Cover sheet for DSAA 2020 tutorial proposal

Label noise: Problems and solutions

Presenter details

Sandhya Tripathi
Affiliation: PhD student, IIT Bombay & Visiting researcher, Washington University at St. Louis
Address: PhD lab, IEOR, IIT Bombay, Mumbai 400076, India
Email Id: sandhya.tripathi@iitb.ac.in; sandhyat@wustl.edu

N. Hemachandra
Affiliation: Professor, IEOR, IIT Bombay
Address: Room 205, IEOR, IIT Bombay, Mumbai 400076, India
Email Id: nh@iitb.ac.in
Phone: +91-22-25767885

Label noise: Problems and solutions

DSAA 2020 Tutorial Proposal *

Sandhya Tripathi & N. Hemachandra

Abstract

With the data coming from various sources and many of them being unreliable, users are presented with the problem of developing robust algorithms or modifying the working of existing algorithms so that they become label noise robust. We would be presenting a study on various aspects of label noise problems and solutions. We would start with the problem setting of standard 0-1 binary classification and outline some historical background. There will be three strands of literature which will be presented: 1) Identification of label noise robust loss functions and corresponding sufficient conditions. We would also discuss about various modifications proposed to surrogate loss functions like unbiasedness, importance re-weighting, etc. These methods are theoretically sound and demonstrate good empirical performance. However, they either assume true noise rates or estimate them, which might not be accurate. The noise considered in this setup is either symmetric across all classes or it is only dependent on class. Also, the setting is empirical risk minimization (ERM) framework. 2) In this part, we generalize the noise setting to instance dependent noise and ERM setup to deep networks. In instance dependent noise case, a major result available is the noise robustness of Isotron algorithm. Another approach taken is via consistency, where k-NN and SVM are shown to be noise robust. With respect to deep networks, various perspectives have been proposed in addition to the existing way of showing inherent robustness of loss function and modification of loss functions. One approach here is to understand the basic working of neural networks in terms of dimensionality and modify the algorithm. Another approach assumes that there is a small amount of clean data available which can be used along with noisy data for training a deep network. 3) In the last part, we will understand the effect of label noise on cases beyond standard classifications like cost sensitive binary classification, generative models, learning from positive and unlabelled data and active learning. For each of these cases, we will briefly explain the problems and the efforts made to solve them in terms of loss functions or the learning approach. Finally, we will conclude by discussing some issues while designing and evaluating the label noise robust algorithms.

Objective: Our objective in this tutorial is to provide details about how this problem of label noise is formulated, in particular in classification and guide across the huge literature accumulated in less than a decade of the label noise problem's revival.

Category: Computer Science

Tutorial structure

The 3 hour slot would be divided as follows:
1.5 hours lecture + 15 minutes break + 45 mins lecture + 30 mins lecture.
The outline is as follows:

*Most of Part 1 and some topics of Part 3 has been taught as part of a regular course at IIT Bombay for the last couple of years.

1. Part 1: Overview and Label noise robustness of loss functions (1.5 hour slot)
    - Motivation and framework.
    - Various methods to deal with label noise problems and our focus.
    - History and the negative result leading to revival of the area ([2, 16]).
    - Details about the two approaches of identifying label noise robust loss functions and modification of loss functions to make them label noise robust compliant ([7, 31, 21]).
    - Some label noise robust loss functions for training deep networks ([1, 34, 33, 32]).
    - Non-conventional approaches to learn in the presence of label noise ([18, 22, 15, 35, 13]).
    - Application driven label noise robustness models ([8]).
    - Use of true noise rates vs noise estimates vs cross validation over noise rates?
2. Part 2: Tailor made approaches for making deep networks label noise robust (45 mins slot)
    - Moving from class dependent label noise to instance dependent label noise.
    - Robustness of Isotron algorithm ([19]).
    - Results from [3, 5] that use consistency arguments to show the robustness of some existing classification algorithms.
    - Learning from label noise using deep networks: identifying label noise robust loss functions to be used in the neural networks ([6]), modification of loss functions ([23]).
    - Method based on the fundamental working of a neural network ([17, 9, 4, 12]).
    - Using a small amount of clean data to improve performance ([10]).
3. Part 3: Learning is more than just 0-1 classification! (1/2 hour slot)
    - Where else does label noise affect?
    - Generative models: Robustness of conditional GANs ([28]).
    - Positive Unlabelled data. Two scenarios: when the PU data has label noise ([11]), when a PU problem is modelled as a label noise problem ([27]).
    - Label noise in active learning setup ([14]).
    - Learning from noisy labelled data when the mis-classifications are differentially costed: identification of robust loss function ([30]), modification of loss function ([20]), using a clean and balanced data ([26]).
    - Some important points to note while implementation and what's ahead? Revival of categorical attribute noise problem ([24])?

