Shallow and Deep Non-IID Learning on Complex Data

Longbing Cao
University of Technology Sydney
Sydney, NSW, Australia

Philip S Yu
University of Illinois Chicago
Chicago, USA

Zhilin Zhao
University of Technology Sydney
Sydney, NSW, Australia

ABSTRACT

Non-IID (i.i.d.) data holds complex *non-IIDness*, e.g., couplings and interactions (non-independent) and heterogeneities (not IID drawn from a given distribution). *Non-IID learning* emerges as a major challenge to shallow and deep learning, including classic statistical learning, mathematical modeling, shallow machine learning, and deep neural learning. Here, we outline the problem, research map, main challenges and topics of shallow and deep non-IID learning.

CCS CONCEPTS

• **Mathematics of computing** → **Probability and statistics**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

Non-IID data, non-IID learning, machine learning, deep learning

ACM Reference Format:
Longbing Cao, Philip S Yu, and Zhilin Zhao. 2022. Shallow and Deep Non-IID Learning on Complex Data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

1 PROBLEM AND RESEARCH MAP

Real-life systems, behaviors and data go beyond the classic i.i.d. (or IID and IIDness, independent and identically distributed) assumption, where data or variables are i.i.d. drawn from a given distribution[1], by following the classic statistical and mathematical terminology. *Non-IIDness* and *non-IID learning* [1, 2] surpass the classic non-i.i.d. settings in the statistical and mathematical sense, where correlation and dependency are typically involved. Non-IID learning is not opposite to the classic i.i.d. analysis following their i.i.d. assumption. *Non-IID* refers to any settings and complexities beyond IIDnesses, where *non-independent* refers to settings such as dependent, correlated, coupled, entangled, and interactive; *non-identically distributed* refers to settings with heterogeneous types, distributions, and relations over variables, sources, samplings, time, space, or heterogeneous results from distinct processes and methods, etc. Together, *non-IIDness* refer to complexities beyond IID, including interaction, coupling relationship, heterogeneity, or non-stationarity over time, space, sampling, source, domain, modality, modelling process, or methods, etc. *Non-IID learning* refers to learning from data with non-IIDnesses. In contrast, *IIDness* refer to data with independence and identical distributions, and *IID learning* refers to learn from data with IIDnesses.

Although shallow and deep Learning have made great success, most methods assume the underlying objects, features, and values are IID and do not involve non-IIDness over sampling, learning processes and methods. In practice, an incorrect understanding and representation of intrinsic non-IIDnesses may result in misleading or incorrect learning and results by IID or near-IID shallow and deep models. Increasing research involves heterogeneity over data sources, modalities, timing, domains and tasks, addressing (1) multi-source, multi-modal, cross-domain, and nonstationary settings, out-of-distribution data, and heterogeneous information networks; and (2) learning methods for nonstationary analysis, domain adaptation, transfer learning, multitask learning, federated learning, and out-of-distribution detection.

Addressing the non-IID nature of complex data, behaviors and systems makes it essential to explore the explicit/implicit interactions and couplings embedded in heterogeneous, dependent, coupled, entangled, interactive or nonstationary data over time, space, sampling, source, domain, modality or by different learning processes, tasks or methods for shallow or deep learning. This results in a comprehensive spectrum of non-IIDnesses in data characteristics, processing, sampling, and learning processes, tasks and methods and from aspects of data types (formats, semantics, etc.) and sources (domains, modalities, networks, views, etc.), etc. Accordingly, Figure 1 illustrates a research map of non-IID learning, which covers a broad-reaching spectrum from data processing to learning. The KDD'22 tutorial[2] presents a comprehensive overview and typical examples of shallow and deep non-IID learning to learn the non-IIDness in complex data. It discussed the limitations of IID learning from complex data, the definitions and frameworks of non-IID shallow and deep learning, and recent learning systems and algorithms.

2 CHALLENGES AND PROSPECTS

The research on non-IID learning covers a wide spectrum of statistical, shallow to deep learning methods and their applications. Figure 1 summarizes some challenges and prospects of non-IID data processing, non-IID feature engineering, non-IID representation, non-IID pattern mining, non-IID statistical learning, non-IID reinforcement learning, non-IID deep learning, non-IID transfer learning, non-IID federated learning, non-IID multi-modal/source/task learning, non-IID vision learning, non-IID natural language processing/document/text analysis, and non-IID behavior modeling, and non-IID applications including non-IID outlier detection, and non-IID recommendation.

[1] https://en.wikipedia.org/wiki/Independent_and_identically_distributed_random_variables.

[2] https://datasciences.org/coupling-learning.

KDD '22, August 14–18, 2022, Washington, DC, USA.

