様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２９ 年　 ８ 月 １０ 日現在

機関番号：６２６０３
研究種目：研究活動スタート支援
研究期間：2015～2016
課題番号：１５Ｈ０６８２３

研究課題名（和文）Onsite Transfer Learning （現地の転移学習）

研究課題名（英文）Onsite Transfer Learning

研究代表者
柳　松（Liu, Song）

統計数理研究所・統計的機械学習研究センター・特任助教

研究者番号：８０７６０５７９

交付決定額（研究期間全体）：（直接経費）　　1,300,000円

研究成果の概要（和文）：本研究によって以下に述べる２つの主要な結果が得られ，また関連する重要な課題を解決した．（１）事後確率密度比の推定量を開発し，それが人工データ・実データに対して有効に働くことを確認した．（２）密度比推定の理論的解析に対する一般的方法を確立し，理論的性質の解明を行った．（３）（関連課題）部分的グラフィカルモデルのスパース構造を発見する新しい方法を提案した．

研究成果の概要（英文）：We have two major achievements throughout the period of this research. Additionally, this project has inspired us to solve another important related problem.
(1) We successfully developed a posterior ratio estimator and it was shown excellent performance on either synthetic or real dataset.(2) A generic theoretical analysis was made for density ratio estimation problems. We investigated the theoretical property of the density ratio estimator for general problems. (3) (Additional) A novel method was proposed for discovering the sparse structure of a partial Graphical Model.

研究分野：統計学

キーワード：Density Ratio Estimation　Transfer Learning　Markov Network　Graphical Model

１．研究開始当初の背景

Transfer learning has always been an important topic in Machine Learning community in the last decades. The idea of using the data from a “source task” to help a similar target task is very natural and practical (Raina et al., 2006). Some of the previous works assume two different tasks share one similar parametric model and propose techniques to guarantee such a similarity is enforced. Another approach is directly reusing the source dataset to help the learning of a target task. A set of weights are also computed to “reweight” the samples such that only the most “helpful” samples are preserved and the harmful points are removed (Sugiyama et al., 2007).

However, both frameworks have drawbacks.

(1) The model reuse technique is simple to implement, but restrictive. To ensure the similarity of two models, it is required that two models are trained simultaneously. As the learning and modelling is so sophisticated nowadays that for a smart device with limited computational capacity, it is hard to conduct such a transfer procedure “onsite”.

(2) For the sample reweighting method, it is not required to store and compute two complex models simultaneously, so weights can be computed efficiently on a smart mobile device. However, such a method does not make use of the similarity between models of two tasks, which can be beneficial to the transfer task.

２．研究の目的

We hope to find an algorithm that can improve the performance of embedded Machine Learning algorithms when facing an unfamiliar environment, using only limited computational power.

Such an adjustment to fit the new environment is done by a transfer learning algorithm that can take the advantages of model similarity and can be computed with efficiency on small embedded devices.

３．研究の方法

$$P_2(Class\mid Data) = \frac{P_2(Class\mid Data)}{P_1(Class\mid Data)} \cdot P_1(Class\mid Data)$$

Onsite Stage　　Learning Stage

D_2　P_1　　D_1

Figure 1 The idea of onsite transfer

In this research, we focus on classification tasks in general, i.e. the task is to predict a discrete outcome (i.e. class labels) using feature data. In probabilistic classification, the task is to learn a class posterior $P(\text{Class}|\text{Data})$ to make future predictions.

In this transfer learning setting, we assume that an accurate predictive pattern

P_1, i.e. $P_1(\text{Class} \mid \text{Data})$

has been obtained by using dataset D_1. The natural idea of the onsite adaptation is to learn a ratio between two class posteriors

$P_2(\text{Class}|\text{Data})/P_1(\text{Class}|\text{Data})$,

where P_2 is the predictive pattern of onsite dataset D_2, and multiply it with P_1. Therefore, the central issue in this research is how to estimate such a ratio accurately and efficiently, using only onsite information P_1 and D_2 alone.

To be precise, Machine Learning experts may utilize enough data in D_1 to create a predictive pattern, a good classifier P_1 in laboratory. However, during the transfer phase, our algorithm tries to adapt and learn a new probabilistic classifier P_2 using onsite information that is no more than D_2 and P_1. This is a challenging problem. During the research period, we slight change the setting which allows **a small proportion** of samples from D_1 to be included.

