A COMPARISON OF TWO STRATEGIES FOR AVOIDING NEGATIVE TRANSFER IN DOMAIN ADAPTATION BASED ON LOGISTIC REGRESSION

A. Paul [a,*], K. Vogt[b], F. Rottensteiner[a], J. Ostermann[b], C. Heipke[a]

[a] Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany
(paul, rottensteiner, heipke)@ipi.uni-hannover.de,
[b] Institut für Informationsverarbeitung, Leibniz Universität Hannover, Germany
(vogt, ostermann)@tnt.uni-hannover.de

Commission II, ICWG II/III

KEY WORDS: Transfer learning, Domain Adaptation, Negative Transfer, Remote Sensing

ABSTRACT:

In this paper we deal with the problem of measuring the similarity between training and tests datasets in the context of transfer learning (TL) for image classification. TL tries to transfer knowledge from a source domain, where labelled training samples are abundant but the data may follow a different distribution, to a target domain, where labelled training samples are scarce or even unavailable, assuming that the domains are related. Thus, the requirements w.r.t. the availability of labelled training samples in the target domain are reduced. In particular, if no labelled target data are available, it is inherently difficult to find a robust measure of relatedness between the source and target domains. This is of crucial importance for the performance of TL, because the knowledge transfer between unrelated data may lead to negative transfer, i.e. to a decrease of classification performance after transfer. We address the problem of measuring the relatedness between source and target datasets and investigate three different strategies to predict and, consequently, to avoid negative transfer in this paper. The first strategy is based on circular validation. The second strategy relies on the Maximum Mean Discrepancy (MMD) similarity metric, whereas the third one is an extension of MMD which incorporates the knowledge about the class labels in the source domain. Our method is evaluated using two different benchmark datasets. The experiments highlight the strengths and weaknesses of the investigated methods. We also show that it is possible to reduce the amount of negative transfer using these strategies for a TL method and to generate a consistent performance improvement over the whole dataset.

1. INTRODUCTION

The key point for successful classification of remote sensing imagery by supervised machine learning is the availability of a sufficient amount of labelled training samples. The underlying assumption is that the training and test data follow the same distribution. If the training samples are taken from a different image than the one to be classified, this assumption may not always hold, e.g. due to different lighting conditions or seasonal effects. This problem can be solved by providing new training samples from the image to be classified, but the collection of the training data is an expensive and time consuming task. An alternative way is to utilize methods of *Transfer Learning* (TL) (Thrun and Pratt, 1998; Pan and Yang, 2010). The goal of TL is to adapt the classifier trained on samples from a *source domain* to a test data set from a *target domain*. One common setting in TL is *Domain Adaptation* (DA). DA assumes labelled data to be only available for a source domain dataset and the source and target domains to differ only by the marginal distributions of the features and the posterior class distributions. At the same time the domains need to be related to a certain degree. If the difference of the distributions of the source and target datasets is too high, the results of an adapted classifier can be degraded compared to the performance without adaptation, i.e. the performance by just applying the classifier trained using source domain samples to the target domain without DA (Eaton et al., 2008). This case is referred to as *negative transfer*. In this paper we address the problem of the relatedness between source and target datasets and investigate different concepts to predict negative transfer without requiring any labelled data in the target domain.

*Corresponding author

The aim of this work is to investigate and compare three different strategies for predicting negative transfer for DA. The first strategy is based on circular validation and was introduced to DA based on support vector machines (SVM) in (Bruzzone and Marconcini, 2010). The second strategy is based on the *Maximum Mean Discrepancy* (MMD) (Gretton et al., 2012), and the third strategy is an extension of MMD by incorporation of knowledge about the class labels in the source domain. None of the three methods requires labelled data in the target domain. The DA framework applied in this paper is based on our previous work (Paul et al., 2016) and uses *Logistic Regression* (LR) as a base classifier to be adapted to the target domain. We use the Vaihingen and Potsdam datasets from the ISPRS 2D semantic labelling challenge (Wegner et al., 2016) consisting of multispectral digital orthophotos (DOP) and digital surface models (DSM). Our scientific contributions can be summarized as follows:

- We propose an extension of the MMD metric for measuring the distances between two distributions so that it can incorporate the knowledge of class posteriors from the source domain.
- We use the MMD, the extended MMD and the circular validation strategy of Bruzzone and Marconcini (2010) to predict negative transfer in DA based on LR (Paul et al., 2016) and compare the performance of these three methods using publicly available benchmark data.

