A Multiple Source based Transfer Learning Framework for Marketing Campaigns

James Brownlow*, Charles Chu*, Guandong Xu†, Ben Culbert*, Bin Fu* and Qinxue Meng*
* Marketing, CFS, Sydney, Australia
† Advanced Analytics Institute, Faculty of Engineering and Information Technology,
University of Technology Sydney, Sydney, Australia
Email: {James.Brownlow, Charles.Chu, Ben.Culbert, Bin.Fu, Qinxue.Meng}@cba.com.au, Guandong.Xu@uts.edu.au

***Abstract*—The rapid growing number of marketing campaigns demands an efficient learning model to identify prospective customers to target. Transfer learning is widely considered as a major way to improve the learning performance by using the generated knowledge from previous learning tasks. Most recent studies focused on transferring knowledge from source domains to target domains which may result in knowledge missing. To avoid this, we proposed a multiple source based transfer learning framework to do it reversely. The data in target domains is transferred into source domains by normalizing them into the same distributions and then improving the learning task in target domains by its generated knowledge in source domains. The proposed method is general and can deal with supervised and unsupervised inductive and transductive learning simultaneously with a compatibility to work with different machine learning models. The experiments on real-world campaign data demonstrate that the proposed method outperform state-of-the-art methods.**

I. Introduction

Nowadays, as business competition has been increasingly fierce, marketing campaigns especially online ones, are delivered more frequently than ever before [1]. As a result, how to efficiently and accurately identify prospective customers who are most likely to respond for the first time in the near future (like one month) after receiving a campaign has become one of the top priority questions in marketing strategy [2]. After all, it is annoying and expensive to spam emails to everyone.

The early work of identifying prospective customers are based on customer segmentation which is heavily relied on traditional data mining and machine learning algorithms including rule-based methods [3, 4, 5], tree models [6, 7, 8], linear and non-linear regression methods with various kernel functions [9, 10, 11]. However, nearly all of these methods share the same assumption that data distribution maintains the same. In fact, as time goes, the profiles and behavior patterns of customers are very likely to change which could result in the change of data distribution. This means that trained models are obsolete and cannot perform well on new data because it has a different data distribution with the training data.

On the other hand, it is impractical to train a model for each campaign. Firstly, recollecting training data, rebuilding models and refine-tuning parameters are time-consuming and costly. Meanwhile, the life-cycle of marketing campaigns is generally short. Over 80% of marketing campaigns are ran less than three months [12]. This makes training a model for each campaign even harder to achieve. Secondly, labeled training samples are often few and sometimes there is even no labeled data, e.g. a new campaign. Without sufficient labeled data, training an accurate model is not possible.

Recently, transfer learning [13] is proposed to deal with the above two issues. This method aims to extract the knowledge of previously trained models from source domains and use it to facilitate the training procedure of the learning tasks in target domains where there may be limited labeled data. Till now, transfer learning has been widely applied in image recognition [14, 15, 16], natural language process [17, 18, 19] and robotics [20], and achieved a big success. Yet, the applications in marketing campaign analysis are not many. The initial work in this field focused on transferring important samples into a untrained dataset. These selected samples from source domains follow the same distribution with untrained dataset in target domains so as to increase the number of labeled samples. Bickel et al. [21] formalized the problem of identifying prospective customers of advertising into a transfer learning problem. This study proposed a transfer learning model to identify customers' sociodemographic features such as gender, age and marital status based on their surfing history and delivering advertisements to users based on their identified sociodemographic features. Their transfer learning procedure focused on resampling data that follows the similar distribution of a given target data. Other methods based on transferring important samples includes active learning based resampling [22], heuristic methods [23] and boosting methods [24]. However, transferring instances from source domains to target domains is inefficient as it is costly to select many and few samples contribute little to improve the learning performance in target domains. Meanwhile these methods can only be applied on the scenario that the learning tasks of source domains and target domains are the same. Otherwise, the labels of selected samples are not valid in target domains. However, in real world, marketing campaigns are very likely to have different purposes. Then another online advertising study [25] considered to use models of source domains to generate new features of data in target domain. This practice can accelerate the training process but fail to deal with the issue of lacking sufficient labeled data and the effectiveness of this transfer heavily depended on the selection of hypeparameters.

To deal with above issues, we proposed a Multiple Source based Transfer Learning Framework for Marketing Campaigns (MS-TLMC) method which can extract knowledge of both data and models from multiple source domains and use it to improve the learning performance in a given target domain. The proposed method has a concise, scalable framework to explicitly control the transferring process and also gives solutions for common issues in campaign data, including imbalance labels and outliers. The main contributions can be summarized as

- The proposed method can work on extract knowledge of multiple source domains simultaneously;
- Instead of selecting important samples that have a similar distribution of the data in a target domain, the proposed method can normalize data of source and target domains to have the same distribution;
- This method is able to deal with supervised and unsupervised inductive and transductive learning tasks in target domains;
- The proposed method is compatible with different traditional classification methods;
- It can deal with issues of imbalance and outliers which are common in campaign data;
- A novel cross validation framework is proposed to evaluate the performance of transfer learning methods.

