

氏 名 佐川 正悟

学位(専攻分野) 博士(統計科学)

学位記番号 総研大甲第 2525 号

学位授与の日付 2024 年 9 月 27 日

学位授与の要件 複合科学研究科 統計科学専攻
学位規則第6条第1項該当

学位論文題目 Gradual Domain Adaptation with Multifidelity Learning and
Generative Model

論文審査委員 主 査 吉田 亮
統計科学コース 教授
日野 英逸
統計科学コース 教授
田中 未来
統計科学コース 准教授
松井 孝太
名古屋大学 大学院医学系研究科 講師

博士論文の要旨

氏 名：佐川 正悟

論文題目：Gradual Domain Adaptation with Multifidelity Learning and Generative Model

Applying machine learning to real-world tasks is often constrained due to the limited number of labeled data. Domain adaptation is a practical approach to mitigate the issue. However, conventional domain adaptation methods do not work well when a large gap exists between the source and target domains. Gradual domain adaptation is one of the approaches used to address the problem. This method utilizes the intermediate domains that are arranged to connect the source domain to the target domain. In previous studies, it is assumed that the number of intermediate domains is large and the gap between adjacent domains is small; hence, the gradual domain adaptation algorithm, involving self-training with unlabeled datasets, is applicable. In practice, however, gradual self-training will fail because the number of intermediate domains is limited and the discrepancies between adjacent domains is large.

For example, automatic driving systems require high object recognition accuracy not only during the daytime but also in the evening and nighttime. If the labeled data are obtained only from the daytime, the object recognition accuracy for the evening and nighttime will deteriorate. Gradual self-training will be effective when we have images that are captured continuously from day to night. However, when the interval of capturing images is long, the gaps between adjacent domains is also large, and gradual self-training may not be effective.

When the applicability of gradual self-training is constrained due to the limited number of available intermediate domains, one option is to discard these intermediate domains and apply conventional domain adaptation only to the source and target domains. We argue that given intermediate domains are valuable and should be used, even when the gaps between adjacent domains are large. In this dissertation, to address this issue, we propose two gradual domain adaptation methods that do not use self-training.

First, we consider a problem setting of semi-supervised gradual domain adaptation, in which we can query the label of a sample from the intermediate domains and the target domain. As a practical problem setting, we consider that querying the label of a sample in each domain comes with some cost, which is not uniform. While a simple query strategy is to devote the entire budget to queries from the target domain, we propose to utilize the notion of multifidelity learning and query from both the intermediate and target domains. Our experimental results demonstrate that the

proposed query strategy is more suited to our problem setting, where the query cost of each domain is not uniform, than this simple query strategy.

Next, we propose a method that addresses the issue of large discrepancies between adjacent domains while maintaining the framework of unsupervised domain adaptation. We utilize the normalizing flow, which is one of the generative models, to capture the continuous change between domains. In general, normalizing flows learn the transformation between the distribution that the given data follows and a parametric distribution, such as the Gaussian distribution. To realize gradual domain adaptation without gradual self-training, we develop a non-parametric transformation method between distributions using normalizing flows. We evaluate the effectiveness of our proposed method on real-world datasets and confirm that it mitigates the problem, in which there are large discrepancies between adjacent domains, and improves the classification performance.

Results of the Doctoral Thesis Defense

博士論文審査結果

Name in Full

氏名 佐川 正悟

Title

論文題目 Gradual Domain Adaptation with Multifidelity Learning and Generative Model

[論文の概要]

論文は5章106ページからなり、英語で書かれている。機械学習を現実のタスクに適用する際、ラベル付きデータの不足が問題となる。ドメイン適応はこの問題を緩和する実用的なアプローチだが、従来の方法ではソースドメインとターゲットドメインの間に大きなギャップがあると効果が低い。本論文は、ソースとターゲットドメインの間に中間的なドメインが存在する場合の、段階的ドメイン適応法を提案するものである。

