Dataset Quality Assessment in Autonomous Networks with Permutation Testing

José Camacho
Dep. of Signal Theory, Telematics and Comm., CITIC
University of Granada
Granada, Spain
josecamacho@ugr.es

Katarzyna Wasielewska
Dep. of Signal Theory, Telematics and Comm., CITIC
University of Granada
Granada, Spain
k.wasielewska@ugr.es

***Abstract*—The development of autonomous or self-driving networks is one of the main challenges faced by the telecommunication industry. Future networks are expected to realise a number of tasks, including network optimization and failure recovery, with minimal human supervision. In this context, the network community (manufacturers, operators, researchers, etc.) is looking at Machine Learning (ML) methods with high expectations. However, ML models can only be as good as the data they are trained on, which means that autonomous networks also require a sound autonomous procedure to assess, and if possible improve, data quality. Although the application of ML techniques in communication networks is ample in the literature, analyzing the quality of the network data seems an ignored problem. This paper presents work in progress on the application of permutation testing to assess the quality of network datasets. We illustrate our approach on a number of simple synthetic datasets with pre-established quality and then we demonstrate its application in a publicly available network dataset.**

***Index Terms*—data quality assessment, permutation testing, anomaly detection, classification, network data, autonomous networks, self-driving networks**

I. Introduction

There is an increasing interest in the development of new machine learning (ML) methods to improve the performance of communication networks in tasks like network monitoring, troubleshooting and optimization [1]. Massive amounts of data can be easily gathered from network deployments [2], and there is the (sometimes naive) belief that every time enough data are available, ML tools can be a suitable problem-solving tool. However, the well-known phrase "garbage in - garbage out" (GIGO) reflects the general agreement that ML tools can only be as good as the data they are trained on, and so we need high-quality datasets [3] [4]. GIGO is a very relevant concept in the study of autonomous networks, which are expected to operate with minimum human intervention. For that purpose, autonomous networks need to generate data to train and update ML models. In this context, data quality problems, like mislabeling, incompleteness or lack of generalization, need to be identified in the generated datasets, which calls for means to assess data quality [5].

A lot of effort has been devoted by the scientific community to improve the quality of ML models from low-quality datasets [6]. In order to improve model quality, operations like data cleaning and optimization of hyperparameters are usually performed. While the effects of duplicates, outliers or missing values are relatively easy to fix, the general characteristics of the dataset such as completeness, accurateness, consistency, variety or uniqueness, remain difficult to assess [7]. Moreover, class imbalance [8], class overlapping [9], noise in data [10], incorrect labeling [11] or even the size of the dataset [12] can also affect the training of ML models. As shown in [13], permutation testing [14] [15] can be a useful tool to evaluate the impact of noisy data on the model performance.

Contrary to previous work, our goal in this paper is to assess the quality of the dataset, rather than the quality of a model. The quality of the dataset affects the development of any model trained from it, and thus we believe it is a more general problem. Dataset quality evaluation is key in the analysis and modeling of big data [16], and it is of interest when developing new benchmarking datasets, critical for network research. Recently, metamorphic testing has been proposed for validating the variety, fidelity and veracity of a model in connection with the data [17]. This approach detects errors in the training dataset verifying the relationships between the input and output data. In turn, crowdsourcing improves the quality of the dataset by taking human knowledge into account [18].

This paper introduces a novel approach for the application of permutation testing to assess the quality of network datasets. Preliminary results show a promising performance of the proposed method in both synthetic and real data and for tasks like anomaly detection and classification. The rest of the paper is as follows: Section II introduces the proposed permutation approach for dataset quality assessment, Section III presents the results of the experiments and Section IV draws the conclusions.

II. Proposed approach

Permutation testing [14] is a simple and intuitive form of non-parametric inferential statistics. Thus, it is used for the computation of probabilities and p-values that are specifically tailored to the dataset at hand, without any assumption about the null distribution. We can use this method to evaluate to which extent a model performance figure (e.g, classification accuracy, correlation coefficients, etc.) is the result of chance.

978-1-6654-0601-7/22/$31.00

The permutation testing is a resampling approach. Take for instance a dataset $\mathbf{X}$ with N observations from two variables. We can test whether the correlation between these variables is significantly high by creating a large number of permuted instances of $\mathbf{X}$ where the order of the observations in (only) the first variable is randomly shuffled. This would produce one hundred new correlation coefficients, one per permuted dataset, representing a null distribution of the correlation coefficients we can expect in a dataset like $\mathbf{X}$. If our true correlation is above, say, 99 of the 100 resamples, then we can say it is statistically significantly high with a p-value ≤ 0.01.