To evaluate the performance of the proposed MS-TLMC method, we firstly compare it with the scalable transfer learning framework [25] on a series of campaign data and further test its compatibility with different classification models including Logistic Regression [10], Support Vector Machines [11] and XGBoost [26] followed by an efficiency evaluation.

II. Literature Review

The research of transfer learning can be traced back to 1990s. The earliest work to know is the discriminability-based transfer (DBT) method proposed by Lorien Pratt [27] in 1993. The later "Learning to Learn" workshop about lifelong machine-learning in NIPS Conference [28] argued that retaining and reusing previously learned knowledge in new learning tasks was a key to improve the learning performance as learning tasks became increasingly more complex. This workshop triggered a wide discussion about transfer learning which has been a major research topic appearing in top machine learning conferences and journals [29].

Till now, most research and successful applications of transfer learning are concentrated in computer vision and natural language processing [30, 31, 32] by combined convolutional neural networks, recurrent neural networks and deep neural networks. For example, in computer vision, Oquab et al. [33] showed that image representations learned by convolutional neural networks on a large-scale annotated datasets can be efficiently transferred to other visual recognition tasks with limited amount of training data. Based on this, Shin et al. [16] successfully transfered the learned neural network of ImageNet[1] into a Computer-Aided system to detect thoraco-abdominal lymph node and interstitial lung disease from axial CT slides. Zoph et al. [34] considered to learn small common block functions which can be used in different convolution neural networks to recognize images. For natural language processing, Huang et al. [35] proposed a cross-language knowledge by building a shared-hidden-layer multi-lingual deep neural network. The research by Hill et al. [36] demonstrated that transferring unsupervised [37] or supervised [36] word meaning learned from context are possible by sharing word embeddings learned by neural machine translation models trained by bilingual texts. Comprehensive surveys of transfer learning in computer vision and natural language processing can be found in [13, 29, 38]. The above research of transfer learning mainly focused on transferring model components such as model inputs (feature space), small functional blocks, or model parameters because training and fine-tune a deep neural network model is challenging. However, deep neural networks are rarely applied on regular datasets. One of the main reasons is that features of images and texts have strong local correlations while the features of regular datasets are often independent. This makes the performance of deep neural networks on regular datasets not satisfied. Thus, proposing an efficient and effective transfer learning framework on regular datasets such as campaign data, has been increasingly urgent because of a massive input data and a growing demand on marketing campaign delivery.

Currently, the research of transfer learning on regular datasets mainly focused on short life-cycle learning tasks such as real-time learning and online learning [39] which allow a very limited time to train a model. The initial research focused on transferring instances or instance feature space to target domains to increase the number of training samples. Wang and Pineau [24] selected samples that have the similar distributions with samples in target domain. Zhao et al. [22] proposed to adjust the weights of selected samples from source domains by boosting. Later, the research focus of transfer learning moves to knowledge transfer (learned models). Perlich et al. [25] proposed an optimization method based on minimizing loss functions and adding regularizations. Long et al. [40] added kernel functions in objective functions. These methods can accelerate the learning process and have objective functions converged quickly in target domains but the learning performance is still heavily relied on a large labeled training dataset.

Indeed, proposing a general framework of transfer learning is challenging as it needs to deal with multiple types of learning questions. According to the difference of source and target domains, transfer learning tasks can be categorized into inductive and transductive learning and according to the availability of class labels in target domains, these learning tasks can be supervised and unsupervised. Most previous research focused on solving one or several questions rather than build a general, scalable framework to address all of them

[1] http://www.image-net.org/

systematically. Real-world applications often need to deal with all these questions simultaneously. Thus, this paper proposes a general, multiple-source based transfer learning framework that can deal with all above mentioned problems. It focuses on making a full use of generated knowledge of both data and models in multiple source domains to improve the learning performance in target domains and it is also compatible with different learning methods.

III. Preliminaries

This section defines notions and terms used in this paper followed by a formal definition of the transfer learning problem in the context of marketing campaign analysis.

A. *Multiple Source based Transfer Learning*

Transfer learning is proposed to improve the learning performance in a target domain based on the learned knowledge from multiple source domains. Specifically, a domain ($\mathcal{D}$) here generally has a feature space $\mathcal{X}$ where instances ($x_i \in X$) follow a distribution P. In this domain, a learning task $\mathcal{T}$ is considered to include a class label Y and a mapping function f parameterized by θ which is learned to represent the relationship between instances X and their corresponding class labels Y. Transfer learning aims to improve the learning efficiency and effectiveness of the mapping function f_T in $\mathcal{D}_T$ based on the knowledge (f_S) learned from n source domains ($\mathcal{D}_{S_1}, \ldots, \mathcal{D}_{S_n}$) where $\mathcal{D}_{S_i} \neq \mathcal{D}_T$ andor $\mathcal{T}_{S_i} \neq \mathcal{T}_T$ [29]. For the transfer of knowledge to be effective, the learning function f_{S_i} and f_T should be correlated. According to the difference of source and target domains and their learning tasks, transfer learning can be roughly categorized into inductive transfer learning [18] and transductive transfer learning [40].