Through a simple intuitive analysis, we can see that separately estimate the P_1 and P_2 and then take the ratio between them would not be a logical choice since

① If one can already obtain an accurate estimate of P_2, there is no need to transfer;

② Given sample size of D_2 is usually much smaller than that of D_1, P_2 obtained by such a procedure is likely too poor to be used.

Therefore, in this research, we plan to use the density ratio estimation criterion to **directly** estimate the ratio between probabilities.

Since the nature of this research is both theoretical and practical, we divide the entire research scheme into two parts:

(1) Developing the estimator of the posterior ratio, and validate it using empirical experiments.

(2) Conducting theoretical analysis showing the proposed ratio estimator have good performance when two tasks are similar.

4．研究成果

We have two major achievements throughout the period of this research. Additionally, this project has inspired us to solve another important related problem. All three projects have now been published.

(1) We successfully develop a posterior ratio estimator for P_2/P_1 and it has shown excellent performance on either synthetic or real dataset.

In this work, we tackle the problem of estimating the ratio between two conditional probability, i.e.,

$$P_2(\text{Class}|\text{Data})/P_1(\text{Class}|\text{Data}).$$

An estimator is proposed that makes use of the sample from the target task and a small proportion of the source task data that is the "most similar" to the target dataset.

The ratio function is modelled using a log-linear model, and the learning criteria is a simple maximum likelihood estimation (MLE). However, we normalize our model using samples from the source task dataset so that the ratio function will be suppressed at the locations where P_1 is high. This is a slight deviation from our original plan as we hoped to estimate P_2/P_1 without using any direct information from D_1 which is usually in huge volume and hard to fit into the memory of a mobile device.

However, in the proposed estimator, it is not required to carry over the entire source dataset, but only the samples that are directly related to the target problem. In fact, using the k-nearest neighbor selection criterion, we only need to prepare kn_2 samples from D_1 where n_2 is the number of samples in D_2 and k is also a small scalar. The computational overhead is still manageable.

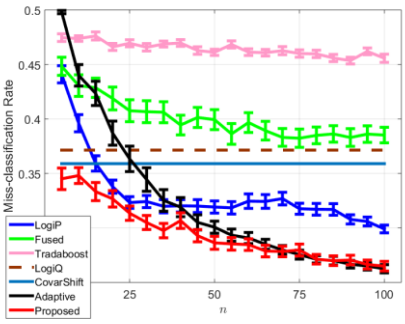

Figure 2 The classification error of various transfer learning methods. The proposed method (Red) achieve the lowest classification error rate.

Experiments were performed on various synthetic and real datasets. Here we only shown an example. In this experiment, we compare the proposed transfer learning method with a few other existing transfer learning methods using Amazon Sentiment dataset, where the task is using the comments of some categories with many customer feedbacks to help the classification of users' sentiment of a category with only a small number of feedbacks.

In Figure 2, we plot the classification error rate of the "Kitchen" category using the customers' feedback from "DVD", "Books" and "Electronics". It can be seen from the plot that the proposed method achieves the lowest classification error rate.

(2) The proposed research also inspired us to investigate the theoretical property of the density ratio estimator in general.

In this work, we consider a generalized joint density ratio estimator. We argued that estimating the ratio of two densities is easier than estimating P_1 and P_2 separately. However, can we justify this intuition rigorously?

As we mentioned, the framework of transfer learning assumes that the models of source and target task are "similar" to each other. Thus, it is sensible to also assume, the difference between two parameters, which determines the behaviors of P_2/P_1, is **sparse**.

We investigate the sufficient conditions for a ratio estimation to "successfully" recover the sparse structure of the parameter and the major discovery is that the number of samples needed for

such a successful sparsity recovery depends only on the sparsity pattern of the differential parameter, rather than the sparsity of density parameters of P_1 and P_2 individually. It means even if we have P_1 and P_2 which are highly complex, if P_1 and P_2 are "similar" to each other in the sense that the difference between two model parameters is sparse, we are guaranteed to have a good density ratio estimation performance with only a few samples.

