氏　　名　南 俊匠

学位(専攻分野)　博士(統計科学)

学 位 記 番 号　総研大甲第 2404 号

学位授与の日付　2023 年 3 月 24 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　Transfer learning with model transformation

論 文 審 査 委 員　主　査　日野 英逸
統計科学専攻　教授
福水 健次
統計科学専攻　教授
藤澤 洋徳
統計科学専攻　教授
吉田 亮
統計科学専攻　教授
松井 孝太
名古屋大学　大学院医学系研究科　講師

（様式３）

博士論文の要旨

氏　　名　南　俊匠

論文題目　Transfer learning with model transformation

Machine learning recognizes patterns from data, and its applications have achieved great success in a wide range of fields. These successes are supported by a huge amount of data. For example, to train models for natural language processing, datasets consisting of more than one billion words are used. For image classification tasks, it is standard to use more than ten million images. In order to obtain a well-performing machine learning model, collecting a huge amount of training data is crucial.

In many real-world problems, however, it is hard to collect a sufficient amount of data. For example, in scientific fields such as materials science, one can obtain extremely limited amounts of data due to several obstacles, e.g., the cost of data collection, the diversity of researchers' needs, and the high level of information confidentiality. For further progress in machine learning, it is necessary to overcome such a bottleneck of a limited supply of data.

Transfer learning, which refers to the problem of applying the knowledge learned in one or more tasks to develop an effective model for a new task efficiently, has received a lot of attention as a methodology for solving the small data problem. In order to obtain a well-performing model in a new target domain, we reuse knowledge such as parameters of models, encoded features, hyperparameters, and samples from related source domains. So far, various methodologies and theories of transfer learning have been studied. Through practical applications, transfer learning has shown to be an effective technique for tasks with small amounts of data, not only in computer vision and natural language processing but also in various scientific fields, such as materials science and biology.

For the success of transfer learning, which knowledge to be transferred is a fundamental issue. Typically, the parameters or intermediate representations of pre-trained source models are reused. The parameter transfer uses the parameters of the pre-trained model as initial values in training in a target domain. In Bayesian statistics, this approach corresponds to using the source parameters as a prior. On the other hand, transfer learning based on the intermediate representations is called feature extraction. When the source model is a neural network, the features encoded in its intermediate layer are used as input for the target prediction task. As a special case of feature extraction, using the output of the source model to predict a shift to the target is called hypothesis transfer learning.

In this thesis, we focus on supervised transfer learning in regression problems, in

particular, the problem of reusing features obtained from the training in the source domain. There are several benefits of focusing on the reuse of source features. One of the most notable is applicability. In general, source models are not always obtained as statistical models, such as neural networks, but may be physical models or may not be explicitly represented as a function of the input. In such cases, approaches relying on a particular statistical model, such as parameter transfer, cannot be used, but approaches based on source features can. Another benefit is computational feasibility. In the case of parameter transfer, all parameters of the source model must be stored, and the entire model must be retrained. When transferring from extremely large models, such as state-of-the-art neural network models with more than a billion parameters, the computational cost becomes enormous. On the other hand, in the case of feature extraction, the computational cost is relatively low because the feature representation is computed only once in advance. For the widespread application of transfer learning, it is essential to develop methods that rely only on the source features.

After reviewing the several methodologies and theories of transfer learning in Chapter 2, this thesis proceeds as follows.

In Chapter 3, we aim to establish a new transfer learning class that is applicable to any regression model. The proposed class unifies different classes of existing transfer learning methods for regression. To model the transition from a pre-trained model to a new model, we introduce a density-ratio reweighting function. The density-ratio function is estimated by conducting a Bayesian inference with a specific prior distribution while keeping the given source model unchanged. Two hyperparameters and the choice of the density-ratio model characterize the proposed class. It can integrate and extend three popular transfer learning methods within a unified framework, including transfer learning based on cross-domain similarity regularization, probabilistic transfer learning using density-ratio estimation, and fine-tuning of pre-trained neural networks.

In general, the model transfer operates through a regularization scheme to leverage the transferred knowledge between different tasks. Conventional regularization aims to retain similarity between the pretrained and transferred models. This natural idea is what we refer to as cross-domain similarity regularization. On the other hand, the density-ratio method operates with an opposite learning objective that we call the cross-domain dissimilarity regularization; the discrepancy between two tasks is modeled and inferred, and the transferred model is a weighted sum of the pre-trained source model and the newly trained model on the discrepancy. These totally different methods can be unified within the proposed framework.