B. *Multiple Source based Inductive Transfer Learning*

Consider n source domains, marked as $\mathcal{D}_{S_1}, \ldots, \mathcal{D}_{S_n}$ with their corresponding learning tasks $\mathcal{T}_{S_1}, \ldots, \mathcal{T}_{S_n}$, for a given target domain $\mathcal{D}_T$ and its learning task $\mathcal{T}_T$, multiple source based inductive transfer learning is proposed to improve the learning performance of f_T in $\mathcal{D}_T$ based on the learned knowledge from $f_{S_1}, \ldots, f_{S_n}$ where not all source learning tasks are the same with the target learning task ($\exists S_i : \mathcal{T}_{S_i} \neq \mathcal{T}_T$).

C. *Multiple Source based Transductive Transfer Learning*

Consider n source domains, marked as $\mathcal{D}_{S_1}, \ldots, \mathcal{D}_{S_n}$ with their corresponding learning tasks $\mathcal{T}_{S_1}, \ldots, \mathcal{T}_{S_n}$, for a given target domain $\mathcal{D}_T$ and its learning task $\mathcal{T}_T$, multiple source based inductive transfer learning is proposed to improve the learning performance of f_T in $\mathcal{D}_T$ based on the learned knowledge from $f_{S_1}, \ldots, f_{S_n}$ where not all source domains are the same with the target domain ($\exists S_i : \mathcal{D}_{S_i} \neq \mathcal{D}_T$) but their learning tasks are the same ($\forall S_i : \mathcal{T}_{S_i} = \mathcal{T}_T$).

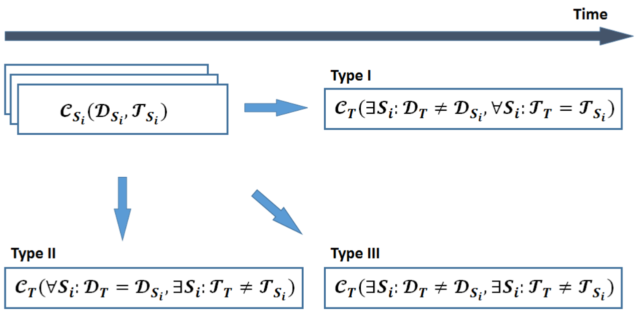
Time
$\mathcal{C}_{S_i}(\mathcal{D}_{S_i}, \mathcal{T}_{S_i})$
Type I
$\mathcal{C}_T(\exists S_i: \mathcal{D}_T \neq \mathcal{D}_{S_i}, \forall S_i: \mathcal{T}_T = \mathcal{T}_{S_i})$
Type II
$\mathcal{C}_T(\forall S_i: \mathcal{D}_T = \mathcal{D}_{S_i}, \exists S_i: \mathcal{T}_T \neq \mathcal{T}_{S_i})$
Type III
$\mathcal{C}_T(\exists S_i: \mathcal{D}_T \neq \mathcal{D}_{S_i}, \exists S_i: \mathcal{T}_T \neq \mathcal{T}_{S_i})$

Fig. 1. Three Types of Transfer Learning Questions in Marketing Campaign Analysis.

D. *Multiple Source based Transfer Learning for Marketing Campaigns*

Marketing campaign analysis is naturally suitable for multiple source based transfer learning. Specifically, a marketing campaign ($\mathcal{C}$) can be considered as one domain ($\mathcal{D}$) with an attached learning task ($\mathcal{T}$). Most marketing campaigns share the correlated objectives such as customer acquisition, customer retention and product promotion which is the prerequisite for effective knowledge transfer from learned domains to unknown ones. However, applying transfer learning is challenging. As illustrated in Fig. 1, there is a marketing campaign pool containing a group of previous learned campaigns including either online or offline ones. For **Type I** transfer, as time goes, same campaigns are conducted repeatedly. But the previous domains ($\mathcal{D}_T \neq \mathcal{D}_{S_i}$ and $\mathcal{T}_T = \mathcal{T}_{S_i}$) may change, e.g. data distribution due to new imported data including new transactions, changes of customer profiles and products. Even in the same time slot (**Type II** transfer), models are required to learn for new campaigns ($\mathcal{D}_T = \mathcal{D}_{S_i}$ and $\mathcal{T}_T \neq \mathcal{T}_{S_i}$). **Type III** transfer refers to transferring old domains and learned campaigns to new domains and unknown campaigns. In fact, **Type II** transferring tasks are few, as the domains are very likely to change because even though they are from the same data sources, different data sampling techniques and filters are very likely to break original data distribution. It can be seen that Type I transfer is transductive learning while the other two are inductive learning. Meanwhile according to the availability of class label $\mathcal{Y}_T$ in $\mathcal{D}_T$, the transfer learning process can be supervised or unsupervised. In this paper, we aim to propose a general multiple source based transfer learning framework to solve these problems.

IV. A Multiple Source based Transfer Learning Framework for Marketing Campaigns

The proposed multiple source based transfer learning framework for marketing campaigns (MS-TLMC) method is a general transfer learning framework for modeling marketing campaigns which mainly consists of three stages including domain transfer, task transfer and a final optimization stage.

