Performance of Machine Learning Models on Criminal Networks (MLMoCN) trained and evaluated with real, synthetic, and hybrid data

JOSÉ CANO-MELANI
EDUARDO SALCEDO-ALBARÁN
LUIS JORGE GARAY SALAMANCA

VORTEX WORKING PAPER No. 63

VORTEX

Vortex Working Paper No. 63.

© *José Arturo Cano-Melani,* jac@scivortex.org - SciVortex Corp, 2023.
© *Eduardo Salcedo-Albarán,* esa@scivortex.org - SciVortex Corp, 2023.
© *Luis Jorge Garay Salamanca*, ljg@scivortex.org - SciVortex Corp, 2023.

Abstract

The objective of this paper is to discuss the performance of Machine Learning Models on Criminal Networks (MLMoCN) trained, implemented, and evaluated in the software VORISOMA with only real data, synthetic data, and hybrid data; the latter, consisting of real data expanded through "layouts" in which variables are replaced with randomized content. This article has four sections. The first section is the introduction. The second section is a discussion of the nature and advantages of using synthetic data for training and evaluating MLMs. The third section is a presentation of the parameters used for generating synthetic and hybrid datasets, and the results of six experiments for training and evaluation of MLMoCNs. The fourth section is the discussion of results.

Introduction: Synthetic and Hybrid Datasets for Criminal Networks Analysis in VORISOMA

1

The size and realism of the datasets used for training Machine Learning Models (MLMs) affects the predictive capacity of the resulting model. For instance, after testing a MLM of Convolutional Neural Networks (CNNs) -commonly used for image pattern recognition (O'Shea & Nash, 2015)- trained with 5, 10, 20, 50, 100 and 200 elements, Choe *et al* (2016) found out that in every case a larger dataset increased the predictive accuracy of the model. Due to this effect, it has been questioned how much data is needed for training a model with "acceptable" predictive capacity, especially when real data is scarce or legally restricted. This question has been commonly raised in the field of MLMs for recognition and analysis of medical phenomena, due to legal regulations on the privacy of this data (Beaulieu-Jones, et al., 2019); however, it also applies to other social phenomena characterized by underreporting of reliable statistical data, such as criminal phenomena.

The scarcity of real data for training MLMs has been observed during the development and use of VORISOMA, a software that assists the analysis of complex Transnational Criminal Networks (TCNs) (Cano-Melani, Salcedo-Albaran, & Garay-Salamanca, 2022) by applying protocols of Criminal Networks Analysis (CNA) (Garay-Salamanca & Salcedo-

Albarán, 2012a; Garay Salamanca & Salcedo-Albarán, 2012). Although the Artificial Intelligence implemented in VORISOMA does not use CNNs for image classification, it uses qualitative data in form of natural language text that informs how individuals and corporations interact and establish TNCs.

Traditionally, CNA (Morselli C. , 2008; Morselli C. , 2012) has been assisted by software for modeling and visualizing networks, while VORISOMA articulates tools for data management, visualization, navigation, and analysis of TCN. However, in recent versions, VORISOMA includes capabilities for identifying and predicting TCN's characteristics, such as the types of agents and interactions that sustain an illicit network. As expected, these capabilities require developing MLMoCNs, a novel research and development field.

In VORISOMA the data is extracted from natural language text such as criminal records and judicial proceedings that sometimes are publicly available in countries characterized by high standards of accountability; however, in other countries opacity is the rule and the data isn't available. In fact, those countries characterized by high levels of opacity in the public information are usually also characterized by high levels of corruption and criminal activity; therefore, countries where real data is scarce are commonly the same hotspots that require detailed analysis. Due to the lack of real contextualized data that informs about those criminal hotspots, training MLMs of Natural Language Processing (NLP) has proven difficult during the recognition and prediction TCNs characteristics such as types of interactions, types of agents involved, locations of the interactions, among other.

Due to the scarcity of real data describing TCNs in specific countries, it has been considered the option of generating synthetic datasets or amplifying datasets based on real data, because synthetic data is a "promising alternative" for training or testing MLMs when the real data is scarce (Jordon, Wilson, & van der Schaar, Synthetic Data: Opening the data floodgates to enable faster, more directed development of machine learning methods, 2020). In other contexts, this generation has been conducted through

frameworks such as Generative Adversarial Networks (GAN) (Goodfellow, et al., 2014) that focus on anonymizing the real data (Jordon, Yoon, & van der Schaar, 2019).

Bearing this in mind, the objective of this paper is to discuss the performance of MLMs trained and implemented in VORISOMA AI with (i) only real data, (ii) only synthetic data, and (iii) hybrid data consisting of real data expanded through "layouts" in which variables are replaced with randomized content. This article has four sections. The first section is the introduction. The second section is a discussion of the nature and advantages of using synthetic data for training and evaluating MLMs. The third section is a presentation of the parameters used for generating synthetic and hybrid datasets, and the results of six experiments for training and evaluation of MLMoCNs. In the fourth section the results are discussed.