Description

Consider the classical classification setup where $\mathcal{X} \subset \mathbb{R}^n$ is the feature space and $\mathcal{Y}$ is the set of labels. The joint data distribution is $\mathbf{X} \times Y$ where $\mathbf{X} \in \mathcal{X},\ Y \in \mathcal{Y}$. Label noise can be symmetric (uniform) or non-uniform. Symmetric label noise (SLN) considers the cases where the probability (noise rate $\rho_{\mathbf{x}}$) with which the label of data point $\mathbf{x} \in \mathbf{X}$ has been flipped is same for all data points irrespective of class, i.e., $\rho_{\mathbf{x}} = \rho$. Class conditional label noise (CCN), a special case of non-uniform noise, is when the noise rates depend only on the class, i.e., $\rho_{\mathbf{x}} = \rho_y,\ y \in \mathcal{Y}$. Non-uniform noise is the most general case where $\rho_{\mathbf{x}}$ could be any arbitrary function of $\mathbf{x}$. Methods dealing

with label noise can be broadly classified into four categories: (1) Cleaning noisy labelled points (2) Elimination of noisy labelled points (3) Heuristics to provide noise tolerance (4) Noise tolerant algorithms. Category (1) and (2) consists of schemes like k-nearest neighbours, outlier detection techniques, etc. We would like to focus on category (3) and (4). Broadly, we would like to give details of current research in the following three topics.

Label noise robust loss functions The problem of learning from noisy examples was first studied by [2] where random classification noise (RCN $\sim$ SLN) model is proposed and theoretical aspects of learnability are considered. It was shown that finiteness of VC dimension and Littlestone dimension continue to characterize learnability even in presence of random noise. Research on label noise problems revived when [16] provided a negative result that any method based on convex potential function are not SLN robust.

The goal in this whole exercise is to provide label noise robust learners. The robustness interpretation is as follows: clean risk of a learner learnt from clean data and noisy data should be equal. One has to achieve this goal by circumventing the negative result of [16]. One direction of work is to identify the loss functions which are not convex potentials and are label noise robust. A sufficient condition for SLN robustness of loss functions is provided by [7, 6]. Loss functions like, 0-1 loss, squared loss, unhinged loss [31], categorical cross entropy for neural networks, etc. are not convex potential and satisfy the sufficient condition. As the robustness is inherent in the loss functions, knowledge of true noise rates or their estimates is not required. For the deep learning framework, [1] propose a theoretically sound SLN robust loss function called bi-tempered loss function that is a non-convex generalization of logistic loss and requires tuning two temperature parameters. [34] propose a loss function, viz., generalized cross entropy loss (GCE) that is shown to be asymptotically robust to SLN uniformly and CCN under some conditions. Also, [33] propose a loss function, viz., L_{DMI} that is invariant to the type and level of noise. [32] shows that symmetric cross entropy is approximately robust to SLN.

Another direction is to modify the loss function, e.g., [21] provide an unbiased estimate of the loss function with the assumption that noise rates are known. This assumption is rarely satisfied and hence one has to resort to estimating the noise rates. Theoretically grounded estimates of noise rates involve estimates of in-class probability, $P(\tilde{Y} = \tilde{y}|X)$ as used by [18] and [15]. Historically, density estimation problem is hard and hence getting accurate estimates of noise rate is a difficult problem. Instead of assuming the knowledge of true noise rates or estimating them, another option is to cross validate over the noise rates as used by [22]. In the regime of loss modification, a special case is of [15], where they use the concept of importance re-weighting to learn from CCN corrupted data. A different perspective is provided in [22] by proposing linear odd loss functions, which are approximately CCN robust. A probabilistic approach is proposed by [35] where the noisy data is grouped into various states of a Markov chain. Approximate stationary distribution of this Markov chain is then used to get the final classification model. This approach doesn't require the knowledge of noise rates. [13] propose an indirect learning method that uses complementary labels and filters out noisy data.

Tailor made approaches for making deep networks label noise robust When the noise is instance dependent, [19] provides consistency results on risk minimization and area under Receiver Operating Curve (ROC) maximization. In addition, if the clean distribution has a structure then, Isotron algorithm is shown to be robust to boundary-consistent noise. Consistency related results in the presence of instance dependent noise are provided by [5] and [3]. It is shown in [3] how k-Nearest neighbour (k-NN) and Support Vector Machines (SVM) are robust to instance and label dependent noise and Linear Discriminant Analysis (LDA) is not, when criteria is whether the convergence rates of the excess risks is same or not.