A. *Campaign Model*

Before describing the proposed MS-TLMC method, we firstly model marking campaigns in the context of transfer learning. Consider a marketing campaign ($\mathcal{C} = (\mathcal{D}, \mathcal{T})$) where instances X in the feature space $\mathcal{X}$ follow a distribution P. The corresponding class label is marked as Y and the mapping

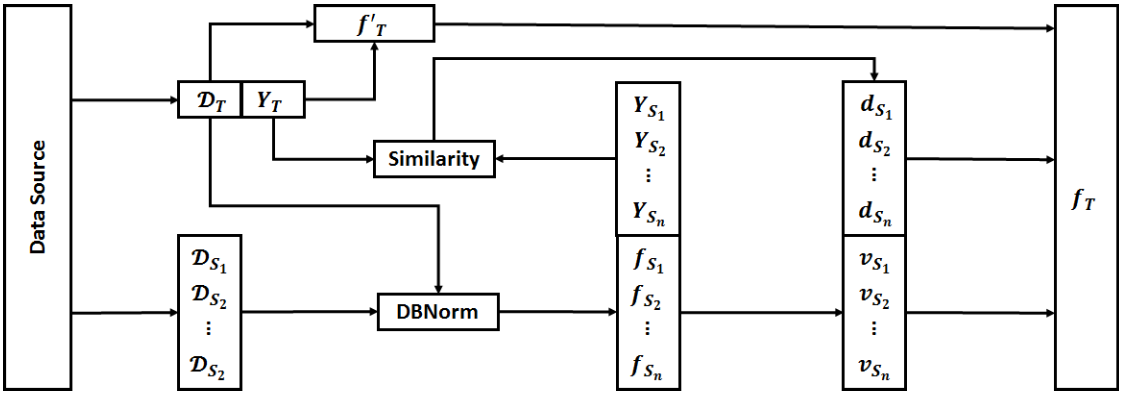

Fig. 2. This figure illustrated the working process of the proposed Multiple Source based Transfer Learning Framework. Unlike the traditional transfer learning method of acquiring knowledge from source domain and applying in a target domain, our method achieved it reversely by mapping target data into source domains.

relation between X and Y can be achieved by a learning function f parameterized by θ. The objective learning function (Ψ) of this campaign is defined as

$$\Psi_{\mathcal{C}} = L\left(f(X;\theta), Y\right) W(Y) + \epsilon(\theta) \tag{1}$$

where L is the loss function between prediction $(f(X;\theta))$ and actual value (Y) and W is the weight function to deal with imbalance of class labels because the campaign data is often very imbalanced with a very low response rate. The term ϵ is the penalty of the model (θ) to avoid over-fitting. In this paper, the loss function (L) of campaign models is Huber loss function [41] which is able to limit the disturbance of outliers and it is defined as

$$L_\delta = \begin{cases} \frac{1}{2}(f(X;\theta) - Y)^2 & |f(X;\theta) - Y| \leq \delta \\ \delta(|f(X;\theta) - Y| - \frac{1}{2}\delta) & \text{otherwise} \end{cases} \tag{2}$$

where δ is a hyper-parameter. The weight function (W) [42] is defined as

$$W(Y_{c_i}) = \frac{|Y_{c_i}|}{|Y|} \tag{3}$$

where $|Y_{c_i}|$ is the number of instances with class label c_i. The proposed multiple source based transfer learning framework uses the knowledge learned from previous campaigns $(f_{S_1}, \ldots, f_{S_n})$ in the learning process (f_T) of target domains.

B. Domain Transfer

The first step of the proposed multiple-source transfer learning is to map target domains into source domains if they are different $(\mathcal{D}_T \neq \mathcal{D}_{S_i})$ which refers to either different feature space $\mathcal{X}$ $(X \in \mathcal{X})$ or different distribution P. The major difference between $\mathcal{D}_T$ and $\mathcal{D}_{S_i}$ lies in data distribution because most marketing campaign systems generally capture customer and product data with the same features. In our previous research, we built a distribution-based normalization (DBNorm) method [43] which are developed as a R package [44] and can be accessed via GitHub[2]. DBNorm enables to normalize P_T into the same distribution with P_{S_i}. Specifically, DBNorm can adjust values from one scale to the other and keep the order of the original values unchanged which means that after normalization, minimum, maximum and median should be the same. Here assume that P_T and P_{S_i} are different, their probability density functions are marked as d_T and d_{S_i} respectively. Given an element $m_1 \in X_T$, the probability of m_1 is $P_T(m_1)$.

$$P_T(m_1) = P_T\left[t <= m_1\right] = \int_{-\infty}^{m_1} d_T(t)dt \tag{4}$$

To maintain the order of data in X_T unchanged after normalizing to X_{S_i}, we need to find an element in X_{S_i} which satisfies the probability of m_1 in X_T being equal to the probability of m'_1 in X_{S_i}. This can be expressed in the following equation where the probability of m_1 under the probability density function d_T is the same with the probability of m'_1 under the probability density function d_{S_i}.

$$P_T(m_1) = P_{S_i}(m'_1) \Rightarrow \int_{-\infty}^{m_1} d_T(t)dt = \int_{-\infty}^{m'_1} d_{S_i}(t)dt \tag{5}$$

It can be found that after domain transfer, transductive transfer learning problems are cast into inductive transfer learning ones if their learning tasks are different or normal learning problems if their learning tasks are the same as they share the same domains $(\mathcal{D}_T = \mathcal{D}_{S_i})$.