In a labelled dataset, where observations belong to a set of predefined classes (like normal vs anomalous) we can compute permutation tests in different ways. A first choice is to permute only the labels, which can be useful to check whether the dataset at hand, $\mathbf{X}$, is able to predict a wider set of different labeling instances. This would indicate that the content of the data in $\mathbf{X}$ is not specific of the true labeling, but rather general, and in turn that the association created in ML models trained with this data is only the (partial) result of chance. Alternatively, we can permute the $\mathbf{X}$ block to check how easily can we predict the true labeling by using randomized data, in turn assessing again the significance of the association of the true data with the labeling. Both possibilities are tailored to the data that remains unaltered: the $\mathbf{X}$ block in the first approach, and the labeling in the second. In this work in progress, we are interested in the first form of permutation, where only labels are permuted.

Since we are interested in assessing the quality of the data, rather than of a particular model, our approach includes a pool of classifiers ranging from simple to complex (non-linear) classifiers. The motivating assumption is simple: if all models perform bad, then the data is of bad quality; but if a subset of models perform good, then the data is of good quality. This assessment is limited by the availability of good modelling approaches. To assess the quality of a dataset (e.g., a benchmark), the pool of classifiers can incorporate state-of-the-art models as they are published, so that we guarantee that the evaluation includes the best possible modelling approaches. In this work-in-progress, we did not look in detail the sensitivity of the approach to the specific choice of classifiers.

Using the classifiers, we obtain a pool of performance measures for the dataset. To test whether this performance is significant, we apply permutation testing to percentages of the observations (from 100% to 1%) so that we can evaluate the loss of performance (if any) when only part of the data is permuted. This allows us to assess the relevance not only of the entire dataset, but also parts of it, in order to assess the accuracy of the labeling throughout the entire data. This approach limits the concept of data quality to problems like unsupervised anomaly detection and supervised classification, and the quality dimension assessed is whether the data contains enough information to predict a specific labelling.

An advantage of permutation testing in comparison to the other methods for data quality assessment mentioned in the introduction, is that the former creates a null distribution that allows us to test the statistical significance of the pool of classifiers fitted from the model.

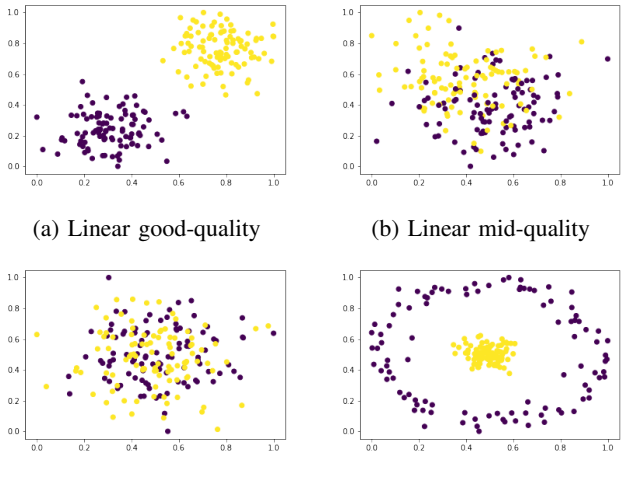
III. Experiments and results

To illustrate our approach, we use the Weles tool [19] and a pool of classifiers with default metaparameters: the *K* Nearest Neighbours (KNN), Gaussian Naive Bayes (GNB), AdaBoost (Adab), Decision Tree (DT), Random Forest (RF) and the Multi-Layer Perceptron (MLP). We apply stratified cross-validation and consider the F-measure (F1) as performance value. Our approach can also be used with other classifiers and performance measures.

A. Toy datasets

We created four random datasets with two variables and 200 observations evenly distributed in two classes. The four datasets are depicted in Figure 1. The first dataset represents good-quality data which can be linearly separated. This is the simplest case for a ML model. The second dataset is arguably of mid-quality, since both classes are partially overlapping. Given that we control the data generation mechanism and that there are no additional variables in the data, we know for certain that no non-linear model can separate the classes in the overlapping region in a systematic way. This situation is even worse in the linear bad-quality dataset, the third example, where observations are drawn from a common distribution and are randomly assigned to a class. Finally, the fourth example is a good-quality non-linear dataset, which requires non-linear ML models for classification. These examples illustrate that our concept of quality is not synonym for complexity: the non-linear data is complex to model, but quality remains good.

