Unleashing the Power of Industrial Big Data through Scalable Manual Labeling

Bruno P. Leao
Siemens Technology
Princeton, NJ 08540
bruno.leao@siemens.com

Dmitriy Fradkin
Siemens Technology
Princeton, NJ 08540
dmitriy.fradkin@siemens.com

Tu Lan
Southern Methodist University
Dallas, TX 75205
tlan@smu.edu

Jianhui Wang
Southern Methodist University
Dallas, TX 75205
jianhui@smu.edu

Abstract

Big Data plays a central role in the remarkable results achieved by Machine Learning (ML) and especially Deep Learning (DL) in the recent years. However, the difficulty in obtaining a reasonable amount of labeled samples limits ML/DL application in various domains, including industrial equipment and system monitoring. In this paper the need for methods that turn manual labeling into a scalable process is highlighted. A real world problem is analyzed for which weak supervision methods, successfully employed in other domains, did not produce acceptable results. An alternative approach based on clustering ensembles is described and tested, achieving good performance.

1 Introduction

Big Data (BD) has played a central role in the recent advances achieved by Machine Learning (ML), especially when considering Deep Learning (DL) [Banko and Brill, 2001, Halevy et al., 2009, Stonebraker and Rezig, 2019]. Not surprisingly, the fields where DL is more successfully applied for supervised learning tasks are those where a large amount of labeled data is available such as image analytics [Sun et al., 2017]. Nevertheless, labeling BD can be considered a major bottleneck in application of DL in various domains [Stonebraker and Rezig, 2019]. Approaches such as semi-supervised learning [Ouali et al., 2020] may help in dealing with a limited number of labels in DL applications, but given a sufficiently complex task and a model of corresponding capacity, more data will in general be the key for improved performance.

When considering tasks such as detection of objects based on image data, domain knowledge required for producing labels as well as access to proper data may be widespread. In such a situation, crowdsourcing may be a good option for obtaining large amounts of labeled samples [Whang and Lee, 2020]. However, that is not the case in many practical applications. Stonebraker and Rezig [2019] present a ML application where the data set comprised 30TB of electroencephalogram data. Such data had to be labeled by trained experts (physicians) which is a much less scalable process than the crowdsourcing for object detection. The focus of the work presented here is on industrial problems where the data labeling process may be considered even less scalable than in the cases above. Such problems present the following key characteristics:

- Multivariate time series of sensor data are used to monitor equipment, systems or processes to identify the occurrence of specific relevant events or states, such as a system health state.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

- Complexity of the problem and the volume of existing sensor data is such that the use of DL is justified, however an adequate volume of labeled data is not available.

This type of application will be referred to hereafter as Big Industrial Data Monitoring (BIDM) for simplicity. It is straightforward to notice that with the ever increasing availability of data driven by the Industry 4.0 [Khan et al., 2017], and the challenges which continue to be faced by companies in terms of labeling such data [Fredriksson et al., 2020], more and more applications will fall into this category. The scalability of the labeling process for BIDM is usually even worse than that of a medical application due to the reduced number of experts who could support it. While a large number of physicians around the world are able to interpret EEG signals, this is usually not the case in industrial processes. Even when a certain company has multiple plants that yield the same product, each of those plants is usually unique in terms of installed equipment and sensors, data acquisition and storage. In such cases it is not unusual that only very few people (or maybe a single person) would be able to adequately interpret historical data from one of those plants to label it for ML purposes. It must be taken into account that many times some sort of logging is available in the BIDM context, providing information about historical events of interest. However, as those were in general not recorded with the purpose of training ML models, they are usually not adequate for the task (e.g. Leao et al. [2020]).

Manually labeling individual samples may be sufficient for standard industrial ML applications, but not for BIDM. In order to make DL use realistic in this context, it is of utmost importance that the information from the subject matter expert (SME) is leveraged to the greatest extent possible. In this work such goal is considered equivalent to maximizing the scalability of the manual labeling process, i.e. maximizing the number of labeled samples produced by a limited effort available from the SME. This is subject to the assumption that labels are of good quality for the task, in the sense that an application that can benefit from more labels can successfully employ those yielded by the scalable process to improve performance. Solutions such as active learning [Settles, 2009] can help in obtaining better labels, but they do not improve scalability. More recently, weak supervision (WS) has been proposed as an approach to solve the scalability problem, achieving good results in a variety of practical applications [Bach et al., 2019]. However, it also has limitations as illustrated here based on a failed attempt to apply it to a real world BIDM problem. An alternative approach is proposed in this work based on clustering ensemble methods. Using clustering to support data labeling is certainly not a new idea. It is used, for instance, in the EEG case mentioned above. The contribution described here is in the use of ensembles specifically for scaling of the manual labeling process, including the interpretation of the model results as a label confidence metric that is used to guide downstream decisions. An application based on a real world BIDM problem is presented to illustrate the concepts. It comprises an unprecedentedly large dataset of measurements from power system sensors (Phasor Measurement Units - PMUs) covering most of the U.S. territory with duration up to two years. The application of the proposed method as well as the attempt at applying WS to the same problem are described.

