Active Learning from Unreliable Data

Zilong Zhao, Sophie Cerf, Robert Birke, Bogdan Robu, Sara Bouchenak, Sonia Ben Mokhtar, Lydia Chen

► **To cite this version:**

Zilong Zhao, Sophie Cerf, Robert Birke, Bogdan Robu, Sara Bouchenak, et al.. Active Learning from Unreliable Data. 13th EuroSys Doctoral Workshop (EuroDW 2019), Mar 2019, Dresde, Germany. hal-02045455

HAL Id: hal-02045455

http://hal.univ-grenoble-alpes.fr/hal-02045455

Submitted on 22 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Active Learning from Unreliable Data

Zilong Zhao*, Sophie Cerf*, Robert Birke†, Bogdan Robu*, Sara Bouchenak‡, Sonia Ben Mokhtar‡, Lydia Y. Chen§

*Univ. Grenoble Alpes, France
{zilong.zhao, sophie.cerf, bogdan.robu}@gipsa-lab.fr
†ABB Research, Switzerland
{robert.birke}@ch.abb.com
‡INSA Lyon, France
{sara.bouchenak, sonia.benmokhtar}@insa-lyon.fr
§TU Delft, Netherlands
{y.chen-10}@tudelft.nl

***Abstract*—Classification algorithms have been widely adopted in big recommendation systems, e.g., products, images and advertisements, under the common assumption that the data source is clean, i.e., features and labels are correctly set. However, data collected from the field can be unreliable due to careless annotations or malicious data transformation. In our previous work, we proposed a two-layer learning framework for continuous learning in the presence of unreliable anomaly labels, it worked perfectly for two use cases, (i) detecting 10 classes of IoT attacks and (ii) predicting 4 classes of task failures of big data jobs. To continue this study, now we will challenge our framework with image dataset.**

The first layer of quality model filters the suspicious data, where the second layer of classification model predicts data instance's class. As we focus on the case of images, we will use widely studied datasets: MNIST, Cifar10, Cifar100 and ImageNet. Deep Neural Network (DNN) has demonstrated excellent performances in solving images classification problems, we will show that two collaborating DNN could construct a more robust and high accuracy model.

***Index Terms*—Unreliable Data; Images; Attacks; Machine Learning; Deep Neural Network**

I. Introduction

A large amount of user-generated data powers up machine learning based applications in our daily life. Labeled data (especially for images) from search engine provides a fast and cheap way to build the large-scale dataset. But it also inevitably introduces some incorrect labels. For examples, if we want to construct an image dataset of "airplane", with keyword "airplane", Google Image can give us a very good first result. But we can also notice that there are not only airplane images, they have also images: airplane interior, airplane runway, airplane structure sketch and Cartoon airplane.

To solve this problem, a simple way to clean the data, it's to find a domain expert to remove or relabel the suspect data in a preprocessing stage. However large-scale annotated datasets with high-quality label annotations are not always available for new domain, due to the significant time and efforts it takes, not to mention that for some online-tuning systems, the user-generated data can be infinite. There is no doubt that DNN is delivering superior results on image classification, but this success is highly tied to the availability of large-scale annotated datasets. So we can see a big contradiction here.

Standard machine learning algorithms typically assume clean labels and overlook the risk of noisy labels. But recent studies have shown that learning from high proportion noisy labels can significantly degrade the DNN's classification accuracy [9].

Existing works of learning from noisy data can be roughly divided into two aspects: (1) filtering out noisy labels and learning only from the predicted clean data, (2) designing noise-aware classification algorithms. The first builds one or multiple filter models, e.g., SVM [8] to clean the data, and only the data instances that predicted label by one or more filters matches its original label, we believe this data instance is clean and can be used to train the classification model. The second type of approach can be summarized by several works, one proposed method is to correct noisy labels to their true labels via a clean label inference step. These methods assume the availability of a small clean dataset to be used [3], [10]. A different approach is to learn directly through noisy data, and in parallel, running a local intrinsic Dimensionality [2] measurement to monitor the stage of training process, to make sure at the end, the test accuracy stabilizes at its highest level [4].

