Model Validation by Increasing Entropy in Datasets

Jimmy Mulder[1][0000- 0001-9681-863X] and Roelant Ossewaarde1[0000-0002-7036-522X]

[1] Hogeschool Utrecht, Heidelberglaan 15, 3584 CS, Utrecht, The Netherlands
jimmy.mulder@hu.nl

Abstract. We present a synthetic data generation method aimed at providing datasets of which the information value/entropy is known, allowing researchers to validate their models' capability to extract all the information in a given dataset, while doubling as a sanity check to catch possible programming mistakes or methodological errors. Our method involves training a neural network to reproduce data, then systematically damaging the networks' connections to create datasets of decreasing quality and increasing entropy.

Keywords: Modeling, Validation, Synthetic Data.

1 Introduction

Datasets contain information about the world, but the exact amount of information contained in a given dataset is often unknown. When data is ambiguous, noisy, flawed or compressed, this increases entropy which negatively affects the information value. For example, two patients could show the same symptoms with a different underlying cause, which would make it harder for a model to predict which disease each patient has. The information content of a dataset limits the theoretical maximal performance that any model trained on the dataset can achieve.

Data scientists try to build models that capture as much of the information in a dataset as possible. The quality of these models is often expressed in metrics regarding true and false negatives and positives, usually accuracy. But since the information value of a dataset is usually unknown, it can be difficult to properly interpret the meaning of these metrics. For example, if a model scores 80% accuracy, what should a data scientist do with that information? Could they perhaps do better by changing or tweaking the model, or is this the best performance they could hope to achieve given the information value of their dataset? The model performance could even be too high – it could be the result of a methodological flaw or programming mistake which inflated the accuracy metrics. Validation of these models is usually done by reserving a part of the dataset, which the model applied to after the training and testing phase, but this method does not catch the issues outlined above.

These are issues that data scientists struggle with on a regular basis. Much research focusses on increasing the accuracy of a new model compared to a baseline model, squeezing every last drop of information from a dataset [1]. If the information value of a dataset isn't known, such research can be a gamble; with the number of papers

published for a 1% accuracy increase, one can only guess at the number of failed experiments.

Similarly, it can be difficult to validate complex models by looking solely at their metrics. Human error in either methodology, code or data can artificially inflate the performance of a model. A famous example is a model that could distinguish wolves from huskies with great accuracy by looking at the background of a photo – if it contained snow, that usually meant the picture was of a wolf [2]. With the size and complexity of models increasing rapidly there is a demand for new validation tools.

In all of these cases it would have been helpful to have a dataset where the information value is known beforehand. We present a proof of concept for the validation of models using a synthetic data approach.

2 Method

Our method involves the addition of entropy to a well-known dataset with high information value, in our case MNIST [3]. Many models have been shown to achieve near-perfect accuracy on this dataset [4], which contains 70.000 handwritten digits. We introduce synthetic variations of this dataset with increasing entropy and show how these can be used as an additional way to validate models.

Our method to decrease the information value in MNIST begins with training a separate neural network to reproduce the images using an auto-encoder structure. The network is first trained to reproduce the data with optimal accuracy. This serves as a baseline. To increase the entropy, we 'damage' the neural network by randomly changing the weights between the neurons. The magnitude of the changes and number of affected neurons is governed by a single parameter which ranges from 1 (no damage) to 0 (damaged to the point where none of the input data is reproduced correctly by the network).

This is similar to pruning but differs in a significant way: the goal of pruning is to reduce model size without affecting performance, whereas our intention *is* to affect performance. 'Damaging' a connection between neurons can even mean increasing its weight rather than lowering or removing it.

Our reasoning for using this method rather than other forms of increasing entropy, such as adding white noise or changing labels, stems from the larger context of this research (modeling and diagnosing neurodegenerative disease), which we do not discuss further for brevity's sake. This method has been described by Lusch et al. [5].

On our poster we show the effects this has on the original dataset, and show how the relationship between information value and model accuracy can be used to assess a model's validity.

References

1. J. Lin, “The Neural Hype and Comparisons Against Weak Baselines,” ACM SIGIR Forum, vol. 52, no. 2, pp. 40–51, Jan. 2019, doi: 10.1145/3308774.3308781.
2. M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should i trust you?’ Explaining the predictions of any classifier,” in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.
3. Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, and E. Bottou, “Gradient-based Learning Applied To Document Recognition - Proceedings of the IEEE,” Proceedings of the IEEE, vol. 86, no. 11, 1998, doi: 10.1109/5.726791ï.
4. S. H. Hasanpour, M. Rouhani, M. Fayyaz, and M. Sabokrou, “Lets keep it simple, Using simple architectures to outperform deeper and more complex architectures,” Aug. 2016, [Online]. Available: http://arxiv.org/abs/1608.06037
5. B. Lusch, J. Weholt, P. D. Maia, and J. N. Kutz, “Modeling cognitive deficits following neurodegenerative diseases and traumatic brain injuries with deep convolutional neural networks,” Brain Cogn, vol. 123, pp. 154–164, Jun. 2018, doi: 10.1016/j.bandc.2018.02.012.