第1章では、機械学習の応用においてラベル付きデータの数が限られていることが多く、ドメイン適応がこの問題を軽減するための実用的なアプローチであることを説明している。従来のドメイン適応手法はソースドメインとターゲットドメインの間に大きなギャップがある場合にはうまく機能しないことを指摘し、段階的ドメイン適応を導入している。従来の研究では、中間ドメインの数が多く、隣接するドメイン間のギャップが小さいことを前提としており、隣接するドメインのデータで学習した予測器による予測結果を擬似的に正解ラベルとみなす「自己学習」を繰り返し適用する。しかし、実際には中間ドメインの数が限られており隣接するドメイン間のギャップが大きいため、段階的自己学習が効果的に機能しないことがあると指摘している。最後に、本論文で扱う問題の重要性と動機を述べた上で、本論文の貢献と研究全体の概要を示している。

第2章では、Kumarら(2020)が提案した段階的ドメイン適応アルゴリズムを詳細に説明し、段階的ドメイン適応に関連する先行研究として重要度重み付けや自己学習などのドメイン適応手法について相違点を明らかにしつつ説明し、それらの利点と限界を議論している。

第3章では、隣接ドメイン間に大きなギャップがある問題に対処するための半教師付き段階的ドメイン適応の問題を扱っている。この問題設定は中間ドメインとターゲットドメインにおいてサンプルのラベルを問い合わせることができるかと仮定し、各ドメインのラベルを問い合わせる際のコストは一律ではない状況を考えている。マルチフィデリティ学習の概念を用いて、各ドメインから効率的にラベルを収集する方法を提案し、その効果を実験により示している。

第4章では、隣接するドメイン間の大きなギャップに対処するため、生成モデルである正規化フローを利用してドメイン間の連続的な変化を捉える方法を提案している。単一の正規分布とデータ生成分布を結ぶ変換を学習するのではなく、複数クラスの判別問題を自然に扱えるように混合正規分布から多クラスデータ分布への変換を段階的に学習する方法を開発した。数値実験結果により提案手法がドメイン間ギャップを軽減し、分類性能を向上させることを確認している。また、中間ドメインがどのような配置にあるときに提案手法が有効であるかを評価している。

最終章である第5章では、本研究の結論を述べて、今後の展望を述べている。

[論文の評価]

機械学習を現実のタスクに応用する際にラベル付きデータの不足が制約となる問題に対して、ドメイン適応は有望かつ重要なアプローチであるが、特にソースドメインとターゲットドメインの間に大きなギャップが存在する場合、通常ドメイン適応手法が効果的に機能しないという問題があった。この問題を解決するために段階的ドメイン適応が提案されているが、従来の手法では中間ドメインが多く存在し隣接するドメイン間のギャップが小さいという理想的な状況が仮定されていた。実際には中間ドメインの数が限られているため、自己学習を用いた段階的ドメイン適応が失敗することがある。本研究ではこの問題に対処するために自己学習を用いない段階的ドメイン適応手法を2つ提案している。第一の手法は中間ドメインおよびターゲットドメインからサンプルのラベルを取得できる半教師ありの設定に基づいており、クエリコストが均一でない現実的な状況を考慮したマルチフィデリティ学習の概念を用いている。第二の手法は教師なしドメイン適応の枠組みを維持しながら、隣接ドメイン間の大きなギャップに対処するために生成モデルである正規化フローを用いたものである。いずれの手法も、実問題への適用において問題となるドメイン間のギャップを異なるアプローチで埋めるものであり、計測や実験コストの削減に貢献する実用的な段階的ドメイン適応手法であるといえる。以上の理由により、本論文は統計科学の博士論文として十分な意義を持つと判断される。なお、第3章の内容は査読付き国際論文誌 *Neural Networks* (vol. 164, pp.731-741, 2023, 第一著者) に掲載されている。