Synthetic Data

2

The concept "synthetic data" often refers to datasets generated by specifying the distributions of element. These datasets are commonly used for validating and comparing the performance of MLMs instead of replacing real data. However, it is argued that the concept "synthetic data" sometimes doesn't correctly describe the nature of a dataset composed by artificially created datasets based on real elements (Jordon, Wilson, & van der Schaar, Synthetic Data: Opening the data floodgates to enable faster, more directed development of machine learning methods, 2020, pág. 2); this is the case because the specified distribution and characteristics of a synthetic dataset should be based on the characteristics of the real data that is related to the MLM. Therefore, to differentiate between a dataset in which real data is expanded, and a dataset that doesn't follow this expansion, the concepts "hybrid" and synthetic" are herein used respectively.

When generating synthetic data, it is expected that the elements preserve a similar distribution to the one observed in the real elements that describe the analyzed and predicted phenomenon. Although similar, the distribution of elements in a synthetic dataset isn't identical to the distribution in a real dataset; therefore, the differences in the performance reflect the differences in the elements' distribution (Jordon, Wilson, & van der Schaar, Synthetic Data: Opening the data floodgates to enable faster, more directed development of machine learning methods, 2020). It is expected that approaches for generating synthetic data address the reduction of differences in the distribution of real versus synthetic or expanded elements.

Training MLMoCN With Real, Synthetic, and Hybrid Data

3

Six MLM training experiments were conducted using Spacy to evaluate the performance of various training datasets and methodologies for a single classification task against a fixed set of pre labeled reference sets of data extracted from judicial records regarding cases of TCNs, which are publicly available in VORISOMA.

The classification task consists of determining if a sentence refers to and describes a monetary transaction between specific individuals or organizations. A positive resulting classification means that the sentence is a transaction, and a negative resulting classification means that it's not. Before conducting the experiments herein described, the classification task was conducted by human analysts, which allowed consolidating the real data.

1.Datasets and parameters

The synthetic and hybrid datasets for training MLMoCN were generated with a framework. Each record is a sentence based on the following generating parameters:

- **Template:** Different ways to refer to a currency amount without excluding number format variations.
- **NumType:** Decimal, numeral, or "no number".
- **MatchType:** This is the transaction classification that defines whether the sentence is a transaction or not.
- **Format:** Various format options for a decimal number.
- **CurrencySymbol:** Various options of symbols or text labels referring to a specific currency.

- **Prefix:** Various options of arbitrary text included before the template.
  Suffix: Various options of arbitrary text included after the template.
- **Range:** Various options of numeric ranges for the main numeric expression used in the sentence to represent the transferred amount.
- **Non currency words:** Not considered as part of the parameters used for permutation; instead, randomly included in part of the prefixes or suffix to provide clues to the model on how to handle new words.

The following definitions allow understanding the consistency of the synthetic and hybrid datasets:

- **Hybrid Synthetic Data:** Consists of synthetic data with a different template set, customized to match some of the most common patterns found in the manually tagged samples. The process consisted of adjusting the synthetic generator to match these intuitively identified patterns.
- **Minimal Training Set:** Consists of the minimum sample set necessary to include synthetic data examples for all the parameter values of the generation parameters. This differs from the full combinatory - which includes all possible combinations- in that it provides examples ensuring that only all possible values for the each of the parameters are included in the generated samples, in contrast to including all possible value combinations, which is a larger set by several orders of magnitude.

  Exceptions to this rule are included in these minimal sets to ensure that some values are used and processed more than once, up to a specific minimum number of samples. As a result, those values are emphasized during training; specifically: (i) MatchType is forced to a minimum of 2500 samples per value, and (ii) CurrencySymbol is forced to a minimum of 4 samples per value.
- **Training Method, Forced Training and Resampling:** Training is implemented using Spacy's routine model updates with the default optimizer, in which data is

chunked in 200-sample batches. The model is repeatedly updated with the batch examples until all predictions are correct, up to a maximum of 15 times per batch.

Forced training repeats the process over the same dataset until the categorizer produces a controlled maximum number of errors of tolerance or no errors at all.

Resampled training is "forced" but keeps repeating the forced training process over new subsets of the training data, either by generating new randomized synthetic data or by extracting new random records from a larger sample set. The process is repeated until it generates no errors on the first pass over a new set.