The remainder of the paper is organized as follows: in section 2 the BIDM problem under consideration is described; methods employed for scaling manual labeling and results of their application are presented in section 3; section 4 is the conclusion. Given the limited space, technical discussions are summarized. Referenced works provide further details [Siemens Corporation, 2021, Leao et al., 2020, Lan et al., 2021].

2 BIDM Problem and Data

The BIDM problem under consideration consists of detection of relevant events that impact power grid operation based on PMU data. The dataset employed for the analysis, which is unprecedentedly large for this type of application, was collected by the U.S. Department of Energy and Pacific Northwest National Laboratory and integrates data from multiple utilities. It consists of $\sim$26TB of compressed time series data containing measurements of three-phase frequency, currents and voltages from 446 locations spreading over the three major interconnections (ICs) of the U.S. territory: Eastern, Western and Texas. The dataset from each IC is independent from the others and each one corresponds to 13 to 24 months of operation. Sampling rate is either 30 or 60Hz, depending on the PMU.

Event logs have been provided for each IC along with the PMU data, also based on integration of information from multiple utilities. However, they had not been originally created with the purpose

of supporting ML development and early in the analysis it was considered that they could not be adequately used as labels. Main reasons for that are:

- Frequent mislabeling: logs indicating events not seen in the data or events seen in the data but not in logs.
- No localization information: even when the log entries corresponded to events in the data, such event usually only affected some of the PMUs. However logs did not provide any association to specific PMUs.
- Mismatch between event categories and patterns in the data: The same log category could correspond to multiple different patterns in the data and the same pattern to multiple log categories.

Based on the above, it was decided that a data-centric AI approach would be taken as obtaining proper labels was considered much more promising for achieving good results than working on development of more advanced modeling approaches. Leao et al. [2020], Siemens Corporation [2021] provide more details about the datasets and event logs, including data quality issues and pre-processing.

3 Scalable Data Labeling Methods and Results

3.1 Weak Supervision

Given the recent successful applications of WS in various contexts [Bach et al., 2019] it was the initial approach tried for making the manual labeling process scalable. Given the large data volume and the dynamic nature of the BIDM problem, Flying Squid tool [Fu et al., 2020] was employed for this task due to its superior computational performance compared to other WS approaches, results in the form of label probabilities and the possibilities of defining dependencies among labels.

Three sources of WS rules were used:

- Based on mapping of event logs to specific types of events of interest. This mapping was performed by a SME based only on the available descriptions. This resulted in 4 rules.
- Based on a SME's analysis of what the patterns in the data should look like for different types of events. This produced 7 rules.
- Based on industry standards associated to the data such as IEEE [2019]. This yielded 6 rules.

The definition of which subset of the rules would be applicable to each type of event of interest have also been performed by a SME.

Those rules were applied to the data from the West IC. The use of dependencies among labels resulted in a large penalty in computational performance and did not result in relevant changes in results based on tests with small subsets of the data, therefore they were not employed. SMEs have manually labeled events happening in a small portion of the data, corresponding to a few days, so that a sanity check could be performed. The event categories for which we had the largest number of labels from this process, corresponding to *short circuit* and *trip with no short circuit*, were employed for this verification. WS results achieved an AUROC below 0.6, indicating that the performance was not much better than random guessing in this case.

3.2 Clustering Ensemble

The steps associated with the clustering ensemble approach are described below. Prior to those, pre-processing and a coarse filtering of the data have been performed. Lan et al. [2021], Siemens Corporation [2021] present a more detailed description.

1. Calculating various features from each sample corresponding to a 10s multivariate time window. Up to 152 features have been extracted from each sample, including basic statistics such as mean and median, temporal characteristics such as slope and number of turning points, and frequency domain metrics such as fundamental frequency. This step optionally also included sample pre-filtering based on DBSCAN for detecting clear matches to defined categories and excluding them from further processing.

2. Training multiple clustering models based on resulting features. Two configurations of the clustering ensemble have been tried: (1) homogeneous ensembles based on K-means models with different numbers of clusters; (2) heterogeneous ensemble combining K-means, Hierarchical Clustering and Gaussian Mixture Models.

3. Manually labeling each resulting cluster. A few different categorizations of events have been tried during the analysis, but ultimately 10 categories were used to manually label each cluster, corresponding to various patterns found in the data as analyzed by SMEs. Multiple categories could be associated with the same cluster.

The method is defined as follows. Given features associated to N_s samples, N_m clustering models are trained, each one with $n_k, k = 1, 2, ..., N_m$, clusters. Therefore, each sample $j = 1, 2, ..., N_s$ gets associated to N_m clusters. Each of the resulting clusters is manually labeled based on N_c categories. A confidence metric $\alpha_{i,j}$ associated with sample j belonging to category $i = 1, 2, ..., N_c$ is defined as $\alpha_{i,j} = (1/N_m) \sum_k I_{k,i,j}$ where $I_{k,i,j}$ is an indicator function which is equal to 1 if sample j is associated to category i in model k and 0 otherwise. Therefore, $\sum_i \alpha_{i,j} = 1 \, \forall j$ and the most conservative selection of samples for category i in terms of avoiding mislabeling is performed by selecting the subset $S_i^p | \alpha_{i,j} = 1 \, \forall j \in S_i^p$ as positive samples and the subset $S_i^n | \alpha_{i,j} = 0 \, \forall j \in S_i^n$ as negative ones.