The remainder of this paper is organized as follows. Section 2 discusses related work about negative transfer in the framework of transfer learning. In Section 3, we give a brief overview about

our DA approach based on LR (Paul et al., 2016) and the strategies used for predicting negative transfer. Section 4 presents our experimental evaluation on real urban test areas. We conclude the article with a summary and an outlook in Section 5.

2. RELATED WORK

In this section we introduce important notations and concepts and give a short overview about the related work. Our notation follows Pan and Yang (2010). A domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$ consists of a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$ with $X \in \mathcal{X}$. In TL, we consider two domains, the source domain $\mathcal{D}_S$ and the target domain $\mathcal{D}_T$. A task for a given domain is defined as $\mathcal{T} = \{\mathcal{C}, h(\cdot)\}$ consisting of a label space $\mathcal{C}$ and a predictive function $h(\cdot)$. The predictive function can be learned from the training data $\{\mathbf{x}_i, C_i\}$, where $\mathbf{x}_i \in X$ and $C_i \in \mathcal{C}$. In TL we consider a source S, from which some knowledge will be transferred to a target T by learning a predictive function $h_T(\cdot)$ under the constraint that either the domains or the tasks, or both, are different but related. There are different settings of TL according to whether labelled training data are available in the target domain or not and according to what is actually transferred. We are interested in the DA setting, where no labelled target data are available. Our DA method is based on *instance transfer*. That is, we successively replace training data from the source domain by data from the target domain, using *semi-labels* obtained from the current state of the classifier for re-training (Paul et al., 2016). For thorough reviews about the different settings of TL we refer the reader to (Pan and Yang, 2010) and (Csurka, 2017).

The performance of Transfer Learning is highly depended on how closely the source and the target domains are related to each other. An insufficient similarity can have a negative impact on the predictive function (Rosenstein et al., 2005). The negative impact results in a reduction in accuracy compared to not transferring any knowledge, referred to as *negative transfer*. Rosenstein et al. (2005) demonstrate that for a hierarchical Naive Bayes classifier the danger of negative transfer decreases with an increasing number of labelled target samples that are available for training. However, it is difficult to find a true measure of relatedness between the source and target domains or to define a robust method for predicting negative transfer if only a few or no labelled samples are available in the target domain. This could be a reason why the area of negative transfer prediction has not been widely researched, in particular in the context of DA, where no such data are available.

Eaton et al. (2008) propose a transferability measure between domains that is defined as the difference in performance of a classifier with and without transfer. They use LR as a base classifier, and transfer is achieved by using the parameters of the classifier learned using samples from domain i as a prior for learning the classifier in domain j. Assuming multiple source domains, the authors build a model transfer graph where each domain corresponds to a node and all nodes are connected with each other; the node weights correspond to the transferability measure. In order to transfer knowledge from the source task to a target task, the transfer graph is expanded to include that new task and an optimal transfer function is learned to avoid the use of irrelevant source tasks. The authors state that the underlying assumption of the symmetry between two tasks do not always hold in practice. Furthermore, the method requires labelled target data, which we assume not to be available in our framework.

As it is difficult to measure the relatedness between any particular source and target domains, several methods propose to transfer knowledge from multiple source domains to minimize the effects of negative transfer from a single unrelated source domain. This strategy increases the chance of finding at least one source that is closely related to the target. For example, Yao and Doretto (2010) use the boosting-based TL framework from (Dai et al., 2007) to deal with multiple source domains. In this work, a weak classifier is trained for each source domain and applied to a particular target domain. Finally, the algorithm selects one target-source pair based on the lowest prediction error rate on the target domain for the current iteration. A similar approach was proposed in (Eaton and desJardins, 2011). The authors use a boosting technique for instance based transfer that selectively chooses the source knowledge to transfer to the target. In the boosting process, higher weights are assigned to source tasks showing positive transferability to the target task by adjusting the weights of individual instances within each source task. Again, the concept of transferability of (Eaton et al., 2008) is used in this context. Ge et al. (2014) also try to consider multiple source domains. For each cluster of features in the target domain, they determine a weight for each source domain to modulate its impact in the transfer process. Experiments are conducted for applications in medicine and spam filtering. However, all methods for using multiple sources cited so far require some amount of labelled target data.