While the recent state-of-the-art solution can have a very good result on noisy label issues, little focus has been given to the online setting with data instances having fluctuating noise levels. Our study presents the initial results on how to build a classifier by selectively and continuously learn from high quality data that leads to a strong classifier.

Figure 1 shows training and predicting process of our system. Training process is triggered by $\mathcal{D}_i$: ith training batch, the quality model will predict labels for each data instance, if the predicted label matches its original label, we believe this data instance is clean, and will pass it to train classification model, these "clean" data will also be used to train quality model itself, if the predicted label doesn't match its original label, we will throw the data instance. From $\mathcal{P}_i$ to $\hat{Y}_i^{\mathcal{P}}$ represents predicting process, only classification model is involved in this process.

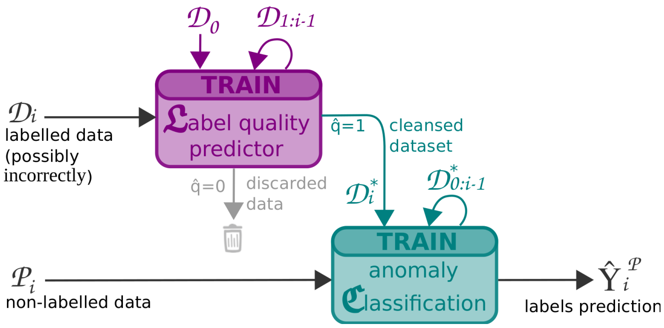
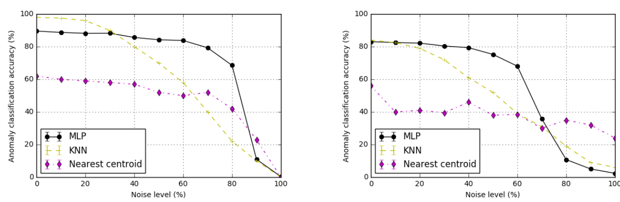


Fig. 1: System framework. The training process is triggered by every batch of $\mathcal{D}_i$. Each bloc represents a machine learning algorithm, the colored arrows on top of bloc represent the coming data used to train the algorithm. $\mathcal{P}_i$ represents a batch of data need to be predicted.



(a) Use case of IoT thermostat device attacks (b) Use case of Cluster task failures

Fig. 2: Impact of noisy level on anomaly classification accuracy

II. Motivating case studies

As the system for image classification is on developing, we list two study cases from our previous work, to demonstrate the impact of noisy data on anomaly classification detection.

- Detecting **IoT device attacks** from inspecting network traffic data collected from commercial IoT devices [5]. This dataset contains nine types of IoT devices which are subject to ten types of attacks. Specifically, we focus on the Ecobee thermostat device that may be infected by Mirai malware and BASHLITE malware. Here we focus on the scenario of detecting and differentiating between ten attacks. It is important to detect those attacks with high accuracy against all load conditions and data quality.
- Predicting **task execution failures** for big data jobs running at Google cluster [6], [7]. This trace contains a month-long jobs execution record from Google clusters. Each job contains multiple tasks, which can be terminated into four different states: *finish*, *fail*, *evict*, or *kill*. The last three states are considered as anomaly states. To minimise the computational resource waste due to anomaly states, it is imperative to predict the final execution state of task upon their arrivals.

Iot and Cluster datasets have perspectively 115 and 27 features, 11 and 4 classes . To detect anomalies in each case, related studies have applied machine learning classification algorithms, e.g., k-nearest neighbor (KNN), nearest centroid and multilayer perceptron (MLP) (a.k.a feed-forward deep neural networks), under the scenario where different levels of label noise are present. Here, we evaluate how the detection accuracy changes relative to different levels of noises. We focus on offline scenarios where classification models are learned from 14000 records and evaluated on a clean testing dataset of 6000 records. We specifically apply KNN, nearest centroid and MLP on IoT device attacks and cluster task failures respectively, and summarize the accuracy results in Figure 2a and Figure 2b.

