<table>
<thead>
<tr>
<th>Title</th>
<th>Combating Negative Transfer From Predictive Distribution Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Seah, Chun-Wei; Ong, Yew-Soon; Tsang, Ivor W.</td>
</tr>
<tr>
<td>Date</td>
<td>2012-09-29</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10220/39657">http://hdl.handle.net/10220/39657</a></td>
</tr>
<tr>
<td>Rights</td>
<td>© 2012 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at: [<a href="http://dx.doi.org/10.1109/TSMCB.2012.2225102">http://dx.doi.org/10.1109/TSMCB.2012.2225102</a>].</td>
</tr>
</tbody>
</table>
Combating Negative Transfer From Predictive Distribution Differences

Chun-Wei Seah, Yew-Soon Ong, and Ivor W. Tsang

Abstract—Domain adaptation (DA), which leverages labeled data from related source domains, comes in handy when the label information of the target domain is scarce or unavailable. However, as the source data do not come from the same origin as that of the target domain, the predictive distributions of the source and target domains are likely to differ in reality. At the extreme, the predictive distributions of the source domains can differ completely from that of the target domain. In such cases, using the learned source classifier to assist in the prediction of target data can result in prediction performance that is poorer than that with the omission of the source data. This phenomenon is established as negative transfer with impact known to be more severe in the multiclass context. To combat negative transfer due to differing predictive distributions across domains, we first introduce the notion of positive transferability for the assessment of synergy between the source and target domains in their prediction models, and we also propose a criterion to measure the positive transferability between sample pairs of different domains in terms of their prediction distributions. With the new measure, a predictive distribution matching (PDM) regularizer and a PDM framework learn the target classifier by favoring source data with large positive transferability while inferring the labels of target unlabeled data. Extensive experiments are conducted to validate the performance efficacy of the proposed PDM framework using several commonly used multidomain benchmark data sets, including Sentiment, Reuters, and Newsgroup, in the context of both binary-class and multiclass domains. Subsequently, the PDM framework is put to work on a real-world scenario pertaining to water cluster molecule identification. The experimental results illustrate the adverse impact of negative transfer on several state-of-the-art DA methods, whereas the proposed framework exhibits excellent and robust predictive performances.

Index Terms—Domain adaptation (DA), logistic regression (LR), negative transfer, predictive distribution matching (PDM), support vector machines (SVMs).

I. INTRODUCTION

In traditional learning tasks, the decision function \( f \) is typically attained by minimizing the expected risk functional of the form

\[
\min_f \int L(x, y, f) dP(x, y)
\]

with respect to the joint distribution \( P(x, y) \) of the target domain and a loss function \( L \), where \( x \) and \( y \) denote the input feature vector and class label, respectively, of the problem of interest. The joint distribution can be further factorized as \( P(x, y) = P(y|x)P(x) \), where \( P(y|x) \) and \( P(x) \) are the predictive distribution and marginal distribution of the target domain, respectively. However, when no label information is available in the target domain, \( P(y|x) \) cannot be well estimated. To address this, some are turning to domain adaptation (DA) techniques which use labeled data from related source domains to augment the performance of learning tasks on the target domain.

Taking the marketing strategy of sales personnel as a motivating example, for instance, sentiment classification serves as an important task to predict the sentiment polarity of a new product (target domain) based on the multiway scale of user reviews collected from other similar products. Each review is graded based on a five-star rating, and the higher the rating, the better the feedback is perceived. Since the user review feedbacks are usually described by some common words, the annotated sentiment reviews from several other categories of products (source domains) may benefit the prediction of the star rating on unannotated sentiment reviews of new products (target domain). Hence, DA methods generally assume that the source domains share a similar predictive function with the target domain. Aside from sentiment classification, DA methods are also widely studied in natural language processing [1]–[3], text categorization [4], computer vision [5]–[8], Wi-Fi localization [4], remote sensing [9], and recommendation systems [10]. Other applications in which DA can be useful also exist in abundance, including gene expression data [11], cell-phenotype images [12], and aerodynamic design [13], where labeled data in the target domain of interest are generally scarce.

Recent DA methods [5], [14]–[16] have been proposed for learning from multiple source domains. In [16], for instance, the authors proposed a multiple convex combination of support vector machine (SVM) using the data from multiple source domains and the target domain. Since the data do not come from the same origin, the distributions \( P(x, y) \) of the source and target domains are likely to differ. In such situation, target
labeled data are often required to measure the relatedness from the source domains [5], [14], [17]; then, source domains can assist in learning the target task. However, when target labeled data are unavailable, DA methods [1], [18]–[20] usually assume that the prediction distribution \( P(y|x) \) is shared among different domains and minimize the dissimilarities among the source and target domains with regard to the marginal distribution \( P(x) \) only. The dissimilarity in marginal distributions among domains is commonly known as covariate shift [20] that adjusts \( x \) and target domains with regard to the marginal distribution \( P(x) \) only. The dissimilarity in marginal distributions among domains is commonly known as covariate shift [20] that adjusts the weight of each source sample by means of \( P^t(x)/P^s(x) \) as a common remedy used to resolve such an issue, with \( P^s(x) \) and \( P^t(x) \) denoting the marginal distribution of the source domain and target domain, respectively. For example, the kernel-mean matching (KMM) method [18] estimates the weight of each source sample by minimizing the maximum mean discrepancy (MMD) criterion [21] between the source samples and target unlabeled samples; then, reweighted source samples are used for training a classifier for the target data.

