dataset*. The output of the whole process is a quality report with results, recommendations, and further optimization.

A. Dataset Quality Evaluation

Even though the proposed framework is in initial phase of implementation and further research with more experiments is planned, we already applied it on several use cases.

Our static quality report explores the structure of provided *dataset/feature dataset*. We analyze the class imbalance ratio, amount of data for each class, and statistic values for each class with respect to provided annotations. This helps to understand attributes of *Good dataset*. In many public datasets we saw low number of data records per class with various imbalance ration or datasets without any updates.

Weakness analysis works on the principle of synthetic data generation. Initially, datasets are created using the data synthesis algorithm, which has to be tailored for a specific use case. One of the biggest advantages of data synthesis is that datasets do not have to be anonymized, and they can be created cheaply and easily. We analyze responses of the existing ML model to synthetic datasets in the form of the accuracy of the ML model. During the analysis, we try to find datasets with certain properties: 1) The resulting dataset is *minimal dataset*; 2) The dataset will be extended by additional records, and thus a *better dataset* will be created. The whole process of creating *minimal* or *better dataset* is depicted in Figure 2.

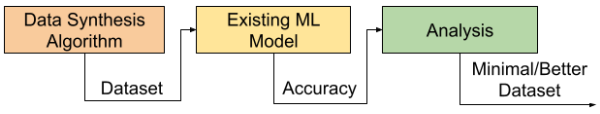

Fig. 2. Process of Better or Minimal dataset generation for DGA domain.

An example of weakness analysis can be an evolutionary algorithm for the creation of a *minimal* or *better* datasets. We tested this setup in the Domain generation algorithms (DGA) domain. DGA generates pseudo-random combinations of characters (a-z, 0-9, -) or words that form strings based on the input DGA seed [15].

The operation of the evolutionary algorithm is as follows. After running the weakness analysis, the population is initialized randomly. The population contains 30 DGA seeds (chromosomes), each seed generates 100 DGA domains. The resulting dataset includes 3000 DGA domains. We do not work with the generated domains but with seeds. A seed is consisted of the individual characters that are called genes. One seed generates a certain number of domain names. The accuracy of the classifier is defined by a fitness function that is applied to the individual chromosomes. Each fitness function evaluation means running a classifier with a certain number of domain names. After evaluating fitness function, two individuals from the population to which the crossover operator and mutation operator are applied are selected randomly. This procedure is applied until the overall accuracy of the population reaches 0 % (*Better* dataset), 100 % (*Minimal* dataset) or reaches the maximum number of generations. The accuracy of the whole population is calculated as the average accuracy of individual chromosomes. *Better* dataset is created by merging the dataset with worth accuracy with an original dataset. *Minimal* dataset is consisted of records reaching the best accuracy.

The experiments were performed by applying weakness analysis to the DGA domain. Each experiment consisted of 50 separate runs. The figures 3 and 4 explain the dependency of the number of the generations of the evolutionary algorithm on the accuracy of the existing ML model. The accuracy of the ML model on the created dataset is shown on the Y-axis. On the X-axis, there is the number of generations of the evolutionary algorithm. The first Figure 3 shows that finding a dataset that achieves the worst results is fast. Subsequently, the created dataset can be used to augment the original dataset and create a *better* dataset. The second Figure 4 looks for a *minimal* dataset, which can also be found very quickly.

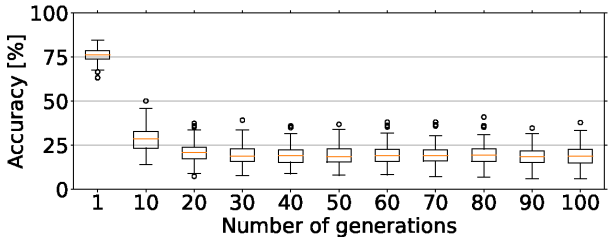

Fig. 3. Minimal dataset created using the weakness analysis.

A prototype of the implementation of weakness analysis is in the beginning, and we would like to achieve 0 % or 100 % accuracy by the evolutionary algorithm. DGA is not the only domain of problems where weakness analysis is applicable. Another area of application we tested is the analysis of existing models for the network traffic classification. The

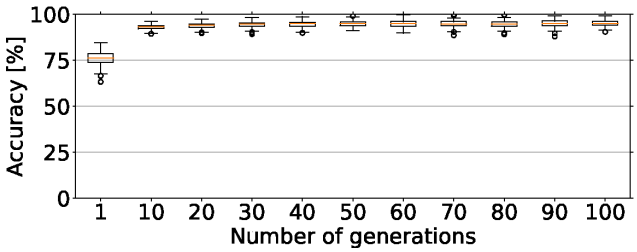

Fig. 4. Worst dataset create using the weakness analysis.

area of network traffic is more complex allow more operations. However, the detailed description is out of the scope of this paper. The weakness analysis would be able to improve existing models if it could find a *better* dataset. In case a *minimal* dataset is found, the learning time would be shortened.