The state-of-the art algorithms in one way or the other are dependent on deep learning and

hence, different approaches have been proposed for learning with deep neural networks in the presence of label noise. Two algorithms by correcting the loss function for label noise are proposed by [23]. Another recent work by [10] uses a small amount of trusted (clean) data while training the deep network. Even though these methods are theoretically grounded, they require noise estimates. After understanding the basic difference in working of deep neural networks with clean labels and noisy labels, [17] provides a dimensionality driven learning algorithm which do not require rate estimates. In another attempt to robustly train deep neural networks, a co-teaching approach using two communicating neural networks is proposed by [9]. [4] propose a relation between noisy data test accuracy and noise matrix.

Learning is more than just 0-1 loss based binary classification! Label noise can create a problem for learning scenarios other than binary classification, for e.g., in conditional Generative Adversarial Networks (c-GANs), in Positive-Unlabelled (PU) framework, active learning, transfer learning, when the mis-classifications are differentially costed, etc. Two variants of c-GANs by modifying the loss function of the generator are proposed by [28] and are also theoretically shown to be label noise robust. Basic GANs can be used to first clean the noisy data and then use an off the shelf algorithm to design a label noise robust classifier [29]. If the PU data available has label noise in positive labels, [11] develop a label noise robust algorithm that estimates the posterior distribution. The idea is to estimate the class prior by transforming the full dimensional input space to a univariate space and hence avoid the curse of dimensionality in density estimation. A PU learning problem can be modelled as a label noise problem by assigning negative label to all the unlabelled examples as in [27]. The performance of algorithms based on such modelling is shown to be better than existing cost sensitive learning based algorithms for PU learning. Loss functions which are noise robust in the setup of active learning are identified in [14]. Along with this, they propose algorithms based on importance re-weighting and novel approaches for noise estimates.

In a classification problem, noisy labels aggravates the difficulty in learning if the data has class imbalance or there is an inherent need to have different costs for false positive and false negative. In [20], it is assumed that data can decide the differential costing by tuning over a parameter. In this setting, they provide two loss correction based algorithms where they cross validate over the CCN noise rates. Another work where the data has class imbalance and domain cost for different classes in presence of SLN is by [30]. The two proposed schemes come under the category of identifying the inherent label noise robustness and do not need the knowledge of noise rates. In [26] it is assumed that some clean balanced validation is always available and then reweighs the noisy and imbalanced training examples to make the scheme label noise robust.

Synthetic data plays a significant role in the evaluation of various algorithms proposed for learning in the presence of label noise. We will outline some schemes available for generating it for our purposes. Research in label noise problems has been growing at a rapid pace due to the real life scenarios it is able to model. A recent survey by [25] gives an exhaustive view of the current situation.

Audience estimate: As label noise is ubiquitous due to plenty of not-so-good quality data, it is a relevant topic for data mining and knowledge discovery community. Hence, we estimate the audience size to be at least 100.

Organizer details

Sandhya Tripathi is a PhD candidate in Industrial Engineering and Operations Research, IIT Bombay, India. Her research interests include machine learning, statistical learning theory, problems arising at the intersection of game theory and machine learning. She has presented her work during

conferences like ECQT at Toulouse, France, 2016, WINE at Bengaluru, India, 2017, short courses like MLSS at University of Kyoto in 2015, IRCN Neuro-inspired computation course at University of Tokyo, 2019, and many workshops and meetings held in India. She has published research papers in the area of scalable classifier design after giving a novel interpretation to exponential loss function and cost sensitive label noise robustness at CODS-COMAD, 2018 and PAKDD 2019 respectively.

She has been a Teaching Assistant throughout her PhD duration for a variety of courses like Engineering statistics, Optimization models, Decision Analysis and Game Theory, Computer programming and algorithms, Online Learning. The work involved was to take tutorials, prepare assignments and grading along with other teaching assistants. `http://www.ieor.iitb.ac.in/sandhya` Currently, she is visiting Department of Anesthesiology, Washington University at St Louis to use transfer learning and cost sensitive learning on clinical data (Feb 2020-Oct 2020).

Nandyala Hemachandra is a Professor in Industrial Engineering and Operations Research, IIT Bombay. His Research and Teaching interests include Statistical Learning Theory, Markov Decision Processes, Queuing Theory, Game Theory, etc., along with their applications to resource allocation problems arising from Communication Networks, Supply Chains, Logistics, etc.