As an alternative to reweighting methods, others have also considered the extraction of useful features from the source domains to augment the original feature space to train the classifier [3], [22]–[26] so that the augmented feature space leads to similar marginal distribution between the source and target domains. For instance, an alternative to KMM that minimizes the MMD criterion for the purpose of minimizing the distributions of the source and target domains, minimizing the quadratic distance [27] and geodesic distance [25], is recently proposed. For another instance, the feature augmentation (FA) approach [3] augments features belonging to the same domain by twice that of the original features to bias the classifier in treating the data of the same domain twice as much than those of differing domains. Furthermore, the FA approach is also considered as a multitask learning algorithm since its model parameters \( \theta_r \) in the \( r \)th domain are decomposed as \( \theta_c \) and \( \theta'_r \), where \( \theta_c \) is shared among all domains and \( \theta'_r \) is for each individual domain.

In general, multitask methods [28], [29] simultaneously learn the models of all the tasks by sharing some common parameters such that the learned model can classify each individual task well, whereas DA methods focus on classifying well on the target task only. Another major difference is that the source and target tasks in DA are the same but different in data distribution, whereas the tasks in multitask are different but related. In particular, DA methods generally address the marginal distribution differences between the source and target domains, and this paper further combats negative transfer from predictive distribution differences. In contrast, task clustering for multitask learning discovers hidden structure within a set of related tasks for a robust learning [29], [30].

Frankly, each domain has its own predictive distribution \( P(y|x) \) in real applications; as a result, the phenomenon of negative transfer [31] is known to creep in, leading to the impediment on the performances of DA approaches [31], [32]. Thus, negative transfer can be deem to have occurred when the DA method is observed to deteriorate over the prediction performance of its respective non-DA counterpart. To this end, more recent works, including domain adaptation SVM (DASVM) [9], maximal margin target label learning (MMTLL) [33], and TARget learning Assisted by Source Classifier Adaptation (TARASCA) [34], have also attempted to maintain the consistency of the joint distributions \( P(x, y) \) across the different domains.

In spite of the recent advancements made in DA, many fundamental problems of negative transfer resulting from the differences in predictive distribution have remained unresolved. In particular, to perform well, reweighting methods require the source and target domains to share similar predictive distributions. Furthermore, a considerable large number of target labeled data are typically required to robustly reweight the training instances reliably. Like general DA approaches, feature DA techniques are also plagued by the issues of predictive distribution dissimilarities among the source and target domains. When many overlapping sources and target data with conflicting class label information exist, general DA approaches such as DASVM do not function well. Moreover, DASVM is unable to deal with multiclass problems as the adopted progressive transductive SVM [35] strategy in DASVM is unable to infer multiclass pseudolabels of target unlabeled patterns. In contrast to the proposed method, MMTLL and TARASCA consider classifier/model-based transfer by choosing the weights of the source classifiers (among many source classifiers with different bias parameters) via cluster assumption that exists in the target unlabeled data for positive transfer.

From our survey of the literature, some of the core roots of negative transfer due to predictive distribution differences (which generally violate the assumptions of many DA methods) can be summarized as follows.

1) **Conflicting class labels between related source domains:** The domains contain sample data or clustered data with conflicting class labels. For instance, a domain with a class label which differs from the majority of related domains having a common class label, in some localized region of the vector space, is established as an outlier.

2) **Sample selection bias due to imbalanced class distribution:** The sparsity of labeled data does not serve as good representations of the general population [9], particularly for the imbalance problem where bias exists when estimating the target predictive distribution (e.g., by logistic regression (LR) or naïve Bayes classifier); this bias is generally known as the sample selection bias.\(^1\)

In most cases, the source domains have differing class distributions from the target domain, and these class distribution differences can easily lead to predictive distribution dissimilarities among the domains. Without sufficient label information on the target domain, the true class distribution of the target domain is generally unknown, and resampling strategies (e.g., Synthetic Minority Over-sampling Technique [37]) that are designed to adapt the class distributions of the source domains to match with the target domain do not apply well in this setting. Even in the event of high similarity in the feature space, any class distribution differences between the target and source domains can still mislead the learning of the target domains.

---

\(^1\)Recently, a work [36] considers that the class distribution of the training set differs from the testing set while both sets are from the same domains.
predictive distribution due to the learning biases toward that of the source domains.

Taking these cues, in this paper, we propose a novel DA method, namely, predictive distribution matching (PDM), to address the challenges that arise from the predictive distribution differences. The main contributions and core ingredients of the proposed framework are outlined in what follows.

1) A criterion of positive transferability is proposed to measure the differing predictive distributions of the target domain and the related source domains. Using this criterion, a PDM regularized classifier is introduced to infer target pseudolabeled data which subsequently assists in identifying the relevant source data in a manner that the predictive distributions of both source and target data are maximally aligned.

2) To our knowledge, the proposed framework serves as the first attempt to combat negative transfer in multiclass problems, where target labeled data are unavailable or scarce. In particular, a new form of LR, established here as the PDM-LR, is proposed for handling the multiclass DA problem.

