Research Statement

Hwanjun Song

Data Mining Lab, Graduate School of Knowledge Service Engineering, KAIST

songhwanjun@kaist.ac.kr *https://songhwanjun.github.io*

My general research interests lie in improving the performance of machine learning (ML) techniques under real-world scenarios. I am particularly interested in designing more advanced approaches to handle *(i) large-scale data* and *(ii) noisy data*, which are two main real-world challenges to hinder the practical use of ML approaches.

More details with regard to the research I have done are provided below, followed by an outline of my future research plan.

Previous Research: ML on Large-scale Data

As the amount of data increases rapidly, many ML algorithms have achieved remarkable performance in numerous tasks such as document categorization and image classification. However, the extremely high computational cost for the large-scale data makes them infeasible in real-world. To this end, many researchers approximately decomposed the algorithm into small ones and then performed them in distributed environment such as Hadoop and Spark. This approach greatly improved the efficiency, but still suffered from the following limitations:

Accuracy Degradation (KDD 2017 [1])

Because most of ML algorithms were designed to run on a single machine, it is not trivial to decompose the algorithm for the purpose of parallelization. Thus, many studies divided the entire data into multiple partitions and then simply applied the algorithm in parallel without any guarantee of accuracy. Taking k-Medoids clustering as an example, each k-medoid object must be found from the entire data, but the approximate k-medoid object found from each partition are used for parallel processing. In this regard, my work aims at providing the tight error bound between the optimal solution and the approximate solution based on the number of objects in each partition. Based on the theoretical foundation, the proposed parallel k-Medoids algorithm called PAMAE shows an accuracy comparable to that of optimal k-Medoids and, at the same time, shows an efficiency comparable to that of the most efficient parallel algorithm.

Load Imbalance (SIGMOD 2018 [2])

In the ML algorithms such as DBSCAN, neighboring objects must be assigned to the same data partition for parallel processing to facilitate calculation of the density of the neighbors. That is, the entire data is divided into multiple contiguous sub-regions. However, such region-based partitioning scheme causes the load imbalance problem because the data distribution in the sub-regions tend to be highly diverse in real-world. In MapReduce paradigm, because the execution time is determined by the slowest worker, balancing the load between data partitions is very challenging problem. To remedy this problem, my work aims at building a compact global summary of the entire data, which eliminates the restriction of region-based partitioning. This design enables to randomly divide the entire data into multiple partitions of the same distribution. By applying these technique, the proposed RP-DBSCAN significantly outperforms the state-of-the-art parallel DBSCAN algorithms by up to 180 times without loss of accuracy.

Current Research: ML on Label Noise (Noisy Data #1)

In standard supervised learning, labels of training data are assumed to be true, but they may not be true in real-world because the labeling process is highly cost and time consuming. Such noisy labels lead to poor performance of supervised ML algorithms. In particular, owing to the high capacity to fit any noisy labels, deep neural networks are known to be extremely vulnerable to such label noise. My recent work focuses on training deep neural networks more robustly under the data with label noise.

Robust Training on Label Noise (ICML 2019 [3])

To address the problem of learning from noisy labels, many studies have adopted two strategies: loss correction and sample selection. Loss correction is to correct the loss of all samples in each mini-batch based on the estimated noise transition matrix, and sample selection is to filter out clean samples with low-loss from the training data. However, both strategies still suffered from either accumulating noise from false correction or ignoring useful but hard samples. In this regard, my work designs a hybrid approach of both loss correction and sample selection, which eliminates the drawbacks of the individual methods. The proposed method called SELFIE selectively correct the loss of samples that can be corrected with a high precision. Then, it combines them together with the loss of clean samples to update the network. The experimental results on both synthetic and realistic noisy data show that SELFIE guides the network to avoid noise accumulation from false correction and allows it to take advantage of full exploration of training data.

Future Research: ML on Out-of-context Samples (Noisy Data #2)

When training a classifier, we assume that all samples in training data are in-context samples highly relevant to our purpose, but it is inevitable to contain many out-of-context samples in the data collection procedure (e.g., when crawling animal images using the keyword jaguar, many cars branded as Jaguar are also fetched). Usually, those out-of-context samples are excluded from training data through additional data cleaning processes, which are highly inefficient and time consuming.

My future research aims at handling the out-of-context training samples during training phase without any pre-processing of cleaning data. Different to my work for label noise, the main challenge of this work is how to classify the out-of-context samples during training. Thus, I plan to conduct research on understanding the negative impact of the out-of-context samples on the model and clarifying different attributes between in-context and out-of-context samples in training data.

References

[1] Song, et al., "PAMAE: Parallel k-Medoids Clustering with High Accuracy and Efficiency", In *Proc. 23rd ACM Int'l Conf. on Knowledge Discovery and Data Mining (KDD)*, pp. 1087 ~ 1096, Aug. 2017.

[2] Song, et al., "RP-DBSCAN: A Superfast Parallel DBSCAN Algorithm Based on Random Partitioning", In *37th Proc. Int'l Conf. on Management of Data (SIGMOD*), pp. 1173 ~ 1187, June 2018.

[3] Song, et al., "SELFIE: Refurbishing Unclean Samples for Robust Deep Learning", In *36th Proc. Int'l Conf. on Machine Learning (ICML)*, pp. 5907 ~ 5915, June 2019.