In Chapter 4, we develop a transfer learning methodology to estimate cross-domain shifts and domain-specific factors simultaneously and separately using given target samples. The framework we employ considers two different transformation functions:

one to represent and estimate domain-specific factors and the other to adapt them to the target domain in combination with the source features. For these transformation functions, we derive a theoretically optimal class based on the assumptions of invertibility and differentiability as well as consistency, i.e., that the optimal predictor does not change through the two transformations. The resulting function class takes the form of an affine coupling of three functions. These functions can be estimated simultaneously using conventional supervised learning algorithms such as kernel methods or deep neural networks. We refer to this framework as the affine model transfer.

The affine model transfer can also be interpreted as generalizing the feature extractor by adding a product term. This additional term allows for the inclusion of unknown factors in the transferred model that are unexplainable by source features alone. Furthermore, this encourages the avoidance of a negative transfer. The usual transfer learning based on feature extraction attempts to explain and predict the data generation process in the target domain using only features from the source domain. However, in the presence of domain-specific factors, a negative transfer can occur owing to a lack of descriptive power. The additional term compensates for this shortcoming.

Several synthetic and real data analyses were performed for each proposed method to highlight practical advantages and features. Through a wide range of prediction tasks, we investigated the applicability of the proposed methods and their behavior according to the cross-domain relationships. Furthermore, we applied the affine model transfer to calibrate the simulated values, and found that, in the estimated model, the domain-common and domain-specific factors were captured consistently with the physicochemical formula.

Form8 : Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis defense

博士論文審査結果

Name in Full
氏　　名　南　俊匠

Title
論文題目　Transfer learning with model transformation

2023 年 1 月 25 日午前 10 時から約 2 時間にわたり南俊匠氏の博士論文審査委員会を開催した．出願者による 1 時間の公開発表による概要説明と質疑応答，さらに約 1 時間の審査委員のみによる審査を行った結果，審査委員会は本論文が学位の授与に値すると判断した．

[論文の概要]

英語で執筆された出願論文は，5 章 126 頁からなる．本研究では，複数の既存手法を包含する転移学習の統一的フレームワークを構築し，様々な手法の特徴付けや理論的性質の考察，転移学習の新しい方法論の提案を行っている．

第 1 章は，転移学習の近年の動向を概説したのち，研究の問題意識と学術的意義・貢献を説明している．

第 2 章は，ベイズ推論，密度比推定，ニューラルネットワークに基づく教師あり転移学習や Hypothesis Transfer Learning (以下 HTL という)，転移学習の汎化誤差解析に関する先行研究のレビューを行っている．

第 3 章では，ドメイン間の類似性を活用するベイズ型転移学習と，ドメイン間の違いを陽にモデリング・推定する密度比推定型転移学習という二つの対照的な手法を二種類のハイパーパラメータでつなぐ転移学習のクラスを定め，ハイパーパラメータ選択によりこれらのハイブリッド型転移学習を実現することを提案している．提案手法の理論的性質に関する考察や数値実験の結果がまとめられている．

第 4 章では，第 3 章の研究から着想を得る形で，アフィン転移学習というもう一つのクラスを提案している．データ変換・モデル変換という二種類の変換関数を用いた転移学習において，ある条件の下でアフィン転移学習は ℓ_2 期待損失を最小にすることが証明されている．このクラスは，深層学習から得られた特徴器による転移学習，密度比推定型転移学習，オフセット型・スケール型の HTL などを包含する．ここでは特に，アフィン型転移モデルを正定値カーネルでモデリングし，汎化誤差や超過リスクの上界を導いた上で理論的性質を論じている．数値実験では，複数のケーススタディを通じて，提案手法が負の転移の抑制に有効であることや，全原子古典分子動力学法による高分子物性計算のキャリブレーションへの適用例などを報告している．

第 5 章では，提案された二つのクラスと既存手法の包含関係をまとめたのち，今後の課題として，転移学習の四つの未解決問題を論じている．

[論文の評価]

転移学習の研究は，長年様々な手法が乱立した状態が続いており，統一的視点に基づいた体系化があまり進展していないことが問題視されてきた．そのような中，限定的な設定の下ではあるが，様々な手法を包含する統一的フレームワークを構築し，それらの特徴付けを行い，理論的性質を明らかにした点は，統計的機械学習の発展に寄与しうる研究成果といえる．また，提案したフレームワークの下で自然な形で新しい方法論を導いており，研究成果の学術的新規性も認められる．以上の理由により，統計科学分野の博士論文の研究としては十分に高い水準に達していると判断する．

[その他]

第 3 章の研究をまとめた論文が査読付き国際会議 AAAI 2021（Proceedings of the AAAI Conference on Artificial Intelligence 2021）に採択されている（第一著者）．また，第 4 章の研究をまとめた論文を arXiv に公開している（第一著者，査読付き国際会議に投稿済）．