This paper extends from the preliminary work [38] and is organized as follows. Section II introduces the proposed PDM framework to match the predictive distributions of the source and target domains in the context of multiclass problems. Instantiations of the PDM framework on LR and SVM are subsequently showcased in Section III. Extensive experimental studies of the PDM framework, pitted against several state-of-the-art DA and traditional algorithms on multiclass and multidomain data sets, including the real-world Sentiment data set, are reported in Section IV. Analysis and discussion pertaining to the experimental results are then provided in Section V. In addition, a novel real-world water-molecule application is showcased in Section VI. The brief conclusions of this paper are then drawn in Section VII. In contrast to this paper, the initial work [38] of this paper focuses on only binary problems and SVM context. The core symbols used throughout the rest of this paper are listed in Table I.

A. PDM Regularization for Multiple Source Domains

Here, our interest is on PDM across multidomains and the multiclass contexts. First, we define the notions of positive transferability and negative transferability.

Definition 1: Positive transferability is introduced as the assessment of the synergy between the source and target domains in their prediction models. In other words, it measures the constructive synergy of the source labeled data in accelerating or enhancing the learning of the prediction model for the target unlabeled data. This is highly plausible when the selected source labeled data and the set of target data share similar predictive models.

Next, the antonym of positive transferability, which is referred as negative transferability, is defined.

Definition 2: Negative transferability is introduced as a measure for the destructive synergy of the source labeled data in accelerating or enhancing the learning of the prediction model for the target unlabeled data. This is highly plausible when the selected source labeled data and the set of target data share similar predictive models.

Next, the antonym of positive transferability, which is referred as negative transferability, is defined.

Definition 2: Negative transferability is introduced as a measure for the destructive synergy of the source labeled data in accelerating or enhancing the learning of the prediction model for the target unlabeled data.

With the notion of positive transferability given by Definition 1, we proposed a criterion to measure the degree of positive transferability between a sample pair of different domains in terms of their predictive distributions.

<table>
<thead>
<tr>
<th>Table I</th>
<th>SYMBOL DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>Total number of domains, the first (m-1) domains represent source domains while m denotes the target domain</td>
</tr>
<tr>
<td>C</td>
<td>Number of Class Labels</td>
</tr>
<tr>
<td>x_r</td>
<td>Feature vector of rth data of rth domain</td>
</tr>
<tr>
<td>y_r</td>
<td>Class Label (1,...,C) of the data x_r. When r == m, it refers to the rth inferred pseudo-label from the target unlabeled data</td>
</tr>
<tr>
<td>n_r</td>
<td>Number of labeled data in rth source domain or the number of pseudo-labeled data in target domain. For simplicity, the index for jth iteration after inferring pseudo-labeled data and removing the irrelevant source data is not shown</td>
</tr>
<tr>
<td>n</td>
<td>(\sum_{r=1}^{m} n_r)</td>
</tr>
<tr>
<td>D_L</td>
<td>(D_L = \cup_{i=1}^{n} {x_i, y_i}), all labeled data in all source domains</td>
</tr>
<tr>
<td>D_L</td>
<td>Remaining of source labeled data in jth iteration. Note, D_L^0 = D_L</td>
</tr>
<tr>
<td>D_U</td>
<td>The set of unlabeled data in target domain</td>
</tr>
<tr>
<td>B_i</td>
<td>Number of inferred pseudo-labeled data from the target (rth domain) unlabeled data during the jth iteration, i.e. ({x_i^{m}, y_i}) where y_i is the pseudo-label of data x_i</td>
</tr>
<tr>
<td>D_U</td>
<td>D_U (\setminus B_i), Note, D_U^0 = D_U</td>
</tr>
<tr>
<td>(\eta)</td>
<td>Number of features</td>
</tr>
<tr>
<td>P_r(x)</td>
<td>Marginal distribution of rth domain</td>
</tr>
<tr>
<td>P_r(y</td>
<td>x)</td>
</tr>
<tr>
<td>I(·)</td>
<td>Indicator function which has a logic of 1 if the predicates hold, otherwise a logic of 0 is given</td>
</tr>
<tr>
<td>W_{ri}</td>
<td>The positive transferability criterion between the rth and jth samples in rth and dth domains, respectively</td>
</tr>
</tbody>
</table>
minimize the predictive function \( \omega \) required to compute only once before the iterative process.

The acquired pseudolabeled data of the \( j \)-th iteration class \( c \) is then

\[
B^j_c = \{ (x_i, y_i) \in T^j_c \mid |1 \leq i \leq p^j_c \}\]

where \( p^j_c = \min(|T^j_c|) \), \( |T^j_c| \) is the cardinality of \( T^j_c \) and \( \sigma \) is a relaxation parameter.\(^3\) \( T^j_c \) denotes the unlabeled data with class labels \( c \) inferred using the PDM regularized classifier \( f^j \).