Bearing in mind these parameters and definitions, the following synthetic and hybrid datasets were generated:

1. Training Set (Original): Real data. This is the dataset used as reference, which is expected to generate a match close to 100%. The error tolerance during the training occasionally generates some mismatches, which explains the difference.
2. Lava Jato (excluding training): Real data. This dataset includes sentences manually tagged as "financial transnational transaction" in a criminal network entitled "Lava Jato", analyzed in a VORISOMA workspace by human analysts.
3. "Extra Tagged Workspaces (excluding training)": Real data. This dataset includes sentences manually tagged as "financial transnational transaction" in other criminal networks different to Lava Jato, analyzed in VORISOMA workspaces by human analysts.
4. Synthetic Variant 1: Hybrid data. Minimal Set, with new random ranges and "new non currency words".
5. Synthetic Variant 2: Hybrid data. Minimal Set, with various and different non currency words.
6. Synthetic Variant 3: Hybrid data. Minimal Set, with different ranges, suffixes, and prefixes, but using the same non currency words as the training set.

Except for the training set, the evaluation sets exclude any samples used during training. For hybrid experiments the synthetic variants use hybrid synthetic data; however, for experiments with fully synthetic (non-hybrid) data, the synthetic variants use the same fully synthetic non-hybrid generator.

2.Training Experiments

1. **Pretrained With Lava Jato + Extra:**
   This is the reference experiment. The pretrained Spacy's language model was used to classify positively if the model detected entities labeled/tagged as 'MONEY', 'CARDINAL', 'QUANTITY'. In this case, the original training set is empty, and there is no actual training.
2. **Train With Synthetic Resampled:**
   An English blank language model was trained using a minimal synthetic dataset, repeating the training with new synthetic data and new samples until an error is produced.
3. **Train With Synthetic:**
   An English blank language model was trained using the full synthetic dataset including all valid combinations with a single pass, not forced nor resampled.
4. **Mixed Train Random (Lava Jato + Extra):**
   An English blank language model was trained using the manually classified samples from various VORISOMA Workspaces, including "Lava Jato".
5. **Train With Synthetic/Hybrid Resampled:**
   An English blank language model was trained using a minimal sample set of hybrid synthetic data. The training process was forced and resampled.
6. **Train With Synthetic/Hybrid (Single pass):**

An English blank language model was trained using the full sample set of hybrid synthetic data with a single pass, not forced nor resampled.

Experiment	Forced	Resampling 1: Yes 0: No	Training Set Size	Training Time (seconds)	Evaluation Time (seconds)
Pretrained With Lava Jato + Extra	FALSE	0	0	0	156,8
Train With Synthetic Resampled	TRUE	1	3855	371,9	74,6
Train With Synthetic	FALSE	0	813253	4907,5	2389,6
Mixed Train Random (Lava Jato + Extra)	40	0	1000	133,9	65,1
Train With Synthetic/Hybrid Resampled	TRUE	1	4280	417,6	720,2
Train With Synthetic/Hybrid (Single pass)	FALSE	0	4280	21,8	625,6

3.Results

The following were the results of each training, based on the datasets described above:

- **avg_ratio:** Average match ratio across all the evaluation datasets, excluding the training set.
- **avg_real_ratio:** Is the average match ratio across the manually labeled/tagged datasets only ("Lava Jato" and "Extra Tagged").

	params > name	params > forced	result > UNIFIED > total > avg_ratio	result > UNIFIED > total > avg_real_ratio
1	Pretrained With Lava Jato + Extra	FALSE	0.52952	0.716260
2	Train With Synthetic Resampled	TRUE	0.82451	0.502419
3	Train With Synthetic	FALSE	0.74888	0.494938
4	Mixed Train Random (Lava Jato + Extra)	40	0.65515	0.682148
5	Train With Synthetic/Hybrid Resampled	TRUE	0.76275	0.557398
6	Train With Synthetic/Hybrid (Single pass)	FALSE	0.64487	0.499598

Final Discussion

4

In general, the best performing training method was "Train With Synthetic Resampled", which consisted of only synthetic data used for repeatedly training until a mistake was generated:

- For real data, the best method was "Pretrained With Lava Jato + Extra" using Spacy's pretrained model.
- For the training methods on real data, the best performing method was "Mixed Train Random (Lava Jato + Extra)".

In various tests the training conducted with hybrid synthetic data performed 1% to 2% better than the fully synthetic data against real data, which is negligible but still relevant. On the other hand, "resampling" plays a significant role when improving synthetic based training. In this sense, trainings based only on synthetic datasets performed significantly worse than training based on real data.

Considering the results, in this case it was decided to use the model trained with real data for a particular predictive application; however, it is important to improve the hybrid generator and use the same comparison method herein described to determine its capacities. It can also be concluded that MLMoCNs trained with "forced" and "resampling" methods generate better results with smaller datasets, when compared to MLMs trained without these methods but with larger datasets, including all possible permutations of generation parameters.