Such a discovery guided an experiment of learning a density ratio over two high dimensional datasets with only small number of samples. The sparsity pattern of the ratio parameter reveals the changes between two complex gene networks (See Figure 3).

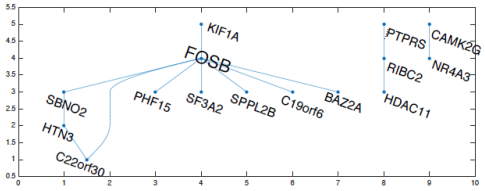

Figure 3 The differential model learned between two gene networks.

Here the gene profiling data are sampled under two different external environments. The "FOSB" gene is a known regulator that controls the expression of other genes and it is switched on and off given different external stimuli. The density ratio estimator can discover such an important regulator gene without any prior knowledge.

(3) A novel method was proposed for discovering the sparse structure of a partial Graphical Model.

We have shown that if P_2 and P_1 are similar, then density ratio estimation can be very efficient. It opens a door of learning the partial pattern of P_2 even if we do not have enough samples to infer the complete pattern of P_2.

We can create a synthetic dataset of $\bar{P}_2$ which has a "similar model" comparing to the original P_2. The simplest case is that if we have $P_2(X,Y)$, where X and Y are multivariate random variables. It is easy to create another distribution $\bar{P}_2 \coloneqq P_2(X)P_2(Y)$ through marginalization. We show that such a distribution preserves all possible interactions within X and Y while removing all interactions between X and Y.

By comparing P_2 with $\bar{P}_2$, we can discover the interactions **between X and Y** while the "synthetic distribution" $\bar{P}_2$ serves as a "mask" to highlight the differences between two distributions.

We perform the partial graphical mode learning on the recorded U.S. Senate voting data and the generated partial graphical model identifies the bipartisanship in the U.S. Senate. See

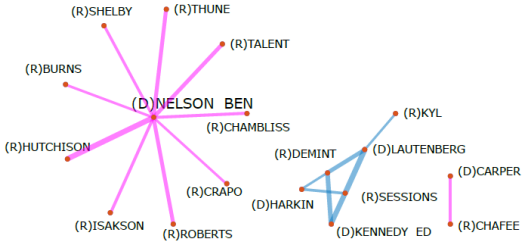

Figure 4 The partial graphical model learned from US senate voting record.

Figure 4.

It can be seen from Figure 4 that one of the conservative democrats "Ben Nelson" is identified as the "hub" which links to multiple republican senators. It is noteworthy that our method also identifies a group of senators that constantly vote against each other (marked with blue links).

References

Raina, R., Ng, A. Y., & Koller, D. (2006). Transfer learning by constructing informative priors. *Inductive Transfer*, 10, 2006.

Sugiyama, M., Krauledat, M., & Müller, K. R. (2007). Covariate shift adaptation by importance weighted cross validation. The Journal of Machine Learning Research, 8, 985-1005.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　2　件）

① Liu, S. Suzuki, T., Fukumizu K., Learning Sparse Structural Changes in High-dimensional Markov Network: A Review on Methodologies and Theories., [with peer review], Behaviormetrika,

44:265, 2017, DOI: 10.1007/s41237-017-0014-z

② Liu, S., Suzuki, T., Relator R., Sese J., Sugiyama, M., Fukumizu, K., Support consistency of direct sparse-change learning in Markov networks., with peer review, Annals of Statistics, Vol. 45, No. 3, 959: 990, 2017, DOI: 10.1214/16-AOS1470

〔学会発表〕（計　2　件）

① **Presenter:** 柳　松
Title: Estimating Posterior Ratio for Classification: Transfer Learning from Probabilistic Perspective.
Conference name: SIAM International Conference on Data Mining (SDM 2016)
Date: 2016/5/7
Place of Conference: Miami, Florida, USA.

② **Presenter:** 柳　松
Title: Structure Learning of Partitioned Markov Networks
Conference name: International Conference on Machine Learning (ICML 2016)
Data: 2016/6/23
Place of Conference: New York City, USA

〔図書〕（計　0　件）

〔産業財産権〕

○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
なし

６．研究組織
(1)研究代表者
柳　松（Song Liu）
統計数理研究所・統計的機械学習研究センター・特任助教
研究者番号：80760579

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
（　　）