\(^3\)A higher value of \( \sigma \) would speed up the process, but at the expense of including less confident pseudolabeled data. Since the efficiency of the process is not a major concern, \( \sigma \) is configured to 1 in our study.
and is sorted in a decreasing order in terms of $P(y = c|x)$ as follows:

$$T_c^j = \left\{ \left( x_i, c \right) | x_i \in D_U^j, h(x_i) = c, P(y = c|x_i) \geq P(y = c|x_{i+1}) \right\}$$

where $D_U^j = D_U^{j-1} \setminus B_j^{j-1}$ and $P(y = c|x_i)$ is the predictive distribution for class $c$ given $x_i$ that is estimated from both $B_j^{j-1}$ and $D_U^{j-1}$ (which is the identified source data in the $(j-1)$th iteration and is presented in the next Section II-C), while the predicted class $h(.)$ is defined as follows:

$$h(x) = \arg \max_{c \in C} \omega_c(x)$$

with $\omega_c(x)$ denoting the predictive output of PDM regularized classifier $f^2$ for class $c$. Hence, $B_c^j$ in (4) (the acquired $p_c^j$ number of pseudolabeled data) represents the data with the highest predictive distribution values in $T_c^j$. After the new set of pseudolabeled data is formed in (3), the PDM framework reestimates $P^m(y|x)$, for instance, by means of LR. The inferred pseudolabeled data formed in (3) can then be used to compute the positive transferability on (2) for the next iteration.

C. Removing Irrelevant Source Data

In practice, some source data may not align with the predictive distribution of the inferred target pseudolabeled data. Hence, in this section, we discuss how these irrelevant source data are removed.

Without loss of generalities, the remaining source labeled data at the $j$th iteration can be defined as

$$D_L^j = D_L^{j-1} \setminus \left\{ \cup_{c \in C} D_c \right\}$$

where $D_L^j$ is the initial set of source labeled data ($D_L$) and $D_c$ is the set of data grouped according to their true class label $y_i$. Each grouped set is then sorted according to their estimated predictive distribution values in ascending order as follows:

$$D_c = \left\{ \left( x_i, y_i \right) | x_i \in D_L^{j-1}, y_i = c, P^m(y = c|x_i) \leq P^m(y = c|x_{i+1}) \right\}$$

where $\pi$ denotes the minimum level of confidence for any data vectors to be retained and $P^m$ is estimated by using the pseudolabeled data $B_j$. The removed source data for each class $D_c$ are the lowest $\gamma$ consistence data points with respect to the pseudolabeled data. After convergence is reached, all source labeled data with $P^m(y|x, f^j) \leq \pi$ are removed by simply setting $\gamma$ to $\infty$.

Since the inferring process is designed to iteratively select the highly confident pseudolabeled data, it is natural to end the PDM process when the inferred labels of the pseudolabeled data fail to measure up to the given confidence level. In this paper, parameter $\delta$ is used to control the level of confidence in the pseudolabeled data as

$$\min_{x_i \in B} P^m(y_i|x_i) \leq \delta.$$  

III. PDM REGULARIZED CLASSIFIER INSTANTIATIONS

In this section, instantiations of the PDM framework with LR (PDM-LR) and SVM (PDM-SVM) are presented as the PDM regularized classifier.

A. PDM LR Classifier (PDM-LR)

LR is primarily popular in the context of text classification. On multiclass classification problems, the predictive distribution $P(y = c|x)$ of a class $c$ is defined as follows:

$$\omega_c(x) = P(y = c|x) = \frac{e^{\beta^T x}}{\sum_{c=1}^{C} e^{\beta^T x}}$$

where $\beta_c$ is the weight vector for class $c$ and $\omega_c(x)$ is the predictive output for class $c$. In LR, both $P(y = c|x)$ and $\omega_c(x)$ have the same value, and $\sum_{c} \omega_c(x) = 1$ can be regarded as a form of probability measure. On multiclass LR, minimizing the negative log likelihood of (10) becomes

$$g_1(\beta) = -\sum_{r=1}^{m} \sum_{i=1}^{n_r} w_i^r \beta^T x_i^r + \sum_{r=1}^{m} \sum_{i=1}^{n_r} w_i^r \log \sum_{z=1}^{C} e^{\beta^T x_i^z}$$

where $w_i^r$ is the weight of the $i$th sample in the $r$th domain and $\beta = [\beta_1, \ldots, \beta_C]$. To prevent overfitting, a regularizer with parameter $C_1$ is typically incorporated

$$g_2(\beta) = \frac{1}{2} C_1 \| \beta \|^2_2.$$  

Hence, a regularized multiclass LR is defined as

$$\arg \min_{\beta} g_1(\beta) + g_2(\beta).$$  

Next, with the $\omega(x)$ of the PDM regularizer in Definition 4 given by $\beta^T x$, the resultant PDM regularizer is then formulated as follows:

$$g_3(\beta) = \frac{C_2}{2n^2} \sum_{r,d=1}^{m} \sum_{i,j=1}^{n_r} \left( \beta^T_i x_i^r - \beta^T_j x_j^d \right)^2 W_{ij} \delta[r \neq d]$$

where $C_2$ denotes the parameter that regulates the importance of PDM. Hence, combining the PDM regularizer with (13), the proposed PDM multiclass LR or PDM-LR in short becomes

$$\arg \min_{\beta} G(\beta) = \arg \min_{\beta} g_1(\beta) + g_2(\beta) + g_3(\beta).$$

Note that a higher $\gamma$ value in (7) speeds up the process, but it will remove high confidence source labeled data that are relevant to the target domain. As speeding up the process is not our main objective here, $\gamma$ is configured as 1 in our study.

This weight can be adjusted to define the importance of each data vector sample to deal with the marginal distribution differences among domains, often known as the reweighting method [18]. Since the focus here is on predictive distribution differences, we treat all data samples equally in this paper, i.e., assigning each $w_i^r$ to $1/n$. 