As stated in the first section, the results herein illustrate that synthetic datasets might be more useful for evaluating the performance of a MLM than for training it; therefore, generating hybrid synthetic data for training purposes is significantly challenging. For instance, it is critical to identify the specific parts of the real data -sentences, in this case- that can be replaced without affecting the classification of the sentence, and then developing and using a custom generator that mixes those replaceable parts randomly. Strict calibration during the selection of those replaceable parts is required.

Regarding the comparison, it is important to highlight that the evaluation method proposed and used herein can be applied to determine if a hybrid data generator correctly represents a sample of real data, in more extended and future applications. This is important because, as discussed, real data regarding criminal networks is scarce; therefore, at some extent the training and evaluation of new MLMoCNs will rely on hybrid data consisting of expanded real data. Bearing in mind the scarcity of real data, in the short-term it will be critical to evaluate the quality of synthetic and hybrid datasets -with the parameters and the process herein proposed- to train MLMoCN.

These experiments and observations illustrate the challenges of building up the datasets for developing MLMoCNs. On one hand, the MLMoCN with best training and predictive performance results of using real data; on the other hand, this data is scarce because the phenomenon is underreported and because public and private stakeholders interested on the subject of TCNs usually don't follow procedures for accumulating, structuring, and labeling such a data. In the short term the lack of real data can be partially addressed by generating hybrid datasets; however, even this generation requieres structuring and consolidating real initial data, which is usually neglected by stakeholders. Without real initial data, it is impossible to reflect and partially explain the characteristics of a complex phenomenon such as TCNs.

Due to the complexity and massiveness of TCNs, it is critical to use Artificial Intelligence techniques -such as Machine Learning and Deep Learning- for analyzing and predicting the characteristics and dynamics the several variables involved in the phenomenon. However, to achieve this objective it is also required that stakeholders apply CNA protocols when conducting empirical analysis of domestic and transnational criminal networks. Otherwise, by training MLMoCNs with synthetic data that doesn't reflect the empirical characteristics of the phenomenon, the analysis will be misleading or distorting due to extreme simplification of variables. Such analysis will be misleading and useless even if AI techniques are applied.

In fact, it was possible to develop the first MLMoCN on VORISOMA because empirical analysis through CNA and consolidation of real data elements/interactions were conducted by human analysts during more than a decade at Vortex Foundation. Without these elements it would have been impossible to train MLMoCNs or even to generate hybrid datasets based on the empirical characteristics of TCNs. Generating hybrid data is similar to simulating the analyzed phenomena, since characteristics of unobserved TCNs are generated; this explains why such a simulation must follow empirical characteristics previously analyzed, to avoid misleading predictions. In conclusion, preserving and using the real predictive capabilities of AI techniques requieres training and evaluating MLMoCNs with real empirical data on interactions and agents involved, or at least generating hybrid data based on real empirical data.

Bibliography

Beaulieu-Jones, B. K., Wu, Z. S., Williams, C., Lee, R., Bhavnani, S. P., Byrd, J. B., & Greene, C. S. (2019). Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing. *Circulation: Cardiovascular Quality and Outcomes*, 1-10.

Cano-Melani, J. A., Salcedo-Albaran, E., & Garay-Salamanca, L. J. (2022). *A model for evaluating AI generated network graphs .* St. Petersburg: SciVortex Corp.

Cho, J., Lee, K., Shin, E., Choy, G., & Do, S. (2016). *How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?* Conference paper at ICLR 2016, Massachusetts General Hospital and Harvard Medical School.

Garay Salamanca, L. J., & Salcedo-Albarán, E. (2012). *Narcotráfico, Corrupción y Estados: Cómo las redes ilícitas han reconfigurado las instituciones de Colombia, Guatemala y México.* Ciudad de México: Random House Mondadori.

Garay-Salamanca, L. J., & Salcedo-Albarán, E. (2012a, March). Institutional Impact of Criminal Networks in Colombia and Mexico. *Crime, Law and Social Change, 57*(2), 177-194.

Goodfellow, I. J., Pouget-Abadie, j., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). *Generative Adversarial Nets.* arXiv:1406.2661.

Jordon, J., Wilson, A., & van der Schaar, M. (2020). *Synthetic Data: Opening the data floodgates to enable faster, more directed development of machine learning methods.* arXiv.

Jordon, J., Yoon, J., & van der Schaar, M. (2019). *pate-gan: generating synthetic data with differential privacy guarantees .* ICLR.

Morselli, C. (2008). *Inside Criminal Networks.* Montreal: Springer.

Morselli, C. (2012). Assessing network patterns in illegal firearm markets. *Crime Law Soc Change*, 129-149.

O'Shea, K., & Nash, R. (2015). *An Introduction to Convolutional Neural Networks.* Aberystwyth University - arXiv.