A multi-source selection method for DA based on instance transfer that does not require any labelled data in the target domain was suggested in (Vogt et al., 2018). The authors propose a similarity measure between the feature distributions of different domains based on the maximum mean discrepancy (Gretton et al., 2012). This distance is well-suited for finding source domains similar to a target domain, but its potential for predicting cases of negative transfer is not evaluated. Furthermore, whereas the method does not require labelled target samples, the paradigm of source selection assumes a large database of labelled source domain data to be available, which may be prohibitive for real applications.

Seah et al. (2013) propose a DA method called predictive distribution matching (PDM) to address the problems that arise from differences in the class posteriors (predictive distributions) of different domains. They propose a transferability criterion to measure the differing predictive distributions of the target domain and the related source domains. In an iterative procedure, the algorithm trains a PDM-regularized classifier that considers the transferability to exclude source samples that are irrelevant for the target domain, and then it uses this classifier to predict pseudo-labels of target samples. Using these pseudo-labels, the transferability criterion is re-evaluated, which leads to an updated set of irrelevant source samples, and the procedure is repeated until convergence. In principle, no labelled target data are required. The framework was tested for SVM and LR as base classifiers on test data sets for text classification, but no results for image classification are given.

Persello and Bruzzone (2016) present a kernel-based feature selection method for hyperspectral data based on a measure for the dataset shift that evaluates the invariance of features across different domains. The method tries to select features that are both discriminant and invariant to the dataset shift between the source and target domains. TL is achieved by just using the selected features for classification. However, this approach requires some amount of labelled data in the target domain.

Bruzzone and Marconcini (2010) applied a heuristic strategy for predicting negative transfer in DA based on SVM that is based on circular validation. They first apply DA from the source to the target domain. Then, DA is applied in the reverse direction, so that the source classifier is determined from the adapted target classifier. The authors argue that cases of negative transfer can

be identified by measuring the classification performance of the resultant classifier in the source domain, where labelled data are available. The method is shown to mitigate the effects of negative transfer to a certain degree. In this work we investigate the circular validation strategy of Bruzzone and Marconcini (2010) and two variants of the MMD similarity metric (Gretton et al., 2012) for negative transfer detection (cf. Section 3).

3. METHODOLOGY

We start this section with the definition of the DA setting. As usual in TL, we assume to have a source and a target domain. In the source domain, we have a training dataset $\mathrm{DS}_S = \{(\mathbf{x}_s, C_s)\}_{s=1}^{N_S}$ consisting of N_S samples with known feature vectors $\mathbf{x}_s$ and corresponding class labels C_s. In contrast, the target domain dataset $\mathrm{DS}_T = \{(\mathbf{x}_t)\}_{t=1}^{N_T}$ contains N_T unlabelled samples $\mathbf{x}_t$. We assume that domains to differ only by the marginal distributions of the features and the posterior class distributions, i.e. $P(X_S) \neq P(X_T)$ and $P(C_S|X_S) \neq P(C_T|X_T)$. From that point of view, DA corresponds to a problem where the source and target domain data are different, e.g. due to different lighting conditions or seasonal effects. We apply the method for DA described in (Paul et al., 2016; Vogt et al., 2018) to transfer a classifier trained using DS_S to classify the data DS_T. In this context, it is our goal to compare different methods for predicting and, consequently, avoiding negative transfer that might occur due to the domains not being sufficiently similar. In Section 3.1, we give a brief summary of the DA approach used in our experiments. After that, we present three concepts for the prediction of negative transfer: circular validation according to Bruzzone and Marconcini (2010) (Section 3.2) and measuring the similarity of two distributions based on the MMD (Gretton et al., 2012; Vogt et al., 2018) (Section 3.3) and our extension of MMD that incorporates knowledge about the class labels in the source domain (Section 3.4). All three strategies are designed for cases where labelled samples are available only for source domain $\mathcal{D}_S$.