Some earlier and ongoing work relevant to this proposal is in the areas of Reinforcement Learning, Two-timescale based simulation based parameter optimization, Design of Classifiers and associated bounds in the PAC-Bayesian framework, Cooperative Game theory approach to classifier design, Change point detection. `https://www.ieor.iitb.ac.in/~nh`

References

[1] Ehsan Amid, Manfred K Warmuth, Rohan Anil, and Tomer Koren. Robust bi-tempered logistic loss based on bregman divergences. In *Advances in Neural Information Processing Systems*, pages 14987–14996, 2019.

[2] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

[3] Timothy I. Cannings, Yingying Fan, and Richard J. Samworth. Classification with imperfect training labels. *https://arxiv.org/abs/1805.11505*, 2018.

[4] Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *International Conference on Machine Learning*, pages 1062–1070, 2019.

[5] Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance and label-dependent label noise. *http://arxiv.org/abs/1709.03768*, 2017.

[6] Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, pages 1919–1925, 2017.

[7] Aritra Ghosh, Naresh Manwani, and PS Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107, 2015.

[8] Sebastian Gündel and Andreas Maier. Epoch-wise label attacks for robustness against label noise. In *Bildverarbeitung für die Medizin 2020*, pages 287–292. Springer, 2020.

[9] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, pages 8536–8546, 2018.

[10] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *Advances in Neural Information Processing Systems*, pages 10477–10486, 2018.

[11] Shantanu Jain, Martha White, and Predrag Radivojac. Estimating the class prior and posterior from noisy positives and unlabeled data. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2693–2701. 2016.

[12] Ishan Jindal, Matthew Nokleby, Daniel Pressel, Xuewen Chen, and Harpreet Singh. Deep neural networks for corrupted labels. In *Deep Learning: Concepts and Architectures*, pages 211–235. Springer, 2020.

[13] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 101–110, 2019.

[14] Jan Kremer, Fei Sha, and Christian Igel. Robust active label correction. In *International Conference on Artificial Intelligence and Statistics*, pages 308–316, 2018.

[15] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2016.

[16] Philip M Long and Rocco A Servedio. Random classification noise defeats all convex potential boosters. *Machine learning*, 78(3):287–304, 2010.

[17] Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah M Erfani, Shu-Tao Xia, Sudanthi Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In *International Conference on Machine Learning*, pages 3361–3370, 2018.

[18] Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*, pages 125–134, 2015.

[19] Aditya Krishna Menon, Brendan van Rooyen, and Nagarajan Natarajan. Learning from binary labels with instance-dependent noise. *Machine Learning*, 107(8-10):1561–1595, 2018.

[20] Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Cost-sensitive learning with noisy labels. *Journal of Machine Learning Research*, 18(155):1–33, 2018.

[21] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, pages 1196–1204, 2013.

[22] Giorgio Patrini, Frank Nielsen, Richard Nock, and Marcello Carioni. Loss factorization, weakly supervised learning and label noise robustness. In *International Conference on Machine Learning*, pages 708–717, 2016.

[23] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2233–2241. IEEE, 2017.

[24] Aditya Petety, Sandhya Tripathi, and N Hemachandra. Attribute noise robust binary classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020.

[25] Ronaldo C Prati, Julián Luengo, and Francisco Herrera. Emerging topics and challenges of learning from noisy data in nonstandard classification: a survey beyond binary class noise. *Knowledge and Information Systems*, pages 1–35, 2018.

[26] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pages 4331–4340, 2018.

[27] Hong Shi, Shaojun Pan, Jian Yang, and Chen Gong. Positive and unlabeled learning via loss decomposition and centroid estimation. In *IJCAI*, pages 2689–2695, 2018.

[28] Kiran K Thekumparampil, Ashish Khetan, Zinan Lin, and Sewoong Oh. Robustness of conditional GANs to noisy labels. In *Advances in Neural Information Processing Systems*, pages 10292–10303, 2018.

[29] Sandhya Tripathi and Nandyala Hemachandra. GANs for learning from very high class conditional noisy labels. Under review. November, 2019.

[30] Sandhya Tripathi and Nandyala Hemachandra. Cost sensitive learning in the presence of symmetric label noise. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining, Macau*. Springer, 2019.

[31] Brendan Van Rooyen, Aditya Menon, and Robert C Williamson. Learning with symmetric label noise: The importance of being unhinged. In *Advances in Neural Information Processing Systems*, pages 10–18, 2015.

[32] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 322–330, 2019.

[33] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In *Advances in Neural Information Processing Systems*, pages 6222–6233, 2019.

[34] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in neural information processing systems*, pages 8778–8788, 2018.

[35] Zijin Zhao, Lingyang Chu, Dacheng Tao, and Jian Pei. Classification with label noise: a Markov chain sampling framework. *Data Mining and Knowledge Discovery*, pages 1–37, 2018.