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.
1) Convergence Analysis: In what follows, the details toward solving (15) and the properties of convergence are presented. Note that the PDM regularizer in Definition 4 represents a special case of [41] which is positive semidefinite, and LR is strongly convex [39]. Hence, the resultant PDM-LR is also strongly convex. Thus, the problem can be solved using convex optimization technique, such as the coordinate descent method due to its simplicity and efficiency, since the computation of the entire Hessian matrix is not required. The coordinate descent method is composed of an outer and an inner loop. The inner loop conducts the Newton descent search on a dimension while the outer loop checks for convergence. An outline of the coordinate descent method is depicted in Algorithm 1.

Algorithm 1 Coordinate descent method

1: repeat
2: for $p = 1 \& c = 1$ TO $\eta \& C$, respectively, do
3: Solve $\min_{\beta_{pc}} G(\beta)$, by means of approximation, to obtain $\beta_{pc}$.
4: $\beta_{pc} = \beta_{pc} + z$
5: end for
6: until $\beta$ is optimal
7: return $\beta$

To attain the gradient and Hessian information of each coordinate, the derivations for LR, (13), follow that of [39]. The gradient and Hessian formulations for the PDM regularizer in (14), on the other hand, are derived in what follows. The first derivative of (14) can be derived as

$$\frac{\partial g_3}{\partial \beta_{pc}} = \frac{C_2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (W_{ij} (\beta'_{c} x_{i} - \beta'_{c} x_{j}))$$

$$\times (I[y_i = c] x_{ip} - I[y_j = c] x_{jp}) = \frac{C_2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (W_{ij} I[y_i = c] (\beta'_{c} x_{i} - \beta'_{c} x_{j}))$$

$$\times (x_{ip} - x_{jp})$$

(16)

since $W_{ij}$ consists of $I[y_i = y_j]$ where the $pc$ of $\beta_{pc}$ is the $p$th dimension and the class $c$ of $\beta$. Note that, for the sake of conciseness in (16), the notation for the PDM regularizer in (14) is simplified with the removal of domain indexes, i.e., $r$ and $d$, since the indicator, $I[r \neq d]$, is a precomputed value and will implicitly inherit the domain indexes. Without loss of generality, the PDM regularizer in (14) simplifies to

$$g_3(\beta) = \frac{C_2}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\beta'_{y_{i}} x_{i} - \beta'_{y_{j}} x_{j})^2 W_{ij}.$$  

(17)

In addition, the second-order information, i.e., the Hessian of (17), is then derived as

$$\frac{\partial^2 g_3}{\partial^2 \beta_{pc}} = \frac{C_2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} I[y_i = c] (x_{ip} - x_{jp})^2.$$  

(18)

In what follows, the details of updating $\beta_{pc}$ using a scaling factor $z$ that meets the sufficient decrease condition will be discussed. $E(p, c) \in \mathbb{R}^{n \times C}$ defines the direction of dimension $p$ and class $c$ for updating the value $\beta_{pc}$ as

$$E_{ij}(p, c) = \begin{cases} 1, & i = p, j = c \\ 0, & \text{otherwise.} \end{cases}$$  

(19)

Updating of new $\beta_{pc}^{new}$ as $\beta_{pc}^{new} = \beta_{pc}^{old} + z$ is equivalent to performing $\beta + zE(p, c)$. With this update, the value of PDM-LR function $G$ in (15) decreases as follows:

$$D(zE(p, c)) = G(\beta + zE(p, c) - G(\beta).$$  

(20)

The sufficient decrease condition [39] of $zE(p, c)$ is given by

$$D(zE(p, c)) \leq \sigma z G'(\beta_{pc})$$  

(21)

where $z = \lambda d$ and $d = -G'((\beta_{pc})/G''((\beta_{pc}))$ denotes the Newton direction. From [39, Theorem 4], there exists the parameter $\lambda = 1, 0.5, 0.5^2, 0.5^3, \ldots$ that satisfies the condition in (21) with $\sigma \in (0, 0.5)$. Here, our search begins with $\lambda = 1$, followed by a check on the sufficient decrease condition of (21). If the condition is violated, $\lambda$ is reduced by half repeatedly, until the inequality of (21) is satisfied. The $D(zE(p, c))$ of each component in (15) can be formulated as $D(zE(p, c)) = D_1(zE(p, c)) + D_2(zE(p, c)) + D_3(zE(p, c))$, where $D_1, D_2,$ and $D_3$ denote the reduction values of the function $g_1, g_2,$ and $g_3$ in (15), respectively. While the derivatives of $D_1$ and $D_2$ follow that of [39], $D_3$, on the other hand, is derived as

$$D_3(zE(p, c)) = g_3(\beta + zE(p, c)) - g_3(\beta) = z \frac{C_2}{n^2} \sum_{i=1,j=1}^{n} 2W_{ij} I[y_i = c]$$

$$\times \left( (x_{ip} - x_{jp}) \beta'_{c} x_{i} + (x_{jp} - x_{ip}) \beta'_{c} x_{j} \right) + z^2 \frac{C_2}{n^2} \sum_{i=1,j=1}^{n} \left( W_{ij} I[y_i = c] (x_{ip}^2 + x_{jp}^2 - 2x_{ip}x_{jp}) \right).$$  

(22)

The stopping criterion of the outer loop is defined as $\|G'\| < \epsilon$, where $\epsilon$ is a predefined parameter (see $G'$ in (21)).