One can see that noisy labels clearly deteriorate the detection results for both IoT attacks and task failures, across all three classification algorithms. For standard classifiers, like KNN and nearest centroid, the detection accuracy decays faster than MLP that is more robust to the noisy labels. Such an observation holds for both uses cases. In IoT attacks, MLP can even achieve a similar accuracy as the case of no label noises, when there is 50 percent of label classes are altered.

III. Perspective

From our new study result, image dataset like cifar-10 (a natural image data set with 10 categories [1]) can also follow the trend with variation of noise level as in Figure 2, the only difference is for image data, the performance of KNN and Nearest centroid algorithms are much worse, only using Convolutional Neural Network (CNN) with specific configuration could maintain the curve as MLP in Figure 2.

From our current result on implementing our two-layer system on Iot and Cluster dataset, we have concluded that the first layer quality model's accuracy is very important for online learning scenario. In our framework, the quality model plays the role of domain expert as we mentioned in section I. The ongoing work is to find out what is the minimum initial clean images to learn, so that we can have a reasonable accuracy for quality model to kick-start our system. To evaluate our system, we need to prepare a testing dataset with all clean data, to see how the classification model's accuracy changes over time.

In a first stage, the result we want to see is that we could use as few as possible initial training data to launch our system, and the classification model's accuracy could converge to the same level just as we train the classification model without noisy data. Secondly, if the result works as we want, we could still make it better by accelerating the converging speed of classification model. As we described in section I, the data instance, which is predicted as not clean by quality model, is thrown away. That decreases the total training data used for classification model. Quality model is not 100% correct, and its accuracy could also increase over time, we should give a second chance to these thrown data instances in later batches. Furthermore, as considering the problem of overfitting, passing only clean data to classification model may not be the best solution, adding a small proportion of noisy data within clean data could also be worth to explore.

References

[1] Krizhevsky Alex and Hinton Geoffrey. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 2009.

[2] Michael E. Houle. Local intrinsic dimensionality I: an extreme-value-theoretic foundation for similarity applications. In Christian Beecks, Felix Borutta, Peer Kröger, and Thomas Seidl, editors, *Similarity Search and Applications - 10th International Conference, SISAP 2017, Munich, Germany, October 4-6, 2017, Proceedings*, volume 10609 of *Lecture Notes in Computer Science*, pages 64–79. Springer, 2017.

[3] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1928–1936. IEEE Computer Society, 2017.

[4] Xingjun Ma, Yisen Wang, Michael E. Houle, Shuo Zhou, Sarah M. Erfani, Shu-Tao Xia, Sudanthi N. R. Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *JMLR Workshop and Conference Proceedings*, pages 3361–3370. JMLR.org, 2018.

[5] Yair Meidan, Michael Bohadana, Yael Mathov, Yisroel Mirsky, Asaf Shabtai, Dominik Breitenbacher, and Yuval Elovici. N-baiotnetwork-based detection of iot botnet attacks using deep autoencoders. *IEEE Pervasive Computing*, 17(3):12–22, 2018.

[6] Charles Reiss, John Wilkes, and Joseph L Hellerstein. Google cluster-usage traces: format+ schema. *Google Inc., White Paper*, pages 1–14, 2011.

[7] Andrea Rosà, Lydia Y. Chen, and Walter Binder. Failure analysis and prediction for big-data systems. *IEEE Trans. Services Computing*, 10(6):984–998, 2017.

[8] Nicola Segata, Enrico Blanzieri, Sarah Jane Delany, and Padraig Cunningham. Noise reduction for instance-based learning with a local maximal margin approach. *J. Intell. Inf. Syst.*, 35(2):301–331, 2010.

[9] Vadim N. Vagin and Marina V. Fomina. Problem of knowledge discovery in noisy databases. *Int. J. Machine Learning & Cybernetics*, 2(3):135–145, 2011.

[10] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5601–5610, 2017.