The results of the data evaluation with our permutation approach is shown graphically in Figure 2 and numerically in Table I. The permutation charts in the figure show the F1 obtained for each classifier on the true data (diamonds) and on the data after partial permutation (circles). Considered



(a) Linear good-quality (b) Linear mid-quality

(c) Linear bad-quality (d) Nonlinear good-quality

Fig. 1: Toy datasets: two variables and 200 observations, half of them belonging to one in two classes.

percentages for permutation are: 100%, 50%, 25%, 10%, 5%, 1%, for each of which we compute 100 resamples (600 in total). In the abscissas, each permutation is located depending on the correlation between the true labeling and the permuted one [20], so that resamples of 100% of permutation are located around the 0 correlation, while those for 1% are close to correlation 1. The tables show the corresponding p-values, obtained as:

$$P = (No.\ of\ (F1^* \geq F1) + 1)/(Total\ no.\ of\ F1^* + 1) \quad (1)$$

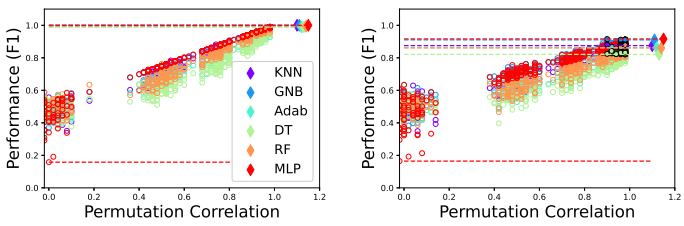
where $F1$ refers to the statistic computed from the real data and $F1^*$ stands for the statistics of the permutations.

Figure 2a shows that all classifiers obtain an optimal performance in the linear good-quality dataset: this is seen in the fact that all diamonds are set to 1 and all permuted results are below that value. The corresponding table shows that results are statistically significant even for 1% of permutation, and therefore when very small inconsistencies in the data are induced. Thus, this dataset is of optimal quality and also simple to model, since all classifiers perform optimally. Figure 2b reflects that the quality of the data is not that good. In the corresponding table we can see that none of the classifiers are significant for 10% or below, which should be understood as the result of the overlapping region in Figure 1b. Figure 2c and the corresponding table show that the performance of the classifiers is not significant for any 100% of the permutation, reflecting that there is no useful information in the bad-quality dataset to predict the labelling. Finally, Figure 2d and the corresponding table show that only a sub-set of classifiers can perform well at all permutation percentages. This reflects that the data is of optimal quality (at least one classifier is significant at 1%) but complex (not all classifiers perform well).

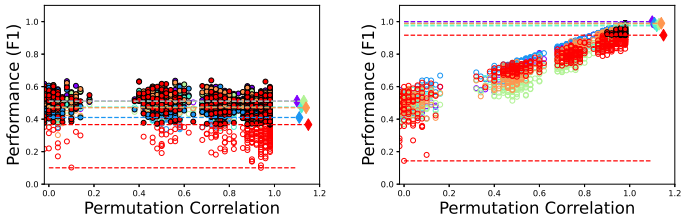
B. Case study 2 - the inSDN dataset

The second experiment was conducted on the publicly available inSDN dataset [21]. This dataset was recently published to provide attack-specific Software Defined Networking (SDN) data to the research community and it is attracting a moderate interest. The data comprises traffic flows of normal (legitimate) traffic and of a number of attacks. We considered the complete dataset, with nearly 350K flows (68.424 normal and 275.465 attacks), and also another instance of the dataset where only DoS attacks are considered (68.424 normal and 53.616 attacks), and evaluated the quality of the data for the problem of classifying attacks from normal traffic. Due to the data size, we compute 25 resamples per each permutation percentage considered before (a total of 150).

The results of the data evaluation with our permutation approach is shown in Figure 3 and in Table II. Looking at the table, the dataset could be considered of high quality, since most classifiers attain significant performance for any percentage of permutation. However, the permutation chart shows an interesting pattern: several permuted models achieve better quality for 100% of permutation than for 50%. This behaviour is particularly pronounced in the dataset with all the attacks,



(a) Linear good-quality (b) Linear mid-quality



(c) Linear bad-quality (d) Nonlinear good-quality

Fig. 2: F1 permutation charts: true performance is shown with diamonds and horizontal lines, resamples with circles, the minimum performance is also represented by a horizontal line.

TABLE I: P-values computed in the permutation. Symbol . should be read as <0.01. All p-values above 0.05 are in red.