2) Computational Complexity of Algorithm 1: From [39, Theorem 5], it can be derived that the gradient, Hessian, and reduction functions of $g_1$ and $g_2$ in (15) have a total computational complexity of $O(n)$. In PDM, (16), (18), and (22) are computed. It is worth noting that the matrix $W$ of PDM is sparse, with at most $k$ nonzero values that denote the $k$ nearest neighbor of each sample [41]. Since matrix $W$ has only at most $kn$ nonzero values, the computations of (16), (18), and (22) equate to $O(kn)$ per $\beta_{pc}$. Since (18) and the second component of (22) are independent of $\beta$, both need to be computed only once and can be cached for subsequent reuse throughout the coordinate descent process. From the aforementioned computational analysis, the computational complexity of each inner loop in Algorithm 1 totals to $O(knC'\eta)$. This, however, can be reduced to $O(kn\eta)$ due to the existence of the $I[y_i = c]$ term in
The parameters of all the methods are configured by means of maximizing the log likelihood on the training data [40]. From (26), $P(y = 1|x) = \frac{1}{1 + e^{A\omega(x) + B}}$

where both $A$ and $B$ are determined by means of maximizing the log likelihood on the training data [40]. From (26), $P(y = 1|x) = 1 - P(y = 1|x)$.

### IV. EXPERIMENTAL SETUP

In this section, an experiment study of the proposed PDM framework, i.e., PDM-LR, is carried out on synthetic multi-class and multidomain data sets and a real-world Sentiment problem.

#### A. State-of-the-Art Algorithms

In this paper, a plethora of supervised, semisupervised, and DA state-of-the-art algorithms are considered for comparison as summarized in Table II, where $D_L$ denotes the data of all the available source domains and $D_U$ is the target unlabeled data. Furthermore, Table II indicates which algorithms can be directly applied to address the multiclass problem. Since the proposed PDM is integrated with LR for multiclass problems, the variants of LR algorithms considered include LR, FA-LR, and PDM-LR in the experimental study.

The parameters of all the methods are configured by means of $k$-fold source-domain cross-validation, which represents an extension of the $k$-fold cross-validation for DA suggested in [45]. Specifically, each partition is a source domain in the $k$-fold source-domain cross-validation. In addition, a linear kernel is used in the SVM considered in this section, due to its popularity in the text classification domain.
In the binary problem, star ratings 1 and 2 form the positive data, while 4 and 5 form the negative data. To study the mismatch of predictive distributions between the source and target domains, five different positive class ratio (PCR) settings are generated here for investigations. The five PCR settings are chosen from 0.3 to 0.7 at an incremental step size of 0.1. Note that the PCR value defines the percentage of positive data in the source domains. A PCR setting of 0.3, for example, indicates that 60 data vectors have positive class labels while the rest have a negative label, out of the 200 selected data in each source domain.

In the multiclass problem, each star rating is equivalent to a class. Coincidently, the data set consists of an even class number; hence, it becomes possible to study the mismatch in predictive distributions between the source and target domains based on binary problem settings. The same five PCR settings are also generated for the source domains, and the PCR value denotes the total percentage of the reviews of star ratings 1 and 2 in the source domain. In addition, the number of star-rating reviews are chosen to be equal for 1 and 2 and 4 and 5. For a PCR value of 0.3, for example, out of the 200 selected data in each source domain, star ratings 1 and 2 each have 30 reviews while star ratings 4 and 5 each have 70 reviews.

2) Multidomain Reuters Data Set: Three out of four main categories of the Reuters data set, namely, People(Peo), Organizations(Orgs), and Exchanges(Excs), are considered in this paper. The Places category is not considered in this paper due to the vast instances belonging to this category that would overwhelm all other categories, thus making the study fruitless. In each task, the $k$th largest subcategory of a main category is considered as the $k$th domain. The groupings of the domains are detailed in Table IV. Mainly, the largest subcategory is used as the target domain (Reuter-1) while the remaining four largest subcategories form the related source domains (Reuter-2 to Reuter-5).

In the binary context, three resultant tasks are investigated in total: Peo versus Orgs, Peo versus Excs, and Orgs versus Excs. For each task, the $m$th domain of a category is labeled as positive, while the $m$th domain in another category forms the negative data. Note that this experimental setting is consistent to the other works [47]–[49] that also considered the Reuters data set.

In the multiclass context, each category is treated as an individual class. All data in the source domains are used as
labeled data, and for the target domain, the entire data set is used as unlabeled data. Note that this data set contains imbalance positive and negative samples in each subcategory; hence, the class distribution of the target domain is imbalanced, and the predictive distributions of the source domains are quite diverse with respect to one another.

3) Multidomain Newsgroup Data Set: The three main categories of the data set are comp, rec, and sci. We considered four subcategories in each main category and grouped these four subcategories into four domains as described in Table V.

In the multiclass context, each category is treated as an individual class. Since each domain has a significant number of data to be used as the target domain, we generated four tasks from these groupings. Task $m$ uses Newsgroup-$m$ as the target domain, and the rest forms the source domains. All data in the source domains form the labeled data while the entire data set of the target domain serves as the unlabeled data.

V. EXPERIMENTAL STUDY

In this section, we present an empirical study of the PDM framework, particularly PDM-LR, on several commonly used DA benchmark text classification data sets.