Employing such method was key to obtaining a reasonable number of good quality labels for multiple event categories. Hundreds of thousands of labels for each event category could be generated based on a few hundreds of labeled clusters. The final set of labels has been considered of good quality based on the same sanity checks performed for WS and additional inspection of sample results by SMEs. They were then used to train deep semi-supervised learning (DSSL) models for detection of events, in order to leverage also the unlabeled data. DSSL models were based on autoencoder architectures using 1-D convolutional layers and fully connected ones, with an additional head appended to the encoder for binary classification of events. Due to the limitations in space, details are presented only for a single model for illustration purposes. This model was trained for detection of *loss of generation* events. 160 clusters have been labeled in an ensemble of 3 clustering models. This was the manual labeling effort required for obtaining all labels, not only the ones in this category. A total of 314050 labeled samples were obtained for the *loss of generation* category employing the most conservative selection approach described above. From those, 11226 were positive ones. Trained DSSL model performance was assessed based on held out data achieving 99.87% precision and 99.96% recall. As additional validation, known events of this type reported by NERC [NERC, 2017, NERC and WECC, 2018] have been correctly identified by this model with no false positives from models associated to other categories.

4 Conclusion

This work discusses the challenges associated with proper use of BD in industrial applications, given the limitations in current manual labeling processes. Identifying the occurrence of specific types of events and states in equipment or systems is one example of a broad category of use cases with ubiquitous applicability and high value to industry for which implementation is limited due to such challenges. As Industry 4.0 paradigms are producing more and more data, BIDM applications, as defined above, become increasingly common.

Existing tools for using BD with a reduced number of labels or improving the information aggregated by labels, such as semi-supervised learning or active learning, can help, but having a larger number of labeled samples provides much broader possibilities as currently verified in the image analytics domain. Therefore, tools are needed to make the manual labeling process scalable. Weak supervision is becoming increasingly popular for this purpose, but it also has its limitations. A failed attempt to apply WS to a real world BIDM problem was presented as well as an alternative approach based on clustering ensembles which has provided good results, achieving scalability and producing means for evaluating the confidence associated with labels. Future work should investigate the limitations of the proposed approach and compare it further to WS methods or other alternatives in the context of additional BIDM applications.

Acknowledgments and Disclosure of Funding

This material is based upon work supported by the Department of Energy under Award Number DE-OE0000917. This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

References

S. H. Bach, D. Rodriguez, Y. Liu, C. Luo, H. Shao, C. Xia, S. Sen, A. Ratner, B. Hancock, H. Alborzi, et al. Snorkel drybell: A case study in deploying weak supervision at industrial scale. In *Proceedings of the 2019 International Conference on Management of Data*, pages 362–375, 2019.

M. Banko and E. Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pages 26–33, 2001.

T. Fredriksson, D. I. Mattos, J. Bosch, and H. H. Olsson. Data labeling: an empirical investigation into industrial challenges and mitigation strategies. In *International Conference on Product-Focused Software Process Improvement*, pages 202–216. Springer, 2020.

D. Fu, M. Chen, F. Sala, S. Hooper, K. Fatahalian, and C. Ré. Fast and three-rious: Speeding up weak supervision with triplet methods. In *International Conference on Machine Learning*, pages 3280–3291. PMLR, 2020.

A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2): 8–12, 2009.

IEEE. Recommended Practice for Monitoring Electric Power Quality. *IEEE Std. 1159-2019*, pages 1–98, 2019. doi: 10.1109/IEEESTD.2019.8796486.

M. Khan, X. Wu, X. Xu, and W. Dou. Big data challenges and opportunities in the hype of industry 4.0. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2017.

T. Lan, Y. Lin, J. Wang, B. P. Leao, and D. Fradkin. Unsupervised power system event detection and classification using unlabeled pmu data. In *2021 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*. IEEE, 2021.

B. P. Leao, D. Fradkin, Y. Wang, and S. Suresh. Big data processing for power grid event detection. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 4128–4136. IEEE, 2020.

NERC. 1,200 MW fault induced solar photovoltaic resource interruption disturbance report: Southern California 8/16/2016 event. Technical report, June 2017.

NERC and WECC. 900 MW fault induced solar photovoltaic resource interruption disturbance report: Southern California event october 9, 2017. Technical report, February 2018.

Y. Ouali, C. Hudelot, and M. Tami. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*, 2020.

B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Siemens Corporation. MindSynchro project - training and test dataset report, US DOE OE FOA 1861, agreement DE-OE0000917. to be published by the US DOE, 2021.

M. Stonebraker and E. K. Rezig. Machine learning and big data: What is important? *IEEE Data Eng. Bull.*, 42 (4):3–7, 2019.

C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

S. E. Whang and J.-G. Lee. Data collection and quality challenges for deep learning. *Proceedings of the VLDB Endowment*, 13(12):3429–3432, 2020.