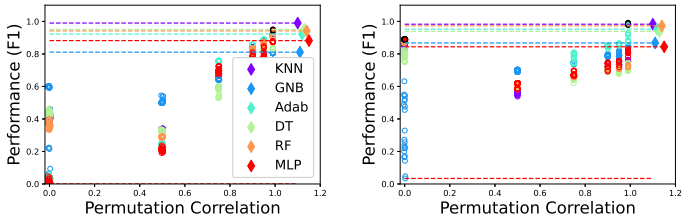
	Linear good-quality dataset						Linear mid-quality dataset					
	100%	50%	25%	10%	5%	1%	100%	50%	25%	10%	5%	1%
KNN	.	.	.	.	.	.	.	.	.	.14	.27	.44
GNB	.	.	.	.	.	.	.	.	.	.06	.16	.26
Adab	.	.	.	.	.	.04	.	.02	.02	.11	.26	.40
DT	.	.	.	.	.04	.15	.	.03	.05	.12	.17	.30
RF	.	.	.	.	.	.	.	.02	.06	.28	.28	.55
MLP	.	.	.	.	.	.	.	.	.	.07	.13	.22

	Linear bad-quality dataset						Non-linear good-quality dataset					
	100%	50%	25%	10%	5%	1%	100%	50%	25%	10%	5%	1%
KNN	.23	.33	.33	.23	.23	.53	.	.	.	.	.	.
GNB	.73	.54	.41	.50	.52	.61	.	.	.	.	.03	.11
Adab	.75	.82	.67	.60	.72	.78	.	.	.	.	.02	.11
DT	.71	.82	.68	.56	.65	.81	.	.	.	.	.	.05
RF	.23	.43	.34	.20	.23	.43	.	.	.	.	.	.02
MLP	.42	.42	.24	.14	.18	.10	.	.	.	.03	.09	.30

where most permuted models at 100% of permutation show an F1 around 0.8. Interestingly, from 50% of permutations, results follow a similar pattern to the one found in good-quality synthetic datasets. Our intuition is that the data is rich in patterns that can predict almost any labelling of the flows, explaining the good performance at 100% of permutations. When we only use partial permutations, the performance is decreased as both correct and permuted labels reflect contradictory patterns. We speculate that this behaviour can pose a relevant problem in unsupervised anomaly detection, when the labelling is not explicitly used in model training. Our initial attempts to build anomaly detection models in this data agree with this conclusion, but more experiments are needed to validate this conclusion.

IV. Conclusions and Future Works

Suitable evaluation of the quality of a dataset is critical for building high-quality machine learning models that can be put



(a) InSDN with DoS (b) InSDN with all attacks

Fig. 3: F1 permutation charts in the inSDN dataset [21].

TABLE II: P-values for the inSDN dataset [21]. Symbol . should be read as <0.05. All p-values above 0.05 are in red.

	InSDN with DoS						InSDN with all attacks					
	100%	50%	25%	10%	5%	1%	100%	50%	25%	10%	5%	1%
KNN	.	.	.	.	.	.	.	.	.	.	.	.
GNB	.	.	.	.	.	.	.	.	.	.	.	.
Adab	.	.	.	.	.	.12	.	.	.	.	.	.35
DT	.	.	.	.	.	.	.	.	.	.	.	.
RF	.	.	.	.	.	.	.	.	.	.	.	.
MLP	.	.	.	.	.	.19	1.00	.	.	.	.	.

into production, which in turn is mandatory for the development of autonomous networks. This paper presents work in progress to develop a methodology to evaluate the quality of a dataset based on a pool of classifiers and permutation testing. We have shown that the approach can successfully differentiate between quality and difficulty of classification of a dataset in both simulated and real data.

The main limitation of this approach is the computational cost, which is intensive and renders the method (as it is now) not applicable with computationally expensive (e.g., deep learning) classifiers or in massive datasets. However, full paralelization is always a posibility to speed up computation. Another limitation is the assumption that the dataset quality depends on the performance of the pool of classifiers. It may be the case that a good quality dataset cannot be classified with state-of-the-art machine learning methods. However, under this circumstance, the dataset itself is of little practical use for the training of those methods. Finally, the approach requires a labelled dataset, which can be hard to obtain but is also required in several practical applications.

As future work, we would like to test our approach on several real data sets more; validate our conclusions on the InSDN dataset; and derive a figure of merit to assess the quality of a dataset using the permutation approach: a scalar value that can complement the visualization and the table with p-values.