A. Binary Classification DA

We begin our study on the performances of various classifiers for each of the four domains in the Sentiment data set, i.e., Book, DVDs, Electronics, and Kitchen Appliances, when used independently as the target domain, are summarized in Fig. 2(a)–(d), respectively. For each of the subfigures, it depicts the testing accuracies obtained for five different PCR settings of the source domains at 0.3 to 0.7. On the other hand, the PCR of the target unlabeled data set is configured at 0.5. Hence, when the PCR of the source domains is also in the region of 0.5, the predictive distributions of the source domains are likely to match that of the target unlabeled data set. Any other PCR settings, on the other hand, would likely result in mismatch of the predictive distributions between the source and target domains.

Note that each of the subgraphs exhibits similar performance trends, where all of the classifiers considered in the study perform optimally at a PCR value of 0.5, while displaying sharp declining accuracies when the PCR is skewed toward either extreme ends, except for the proposed PDM-LR which is designed to handle data sets with unbalanced class labels. It is notable that the larger the discrepancies in predictive distributions between the source and target domains, the greater is the bias found in the target prediction accuracies by the classifiers, which are geared toward the source domains. At both extreme ends of the PCR settings, LapSVM which represents an extension of the traditional SVM is noted to generally fare better than SVM since the former acts to evolve the predictive distribution of the labeled data toward that of the unlabeled data. However, just using the unlabeled data alone in LapSVM does not resolve the issues pertaining to differing predictive distributions between the source and target domains (as denoted by the values of PCR $\neq 0.5$) since LapSVM is observed to underperform PDM-LR in Fig. 2.

At a PCR of 0.5, the target domain shares similar predictive distribution to the source domain. Thus, it is natural to expect the DA algorithms to exhibit similar performances with the traditional classifiers. Nevertheless, FA-LR is observed to obtain improved accuracy over LR in all the subgraphs. In addition, KMM and DASVM also outperform SVM in Fig. 2(a)–(c), respectively. Recall that FA-LR, KMM, and DASVM are the DA versions of the LR and SVM, respectively. The observed improvements thus suggest that FA-LR, KMM, and DASVM can only be beneficial when the predictive distributions among domains match. In either extreme ends of the PCR values, the FA-LR, KMM, and DASVM adaptation methods have reported poorer accuracy compared to their respective non-DA counterparts. Hence, as a summary, it is notable that, when the predictive distributions among domains do not match, DA algorithms generally report lower accuracies than their respective non-DA counterparts; thus, DA algorithms are prone to the effects of negative transfer.

In contrast to existing DA approaches, which suffer from performance degradation due to the effects of negative transfer, PDM-LR effectively unearth the useful knowledge that lies inherent within the multisource domains by means of prediction distribution matching, to arrive at the robust prediction performance observed on the Sentiment data target testing set. In particular, while DA methods in either extreme ends of the PCR values displayed poor prediction accuracies of around 55%, PDM-LR, on the other hand, shows an impressive gain of 15% improvements at an accuracy of 70%. Furthermore, PDM-LR is able to deliver stable results with accuracies that do not deviate over 5% across all the PCR settings. This demonstrates the robustness and reliability of the PDM-LR under different PCR settings by benefiting from the positive transferability facilitated in the proposed framework.

We further experimented the classifiers on the Reuters data set. The results obtained are summarized in Fig. 3. Overall,
it is observed that PDM-LR, KMM, and DASVM outperform all other classifiers considered on the Reuters data set. On the other hand, FA-LR which can also be considered as a multitask learning method for learning a shared parameter model under the unique predictive distribution of each domain can lead to negative transfer since FA-LR underperforms LR in Fig. 3(b) and (c). Notably, Fig. 3(a) shows significant accuracy improvements of PDM-LR over KMM, DASVM, and all others. In Fig. 3(b) and (c), PDM-LR is also shown as competitive to KMM and DASVM. Note that SVM is observed to reach a near full score of 100% accuracy in Fig. 3(c). This suggests the high similarities in predictive distributions among the source and target domains. It is thus reasonable for PDM-LR to perform close to KMM and DASVM. Nevertheless, PDM-LR is generally superior to LR.

**B. Multiclass Classification DA**

The prediction trends of the DA and LR classifiers in the multiclass setting of the Sentiment problem in Fig. 4 are observed to be in agreement with those obtained on the two-class setting (as shown in Fig. 2). Fig. 4 shows that PDM-LR generally outperforms the other counterpart algorithms in almost all the different PCR configurations on the multiclass Sentiment data set. Furthermore, PDM-LR delivers stable results with accuracies that do not deviate over 5% across all the PCR settings. In the multiclass setting, the negative transfer phenomenon of DA is more adverse as shown in the results of FA-LR. This is because the likelihood for differing predictive distributions among domains is likely to happen since more class distributions are considered. Hence, PDM becomes ever more challenging in the multiclass classification context.