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/XXXXXX.XXXXXX

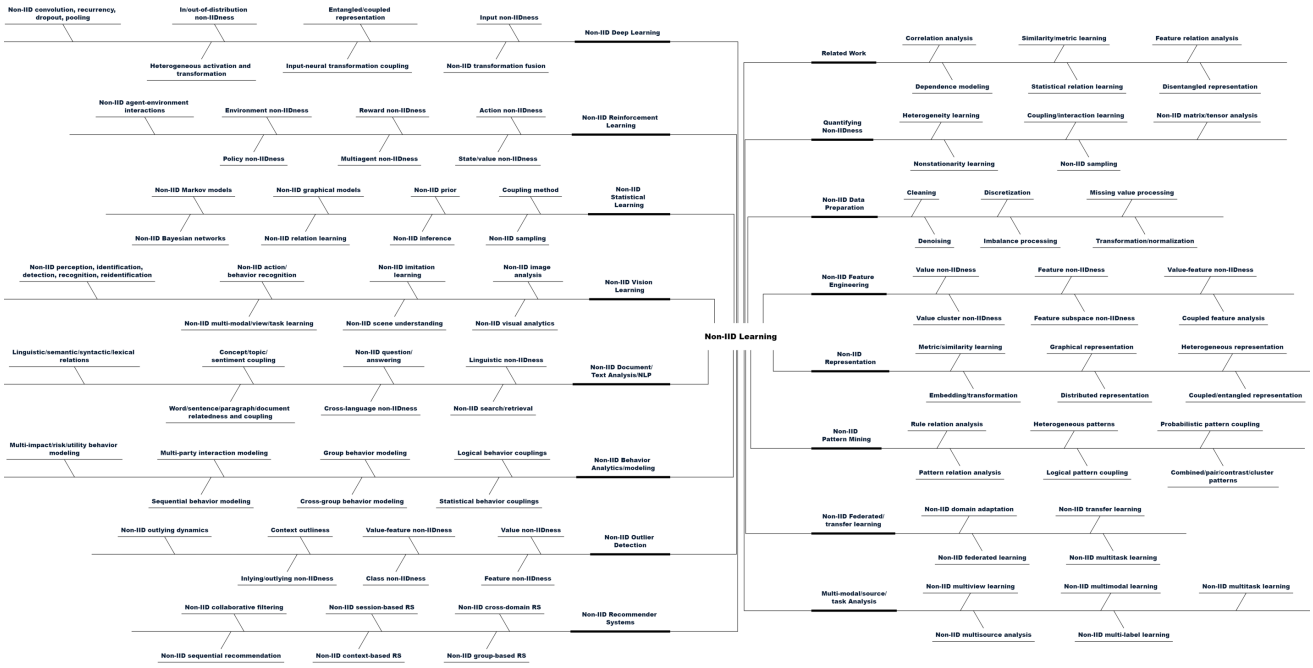

Figure 1: Research map of non-IID learning: In both shallow and deep learning, and from data manipulation and feature engineering to learning process, task, method and result evaluation (please zoom in the diagram for clarity and details).

Non-IID learning reveals many open challenges and prospects to almost every aspect of data understanding to learning. It covers non-IID design, non-IID algorithms, and non-IID system development in different learning paradigms, mathematical and statistical methods, and specific learning theories and methods. Examples are:

- *Quantifying non-IIDnesses* of complex systems, networks, behaviors and their data, such as local/global, static/dynamic, explicit/implicit, flat/hierarchical, vertical/horizontal couplings, interactions, connections, heterogeneities, and non-stationarities between entities, properties, contexts, sources, networks, domains, and modalities;
- *Non-IID learning architectures and frameworks* for non-IID representation, unsupervised, semi-supervised, supervised, and reinforced learning tasks, etc.;
- *Non-IID feature engineering* with value-feature-object-subspace-source couplings, heterogeneities;
- *Non-IID representation learning* with distributed, coupled, entangled, heterogeneous embeddings and transformations;
- *Coupling learning* and *interaction learning* [2] with static/dynamic, explicit/implicit, local/global, structural/semantic etc. couplings and interactions;
- *Heterogeneity learning* [3] with hierarchical and heterogeneous data sources (domains, modalities, networks), distributions, structures, couplings, interactions, samplings, learning processes, learning tasks, etc.;
- *Non-IID statistical learning* with hierarchical, heterogeneous, sparse, and nonstationary factors, relations, distributions, couplings and non-IID prior, sampling, inference, Markov or Bayesian networks;
- *Non-IID network modeling* and *graph learning* with attribute-to-node-to-path couplings and heterogeneities in static/dynamic, undirected/directed, hierarchical/flat, low/high order graph-based and networking settings;
- *Non-IID deep neural learning* with input non-IIDnesses, entangled features, relations and representations, heterogeneous representations and transformations, non-IID convolution, dropout, pooling, recurrency and integration, and in/out-of-distribution settings;
- *Non-IID transfer learning* with heterogeneous and coupled, interactive source/target domains or multiple heterogeneous but coupled tasks;
- *Non-IID federated learning* with non-IID local/global data, source, learning tasks, objectives, and models etc.;
- *Non-IID behavior modeling* with coupled group behaviors, multiparty interactions, heterogeneous multiple time series, sequential behavior matrix, and heterogeneous behavior impact, risk and utility;
- *Non-IID recommendation* with coupled, heterogeneous and evolving users, items, user-item interactions, and rating;
- *Non-IID outlier detection* with asymmetry, heterogeneity, coupling, interactions between majority and minority classes, and between outlying/normal features or samples, etc.

REFERENCES

[1] Longbing Cao. 2014. Non-IIDness Learning in Behavioral and Social Data. *Comput. J.* 57, 9 (2014), 1358–1370.
[2] Longbing Cao. 2015. Coupling learning of complex interactions. *Inf. Process. Manage.* 51, 2 (2015), 167–186.
[3] Chengzhang Zhu, Longbing Cao, and Jianping Yin. 2022. Unsupervised Heterogeneous Coupling Learning for Categorical Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 1 (2022), 533–549.