Acknowledgement

This work is partially funded by the Agencia Estatal de Investigación in Spain, grant No PID2020-113462RB-I00, and by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 893146. We would like to thank Dominik Soukup and Tomáš Čejka for their useful feedback and Szymon Wojciechowski for his support on the Weles tool.

References

[1] P. Kalmbach, J. Zerwas, P. Babarczi, A. Blenk, W. Kellerer, and S. Schmid, "Empowering self-driving networks," in *Proceedings of the afternoon workshop on self-driving networks*, 2018, pp. 8–14.

[2] H. Song, F. Qin, P. Martinez-Julia, L. Ciavaglia, and A. Wang, "Network Telemetry Framework," Internet Engineering Task Force, Internet-Draft draft-ietf-opsawg-ntf-13, Dec. 2021, work in Progress. [Online]. Available: https://datatracker.ietf.org/doc/html/draft-ietf-opsawg-ntf-13

[3] L. Caviglione, M. Choraś, I. Corona, A. Janicki, W. Mazurczyk, M. Pawlicki, and K. Wasielewska, "Tight arms race: Overview of current malware threats and trends in their detection," *IEEE Access*, vol. 9, pp. 5371–5396, 2021.

[4] I. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: an overview from machine learning perspective," *Journal of Big Data*, vol. 7, 07 2020.

[5] D. Soukup, P. Tisovčík, K. Hynek, and T. Čejka, "Towards evaluating quality of datasets for network traffic domain," in *2021 17th International Conference on Network and Service Management (CNSM)*, 2021, pp. 264–268.

[6] C. F. Caiafa, S. Zhe, T. Toshihisa, M.-P. Pere, and J. Solé-Casals, "Machine learning methods with noisy, incomplete or small datasets," *Applied Sciences*, vol. 11, no. 9, 2021.

[7] V. Gudivada, A. Apon, and J. Ding, "Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations," *International Journal on Advances in Software*, vol. 10, pp. 1–20, 07 2017.

[8] A. Sahu, Z. Mao, K. Davis, and A. E. Goulart, "Data processing and model selection for machine learning-based network intrusion detection," in *2020 IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR)*, 2020, pp. 1–6.

[9] M. Dudjak and G. Martinovic, "An empirical study of data intrinsic characteristics that make learning from imbalanced data difficult," *Expert Syst. Appl.*, vol. 182, p. 115297, 2021.

[10] S. Gupta and A. Gupta, "Dealing with noise problem in machine learning data-sets: A systematic review," *Procedia Computer Science*, vol. 161, pp. 466–474, 2019, the Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050919318575

[11] F. R. Cordeiro and G. Carneiro, "A survey on deep learning with noisy labels: How to train your model when you cannot trust on the annotations?" in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2020, pp. 9–16.

[12] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 843–852.

[13] M. Ojala and G. Garriga, "Permutation tests for studying classifier performance," *Journal of Machine Learning Research*, vol. 11, pp. 1833–1863, 06 2010.

[14] F. Pesarin and L. Salmaso, "A review and some new results on permutation testing for multivariate problems," *Statistics and Computing*, vol. 22, no. 2, pp. 639–646, 2012.

[15] F. Pesarin and L.Salmaso, "The permutation testing approach: a review," *Statistica*, vol. 70, no. 4, pp. 481–509, 2010.

[16] F. A. Batarseh, L. Freeman, and C. Huang, "A survey on artificial intelligence assurance," *Journal of Big Data*, vol. 8, no. 1, 2021.

[17] J. Ding and X. Li, "An approach for validating quality of datasets for machine learning," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 2795–2803.

[18] M. Bergman, T. Milo, S. Novgorodov, and W.-C. Tan, "Query-oriented data cleaning with oracles," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1199–1214. [Online]. Available: https://doi.org/10.1145/2723372.2737786

[19] K. Stapor, P. Ksieniewicz, S. García, and M. Woźniak, "How to design the fair experimental classifier evaluation," *Applied Soft Computing*, vol. 104, p. 107219, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494621001423

[20] F. Lindgren, B. Hansen, W. Karcher, M. Sjöström, and L. Eriksson, "Model validation by permutation tests: applications to variable selection," *Journal of Chemometrics*, vol. 10, no. 5-6, pp. 521–532, 1996.

[21] M. S. Elsayed, N.-A. Le-Khac, and A. D. Jurcut, "Insdn: A novel SDN intrusion dataset," *IEEE Access*, vol. 8, pp. 165 263–165 284, 2020.