The multiclass Newsgroup and Reuters experimental results are next summarized in Fig. 5. Although the class distributions of the Newsgroups are similar across domains, there are signs of the DA methods suffering from negative transfer since FA-LR is observed with lower accuracies than the traditional LR. In addition, KMM also has poorer accuracies than LR. In particular, the causes of low prediction accuracy on the testing set for FA-LR and KMM are likely to be a result of the differing predictive distributions between the training and testing sets and also likely the reason for the lack of robustness in the performances of KMM and FA-LR when compared to LR across the range of PCR values in both the Sentiment and Newsgroup data sets (see Figs. 2, 4, and 5). It is also worth noting in Fig. 5 that LR fares generally better in prediction accuracy on the Reuters than the Newsgroup data sets. Taking this cue, we infer the predictive distributions of the domains in the Reuters data set to bare greater similarities. In comparison to all algorithms, PDM-LR overall exhibits at least 5% improvements in accuracy on all the data sets while attaining at least 10% accuracy enhancements on four out of the five data sets considered. Furthermore, an impressive improvement of up to 20% is observed on Newsgroup-1.

**C. Computational Complexity of PDM-LR Regularized Classifier**

In this section, we analyze the computational complexity of the PDM-LR regularized classifier via empirical study. Fig. 6 summarizes the computational effort involved in the training of a PDM-LR regularized classifier on the various data sets considered. It is observed that the coordinate descent method as described previously in Section III-A2 takes around ten iterations in the outer loop to solve the PDM-LR regularized classifier. The experimental results thus confirm our theoretical complexity analysis of PDM-LR as $O(n^2)$. In addition, the time taken by the classifier in each algorithm is shown in Fig. 7. In particular, the PDM-LR regularized classifier takes an addition of 0.4 s or twice the time for training compared to LR where the additional time is mainly for computing the regularizer term of PDM-LR.
VI. REAL-WORLD COMPLEX PROBLEM: WATER MOLECULES

In this section, we present the application of the proposed approach to the water isomer discovery problem [50]. Water clusters are crucial for understanding the enigmatic properties of water. They are analyzed in biology to study hydrophobic and hydrophilic interactions and elucidate water’s role in biochemical processes which include ligand docking and protein folding [50]. Water clusters are also investigated in physical chemistry to discover the fundamental molecular interactions and collective effects of the condensed phase (liquid and ice) [51]. The identification of water isomers, which are low-energy stable and metastable molecular structures of pure water clusters, is important to study the key properties of the structures.

Obtaining the true computational design of water isomers using mechanical calculation, such as B3LYP [52], is often computationally intractable without the availability of some supercomputing facilities, particularly on large-scale water clusters. To overcome the issue of computational intractability, cost-effective empirical models, including OSS2 [53] and TTM2.1-F [54], have been developed and employed as alternatives to their computationally expensive counterparts. Using the sample sets of isomers collected from past sampling on the different models, we aim to predict the true water isomers in B3LYP, thus reducing the time effort that would otherwise be spent on exhaustive sampling using the expensive mechanical calculations. Here, four source water isomer data sets have been collected from past sampling processes on the different potential energy models, which are summarized in Table VI. Data Sources 1 and 2 were obtained from OSS2 while Sources 3 and 4 were obtained from TTM2.1-F. Here, the sparse data set archived from the past computational design of water isomers via B3LYP is then referred to the target domain of interest. In each domain, all water isomers are denoted as positive labeled data while the unstable water-molecule structures are assigned with negative labels.

Aside from PDM-SVM, here, the traditional methods, SVM and LapSVM, and DA methods, KMM and DASVM, as described in Table II, are also used to address the water isomer prediction problem. In addition, FA on the SVM classifier (23), which is denoted in this study as FA-SVM, is further considered. In all the methods, the Gaussian kernel is employed. Fig. 8 summarizes the accuracies obtained by PDM-SVM and all other algorithms considered for predicting the water isomers. Overall, PDM-SVM showcased superiority over all the methods considered, with rewarding performance of at least 5% accuracy improvements. Analysis shows that the improvements are attained due to class distribution differences between the source and target domains, as denoted in the last column of Table VI, and also the conflicting class labels among domains, which is caused by the low fidelity of the empirical models. In addition, PDM-SVM is shown to identify suitable source samples as depicted by the circle-enclosed substructures in Fig. 9(a) and (b) that would lead to positive transferability. The dotted-rectangle-enclosed substructures in Fig. 9(a) and the dotted-triangle-enclosed substructures in Fig. 9(b) denote the positive information transferred from source domains in the inference of the target B3LYP molecule of Fig. 9(c). This learning process allows the target structure to be inferred while avoiding source samples that conflict with the target domain.
Eventually, PDM-SVM learns the predictive distribution of the target domain to select positive transferability data samples for enhanced prediction.

VII. SUMMARY

In practice, the true predictive distributions of the source and the target domains often differ. When the predictive distributions among domains do not match well, DA algorithms that attempt to match the marginal distributions generally fail to function well due to the phenomenon of negative transfer. The causes of predictive distribution differences among related domains are mainly due to the differing class distributions and conflicting class labels among domains in specific regions of the vector space. In addition, the challenges pertaining to the differing predictive distributions among domains are known to increase in the multiclass context since more class distributions need to be considered.

To address the issues of predictive distribution differences among domains, we first present a criterion of positive transferability, which measures the similarity between two samples among domains, we first present a criterion of positive transferability. With this criterion, a PDM regularizer is proposed to enforce data that are similar according to the notion of positive transferability, which measures the similarity between two samples among domains.

To facilitate positive transferability with significant accuracy improvements attained on both binary and multiclass contexts, thus verifying the success of the proposed PDM regularizer in the identification of relevant data and transfer of useful knowledge across source domains.

REFERENCES