C. Task Transfer

The performance of knowledge transfer lies in the similar of learning task. In this paper, the similarity of learning functions are quantitatively measured by a modified cosine similarity [45] of class labels. Here the similarity of the learning task $\mathcal{T}_T$ in target domain $(\mathcal{D}_T)$ and the learning task $\mathcal{T}_{S_i}$ of ith source domain $(\mathcal{D}_{S_i})$ are measured via the similarity of Y_T and Y'_{S_i} which is defined as

$$\begin{aligned} d'_{S_i}(\mathcal{T}_T, \mathcal{T}_{S_i}) &= \max(0, \text{sim}(Y_T, Y'_{S_i})) \\ &= \max(0, \text{sim}(Y_T, f_{S_i}(X_T; \theta_{S_i})) \\ &= \max\left(0, \frac{Y_T \cdot f_{S_i}(X_T; \theta_{S_i})}{||Y_T|| \; ||f_{S_i}(X_T; \theta_{S_i})||}\right) \end{aligned} \tag{6}$$

[2] https://github.com/mengqinxue/DBNorm

And the weights of learning tasks from source domains are normalized as

$$d_{S_i} = \frac{d'_{S_i}}{\sum_{i=1}^{n} d'_{S_i}} \tag{7}$$

where Y'_{S_i} is the learning results generated by learning function f_{S_i} from source domain $\mathcal{D}_{S_i}$ on target domain instances X. The reason of introducing a threshold function $(\max)$ is that the value range of cosine is from -1 to 1 where negative values mean that vectors are negatively correlated while positive values mean they are positive correlated. Then the threshold function let the proposed multiple source based transfer learning model can only consider those learning tasks of source domains that positively correlated to $\mathcal{T}_T$.

D. The Learning Framework of MS-TLMC

After transferring a given target domain $\mathcal{D}_T$ and its learning task $\mathcal{T}_T$ into n source domains $(\mathcal{D}_{S_1}, \ldots, \mathcal{D}_{S_n})$ and their corresponding learning tasks $(\mathcal{T}_{S_1}, \ldots, \mathcal{T}_{S_n})$, it is possible to apply learned knowledge to facilitate the learning process of the target learning task. Then objective function (Δ) of the target learning task is

$$\begin{aligned}\Delta = &\frac{l_T/|X_T|\ f'_T(X_T, d_{S_1}v_{S_1}, \ldots, d_{S_n}v_{S_n}; \theta_T)}{l_T/|X_T| + \sum_i^n l_{S_i}/|X_{S_i}|} \\ &+ \sum_{i=1}^{n} \frac{l_{S_i}/|X_{S_i}|\ d_{S_i}v_{S_i}}{l_T/|X_T| + \sum_i^n l_{S_i}/|X_{S_i}|}\end{aligned} \tag{8}$$

where $v_{S_i} = f_{S_i}(N_{S_i}(X_T); \theta_{S_i})$. The symbol l_T is the number of labeled samples and $|X_T|$ is the number of all samples in the target domain while l_{S_i} is the number of labeled samples and $|X_{S_i}|$ is the number of all samples in the ith source domain. Function f'_T is a model learned from the target domain instances X_T and new features generated by source domains and d_{S_i} is the similarity between the learning task of target domain and the learning task of ith source domain. Function $N_{S_i}(X_T)$ is a normalization process to normalize target instances X_T into ith source domains so as to let them have the same distribution and f_{S_i} is the learned model in ith source domain. It can be found that f_T have two parts, one is from target domain and the other one is from source domains. The contributions of the models learned from the target domain and source domains are determined by the ratio of the number of labeled samples to the number of all samples and the weights of source domains are determined by the similarity of learning tasks in source domains and it of the target domain. This practice is able to make a full use of the data from the target domain and knowledge from source domains. Another advantage is that it can make the proposed MS-TLMC general to deal with supervised and unsupervised learning tasks. Specifically, if there are labeled samples in the target domain, the model result is determined by both data in target domain and learned knowledge from source domains; and if there is no labeled data in the target domain, $l_T = 0$, the model result is determined by source domains which is unsupervised. In this method, f'_T can be learned by the following objective function

$$\begin{aligned}\Psi = &L\left(f_T(X_T, d_{S_1}v_{S_1}, \ldots, d_{S_n}v_{S_n}; \theta_T), Y_T\right) W(Y_T) \\ &+ \frac{1}{2}\lambda||\theta_T||^2\end{aligned} \tag{9}$$

where L is the loss function, X_T represents instances with a class label Y_T, $d_{S_i}v_{S_i}$ is the new feature generated by the ith source domain, the learning model is f_T parameterized by θ_T and W is the function to deal with the imbalance issue of the class label. The term $||\theta^2||$ is L2-regularization of f_T and λ is a hyper-parameter to adjust the weight of regularization.