V. Next Steps

This paper includes initial definitions of analyzing the quality of datasets. In the following steps, we will deeply explore open research questions using the suggested framework.

With our current ability to verify one dataset, we will investigate possibilities to compare two datasets and find the best way to merge them efficiently. This will require a more statistical approach to identify proper metrics.

We have already started with preliminary experiments of semi-automatic network classification. An essential part of this workflow is also the evaluation of the quality of created datasets. Automated pipelines can create datasets with feasible size and expected performance of machine learning models that use them.

Once we finalize our research and experiments for the challenges above, we will continue with further research questions to complete our designed framework.

VI. Conclusion

Recently, machine learning technology has begun to be intensively applied by researchers in the field of network traffic analysis. An essential prerequisite of this approach is the availability of proper datasets that affect training and testing machine learning models. Nevertheless, existing works do not pay enough attention to the quality of datasets. Therefore, machine learning techniques might not always work correctly, and their production deployment is challenging. Research on the quality of such datasets is beneficial to predict the performance of the trained models or their usefulness.

Evaluation of the quality of datasets was identified as an open problem and a non-trivial research challenge. The paper mentioned use cases that can benefit from an existing working methodology covering this task, e.g., evaluation of (automatically created) datasets, comparing datasets from different times or sources. This paper listed the current related works. However, mentioned authors are focused rather on the quality of data generally, but some targeted research for evaluating network traffic datasets for machine learning is missing.

The main contribution of this paper is the identification and description of the open research area, which is being partially elaborated in this paper. The paper presented several formalized definitions of terms related to the evaluation of the quality of datasets for machine learning applications, primarily in the network traffic domain. The defined terminology was used to describe a new framework for the evaluation of datasets. As an example, the framework was applied at the datasets for DGA detection and network traffic classification. Using the described framework, we were able to identify flaws in the datasets that would affect the performance of the trained models.

Even though the current results are rather preliminary, the idea of evaluation of the quality of datasets can be applied in many areas. It is essential, especially for the scenario of automatic dataset creation and annotation, where the size of the datasets may grow very quickly. There is currently no existing mechanism that would measure their feasibility, i.e., quality.

Acknowledgement

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833418 and also by the Grant Agency of the CTU in Prague, grant No. SGS20/210/OHK3/3T/18 funded by the MEYS of the Czech Republic.

References

[1] Z. Rosa *et al.*, "Building a feedback loop to capture evidence of network incidents," in *12th International Conference on Network and Service Management (CNSM)*, Montreal, Canada, 2016.

[2] B. Mokbel *et al.*, "How to quantitatively compare data dissimilarities for unsupervised machine learning?" in *Artificial Neural Networks in Pattern Recognition*, 2012.

[3] N. Laranjeiro *et al.*, "A survey on data quality: Classifying poor data," in *2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC)*, 2015, pp. 179–188.

[4] IOS/IEC, "Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality mode," 2008.

[5] Y. W. Lee *et al.*, "AIMQ: a methodology for information quality assessment," *Information & Management*, 2002.

[6] J. Byabazaire *et al.*, "Using trust as a measure to derive data quality in data shared iot deployments," in *29th International Conference on Computer Communications and Networks (ICCCN)*, 2020.

[7] I. Caballero *et al.*, "A data quality in use model for big data," in *Advances in Conceptual Modeling*, 2014.

[8] J. Hatwell *et al.*, "CHIRPS: Explaining random forest classification," *Artificial Intelligence Review*, 2020.

[9] J. Brabec, T. Komárek, V. Franc, and L. Machlica, "On model evaluation under non-constant class imbalance," 2020.

[10] I. Žliobaitė, M. Pechenizkiy, and J. Gama, *An Overview of Concept Drift Applications*, 2016.

[11] A. H. Renear *et al.*, "Definitions of dataset in the scientific and technical literature," *Proceedings of the American Society for Information Science and Technology*, 2010.

[12] N. J. Nilsson, "Introduction to machine learning," 1998. [Online]. Available: https://ai.stanford.edu/ nilsson/MLBOOK.pdf.

[13] A. Kenyon *et al.*, "Are public intrusion datasets fit for purpose characterising the state of the art in intrusion event datasets," *Computers & Security*, 2020.

[14] A.-F. Domensino *et al.*, "Defining the content of a minimal dataset for acquired brain injury using a delphi procedure," *Health and Quality of Life Outcomes*, 2020.

[15] R. Fangli, J. Zhengwei, W. Xuren, and J. Liu, "A dga domain names detection modeling method based on integrating an attention mechanism and deep neural network," *Cybersecurity*, vol. 3, no. 1, 12 2020.