E. Evaluation

Inspired by Leave-One-Out-Cross-Validation [39], this paper also proposed a similar cross-validation method, named Leave-One-Domain-Out-Cross-Validation (LODOCV), to evaluate the performance of transfer learning method. Specifically, assume that there are m domains, this evaluation method iteratively selects one of the domains as the target domain and considers the other $m - 1$ domains as source domains to evaluate the performance of the proposed MS-TLMC on the chosen target domain based on the other source domains until all domains are selected. Then the performance of the transfer learning method is measured by the averaged result of all iterations.

V. Datasets

The experimental dataset is provided by Commonwealth Bank of Australia[3] who has over 11 million customers. In this paper, we randomly selected ten marketing campaigns from 2015 to 2017 and 57 features of customers are collected covering their demographics, financial status, income & employment, products & super contributions and engagement. For each marketing campaign, the snapshot data of customers are captured before it delivered and different marketing campaigns share the same feature space.

VI. Experiment

The experiment firstly compares the performance of the proposed Multiple Source based Transfer Learning for Marketing Campaigns (MS-TLCM) method against one state-of-the-art transfer learning method (the Scalable Transfer Learning, STL). Then we evaluate its compatibility with Logistic Regression (LR), Support Vector Machines (SVM) and XGBoost. Meanwhile, in order to demonstrate the capability of the proposed method in dealing with small training dataset, the above experiments are ran with different proportions of training data. The last part investigates the model efficiency with different number of source domains and training samples.

A. Comparison Experiment

This section compares the performance of our proposed MS-TLCM method to STL on the collected campaign data. Here we do Leave-One-Domain-Out-Cross-Validation (LODOCV)

[3] https://www.commbank.com.au/

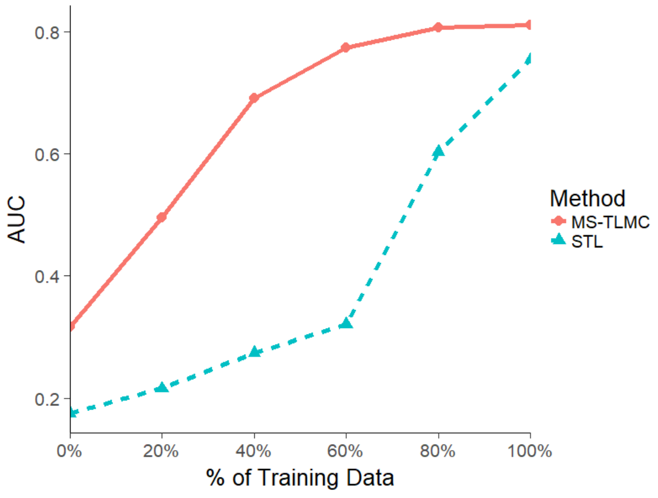

Fig. 3. Averaged AUC Results by STL and MS-TLMC in terms of different proportions of training data.

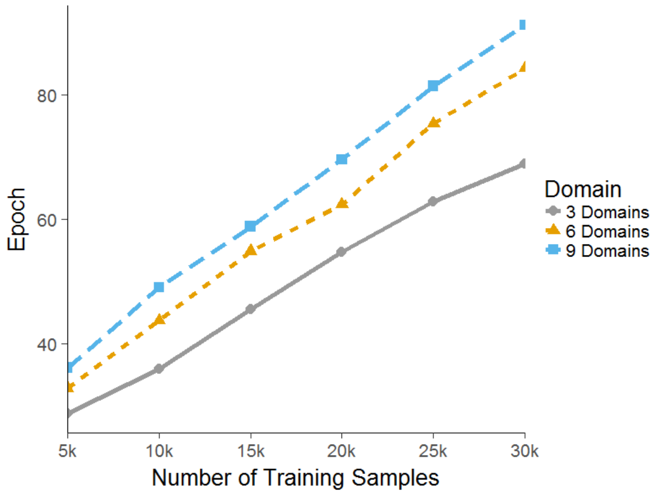

Fig. 4. Averaged AUC Results by STL and MS-TLMC in terms of different proportions of training data.

TABLE I
THE AVERAGED AUC RESULTS BY LODOCV

Method	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**
STL	0.175	0.216	0.274	0.321	0.603	0.754
MS-TLMC	0.317	0.496	0.691	0.774	0.807	0.811

to iteratively choose one campaign as a target domain and take the rest of them as source domains. Because the class labels of the campaign data is very imbalanced with a low response rate ranging from 0.2% to 7%, the binary classification result of the learning task in a chosen target domain are evaluated by Area Under the Curve (AUC) [46] regarding to the minor class. In this experiment, we use different proportions (0%, 20%, 40%, 60%, 80% and 100%) of training data to train models. If the proportion of training data is 0%, it means that all samples are unlabeled and the learning task is unsupervised; otherwise, there are labeled data and learning task is supervised. This setting allows us to test the model performance on different types of learning tasks. The averaged AUC results of these two transfer learning methods by LODOCV are listed in Tab. I and Fig. 3. It can be found that the proposed MS-TLCM method outperforms STL in terms of the averaged AUC values on the minor class and the advantage of MS-TLMC is more distinguished when training data is small. In unsupervised learning where the proportion of training data is 0, both methods achieves the lowest averaged AUC values at 0.317 and 0.175 for MS-TLMC and STL respectively and the their performance gets better as the proportions of training data increase. For 20%, the averaged AUC values of MS-TLMC and STL are 0.469 and 0.216 and the gap between them grows until the proportion of training data is 60%. This demonstrated that compared to STL, the proposed MS-TLMC method can achieve a better learning result when training data is small which can dramatically increase the efficiency of training a model and decrease the workload of labeling training samples. This also means that the proposed MS-TLMC method has a potential to facilitate the sample labeling work. After 60%, the gap between them decreases gradually and their performance are close.

B. Compatibility Experiment

Then we further evaluate the compatibility of the proposed transfer learning method with other machine learning methods including Logistic Regression (LR), Support Vector Machines (SVM) with Gaussian Kernel function and XGBoost. Specifically, we replace learning function f'_T in Equ. 8 and change the loss functions in Equ. 9 correspondingly. For a selected learning method, its performance is evaluated by the averaged AUC of all domains and for each domain, its performance is evaluated by ten-fold-cross-validation while the performance of the corresponding customized MS-TLMC method is evaluated by LODOCV. Then the comparable results of these three machine learning methods are listed in Tab. II. It can be found that the customized MS-TLMC methods generally outperformed their corresponding machine learning methods and the huge gap of AUC values stays at around 40% to 60% which demonstrates that the proposed transfer learning method can efficiently train a model with a small training data. The performance of both LR and SVM is worse as they are weak to deal with imbalanced data while their customized MS-TLMC methods worked much better as there is weighting function to balance class labels. XGBoost achieves similar performance with its customized MS-TLMC method when training data is large. This shows that the proposed transfer learning method can achieve similar learning performance with training a new model. Another interesting finding from this table is that the proposed MS-TLMC works stable with different machine learning methods because it is constructed as an ensemble classification framework.

TABLE II
AUC RESULTS OF THE COMPATIBILITY EXPERIMENT

Method	20%	40%	60%	80%	100%
LR	0.114	0.208	0.247	0.477	0.513
MS-TLMC (LR)	0.378	0.548	0.716	0.721	0.733
SVM	0.130	0.199	0.26	0.479	0.493
MS-TLMC (SVM)	0.447	0.653	0.715	0.733	0.777
XGBoost	0.197	0.264	0.274	0.572	0.748
MS-TLMC (XGBoost)	0.459	0.658	0.744	0.751	0.759

C. Model Efficiency Test

Finally, we evaluate the efficiency of the proposed transfer learning methods in terms of using different number of source domains and training samples. Specifically, its efficiency is measured by the number of training epochs to converge and the convergence threshold is chosen as 0.0001. Fig. 4 illustrates that the proposed method run for more epochs to converge as the number of domain sources and training samples increases. We also find that with the same number of training samples, increasing the number of source domains did not cost much more epochs to converge.

VII. CONCLUSION

In this paper, we proposed a multiple source based transfer learning framework which can use previous knowledge generated from source domains and apply it on target domains. To fully utilize knowledge generated by source domains, the proposed method considers the transfer of both instances and models. In this way, the learning process in target domains is very efficient and a small training set is sufficient to significantly improve model performance. The experimental results demonstrated that the proposed MS-TLMC outperformed the scalable transfer learning framework on a set of campaign data in supervised and unsupervised inductive and transductive learning. The proposed transfer learning framework is also flexible. It can be compatible with different machine learning models including Logistic Regression, Support Vector Machines and XGBoost. In the future, we plan to further explore its usage to other fields and test its performance on balanced datasets. Another potential research direction is to extend it to deal with regression problems.

REFERENCES

[1] L. de Vries, S. Gensler, and P. S. Leeflang, "Effects of traditional advertising and social messages on brand-building metrics and customer acquisition," *Journal of Marketing*, vol. 81, no. 5, pp. 1–15, 2017.

[2] M. R. Solomon, *Consumer behavior: Buying, having, and being*. Prentice Hall Upper Saddle River, NJ, 2014, vol. 10.

[3] S. Brin, R. Motwani, and C. Silverstein, "Beyond market baskets: Generalizing association rules to correlations," in *Acm Sigmod Record*, vol. 26, no. 2. ACM, 1997, pp. 265–276.

[4] K. W. Wong, S. Zhou, Q. Yang, and J. M. S. Yeung, "Mining customer value: From association rules to direct marketing," *Data Mining and Knowledge Discovery*, vol. 11, no. 1, pp. 57–79, 2005.

[5] G. S. Linoff and M. J. Berry, *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, 2011.

[6] T. L. Albrecht, "Advances in segmentation modeling for health communication and social marketing campaigns," *Journal of health communication*, vol. 1, no. 1, pp. 65–80, 1996.

[7] Y. H. Cho, J. K. Kim, and S. H. Kim, "A personalized recommender system based on web usage mining and decision tree induction," *Expert systems with Applications*, vol. 23, no. 3, pp. 329–342, 2002.

[8] P. Wang, W. Sun, D. Yin, J. Yang, and Y. Chang, "Robust tree-based causal inference for complex ad effectiveness analysis," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 2015, pp. 67–76.

[9] B. Schmitt, "Experiential marketing," *Journal of marketing management*, vol. 15, no. 1-3, pp. 53–67, 1999.

[10] A. Plant, J. A. Montoya, H. Rotblatt, P. R. Kerndt, K. L. Mall, L. G. Pappas, C. K. Kent, and J. D. Klausner, "Stop the sores: the making and evaluation of a successful social marketing campaign," *Health Promotion Practice*, vol. 11, no. 1, pp. 23–33, 2010.

[11] K. Coussement, F. A. Van den Bossche, and K. W. De Bock, "Data accuracy's impact on segmentation performance: Benchmarking rfm analysis, logistic regression, and decision trees," *Journal of Business Research*, vol. 67, no. 1, pp. 2751–2758, 2014.

[12] T. Dietrich, S. Rundle-Thiele, and K. Kubacki, *Segmentation in Social Marketing: Process, Methods and Application*. Springer, 2016.

[13] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, 2016.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[16] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[17] F. Huang, A. Ahuja, D. Downey, Y. Yang, Y. Guo, and A. Yates, "Learning representations for weakly supervised natural language processing tasks," *Computational*

Linguistics, vol. 40, no. 1, pp. 85–120, 2014.

[18] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, "Transfer learning using computational intelligence: a survey," *Knowledge-Based Systems*, vol. 80, pp. 14–23, 2015.

[19] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *arXiv preprint arXiv:1705.02364*, 2017.

[20] C. Devin, A. Gupta, T. Darrell, P. Abbeel, and S. Levine, "Learning modular neural network policies for multi-task and multi-robot transfer," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2169–2176.

[21] S. Bickel, C. Sawade, and T. Scheffer, "Transfer learning by distribution matching for targeted advertising," in *Advances in neural information processing systems*, 2009, pp. 145–152.

[22] L. Zhao, S. J. Pan, and Q. Yang, "A unified framework of active transfer learning for cross-system recommendation," *Artificial Intelligence*, vol. 245, pp. 38–55, 2017.

[23] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer joint matching for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1410–1417.

[24] B. Wang and J. Pineau, "Online boosting algorithms for anytime transfer and multitask learning." in *AAAI*, 2015, pp. 3038–3044.

[25] C. Perlich, B. Dalessandro, T. Raeder, O. Stitelman, and F. Provost, "Machine learning for targeted display advertising: Transfer learning in action," *Machine learning*, vol. 95, no. 1, pp. 103–127, 2014.

[26] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.

[27] L. Y. Pratt, "Discriminability-based transfer between neural networks," in *Advances in neural information processing systems*, 1993, pp. 204–211.

[28] S. Thrun and T. M. Mitchell, "Lifelong robot learning," *Robotics and autonomous systems*, vol. 15, no. 1-2, pp. 25–46, 1995.

[29] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[30] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1633–1685, 2009.

[31] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[32] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 5, pp. 1019–1034, 2015.

[33] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1717–1724.

[34] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," *arXiv preprint arXiv:1707.07012*, 2017.

[35] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7304–7308.

[36] F. Hill, K. Cho, S. Jean, and Y. Bengio, "The representational geometry of word meanings acquired by neural machine translation models," *Machine Translation*, pp. 1–16, 2017.

[37] F. Hill, K. Cho, and A. Korhonen, "Learning distributed representations of sentences from unlabelled data," *arXiv preprint arXiv:1602.03483*, 2016.

[38] B. McCann, J. Bradbury, C. Xiong, and R. Socher, "Learned in translation: Contextualized word vectors," in *Advances in Neural Information Processing Systems*, 2017, pp. 6297–6308.

[39] T. Hastie, R. Tibshirani, and J. Friedman, "Overview of supervised learning," in *The elements of statistical learning*. Springer, 2009, pp. 9–41.

[40] M. Long, J. Wang, G. Ding, S. J. Pan, and S. Y. Philip, "Adaptation regularization: A general framework for transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 5, pp. 1076–1089, 2014.

[41] M. Zinkevich, M. Weimer, L. Li, and A. J. Smola, "Parallelized stochastic gradient descent," in *Advances in neural information processing systems*, 2010, pp. 2595–2603.

[42] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy, "Training deep neural networks on imbalanced data sets," in *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016, pp. 4368–4374.

[43] Q. Meng, D. Catchpoole, D. Skillicorn, and P. J. Kennedy, "Dbnorm: normalizing high-density oligonucleotide microarray data based on distributions," *BMC bioinformatics*, vol. 18, no. 1, p. 527, 2017.

[44] B. D. Ripley, "The r project in statistical computing," *MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network*, vol. 1, no. 1, pp. 23–25, 2001.

[45] G. Sidorov, A. Gelbukh, H. Gómez-Adorno, and D. Pinto, "Soft similarity and soft cosine measure: Similarity of features in vector space model," *Computación y Sistemas*, vol. 18, no. 3, pp. 491–504, 2014.

[46] E. Alpaydin, *Introduction to machine learning*. MIT press, 2014.