3.1 Domain Adaptation based on Logistic Regression

Our DA approach is based on LR. LR is a discriminative probabilistic classifier that directly models the posterior probability $P(C \,|\, \mathbf{x})$ of the class labels C given the data $\mathbf{x}$. In the multiclass case we distinguish K classes, i.e. $C \in \mathcal{C} = \{C^1, \ldots, C^K\}$. A feature transformation $\Phi(\mathbf{x})$ is applied to achieve non-linear decision boundaries in the original feature space. That is, LR is applied to a higher dimensional vector $\Phi(\mathbf{x})$ whose components are functions of $\mathbf{x}$. The first element of $\Phi(\mathbf{x})$ is a bias value assumed to be equal to 1 without loss of generality. In the multiclass case, the model of the posterior is based on the softmax function (Bishop, 2006):

$$p\left(C = C^k|\mathbf{x}\right) = \frac{exp\left(\mathbf{w}_k^T \cdot \Phi(\mathbf{x})\right)}{\sum\limits_j exp\left(\mathbf{w}_j^T \cdot \Phi(\mathbf{x})\right)}, \qquad (1)$$

where $\mathbf{w}_k$ is a vector of weight coefficients for a particular class C^k. As these weight vectors are not independent, we set $\mathbf{w}_1 = \mathbf{0}$. The remaining weight parameters $\mathbf{w}_k$ for $k \in \{2 \ldots K\}$ are collected in a weight vector $\mathbf{w} = (\mathbf{w}_2^T, \ldots, \mathbf{w}_K^T)^T$ determined using a training dataset $\overline{\mathrm{TD}}$.

We start the instance based iterative DA process by using the initial training set $\overline{\mathrm{TD}}^0 = \mathrm{DS}_S$ to train our initial classifier. In each further iteration i of DA, a predefined number of source samples is removed from and a number of target domain samples is included into the current training data set. Thus, in iteration i, the current training data set $\overline{\mathrm{TD}}^i$ consists of a mixture of N_S^i remaining source samples and N_T^i included target samples: $\overline{\mathrm{TD}}^i = \{\, \{(\mathbf{x}_s, C_s)\}_{s=1}^{N_S^i} \cup \{(\mathbf{x}_t, \widetilde{C}_t)\}_{t=1}^{N_T^i}\}$, where $\widetilde{C}_t$ denotes the *semi-labels* of the target samples, which are determined automatically using *knn* (k-nearest neighbourhood) analysis. As i is increased, N_S^i becomes smaller and N_T^i increases, until finally, only target samples with semi-labels are used for training, thus $\overline{\mathrm{TD}}^{i_{\mathrm{end}}} = \{(\mathbf{x}_t, \widetilde{C}_t)\}_{t=1}^{N_T'}$ with $N_T' \leq N_T$. For more detailed information about our DA procedure we refer the reader to our previous work (Paul et al., 2016).

3.2 Circular Validation Strategy

The main idea of circular validation relies on the assumption that there exists an intrinsic relationship between solutions that are satisfactory for the two domains. If knowledge transfer performs well, the DA algorithm is expected to be able to move from modeling the source domain problem to modeling the target domain problem and vice versa. On the other hand, if knowledge transfer is impossible, this will be very likely to be true for transfer in both directions, and DA will not yield acceptable results (Bruzzone and Marconcini, 2010). Based on these considerations, Bruzzone and Marconcini (2010) propose to apply DA first and to use the adapted classifier to classify the data in the target domain. After that, DA is performed again, but in the reverse direction (cf. Figure 1). That is, the reverse DA starts by using the results of the classification in the target domain for initial training and then adapts the classifier to the source domain without using the known training labels from that domain. If the source data are classified well after the reverse DA, the distributions of source and target domain are assumed to be related to each other sufficiently well for DA.

We use the overall accuracy $\mathrm{OA}_{\mathrm{DAR}}$ determined from the results of the adapted classifier after reverse DA in the source domain as a feature to predict cases of negative transfer. Given a threshold τ_{OA}, the classification accuracy is considered to be acceptable if $\mathrm{OA}_{\mathrm{DAR}} \geq \tau_{\mathrm{OA}}$. This means that the reverse classifier moves to the state A_1 (cf. Figure 1). In contrast, if $\mathrm{OA}_{\mathrm{DAR}} < \tau_{\mathrm{OA}}$, the solution is not acceptable and the reverse classifier corresponds to state A_4. Note that, if $\mathrm{OA}_{\mathrm{DAR}}$ is used to predict negative transfer, it is unclear whether the problem occurred in the original or in the reverse DA process. That is, it may lead to the rejection of cases in which DA from the source to the target was actually successful (arrow from state A_2 to A_4 in Figure 1). Bruzzone and Marconcini (2010) claim that the proposed validation technique is effective if starting from state A_3 the system never moves back

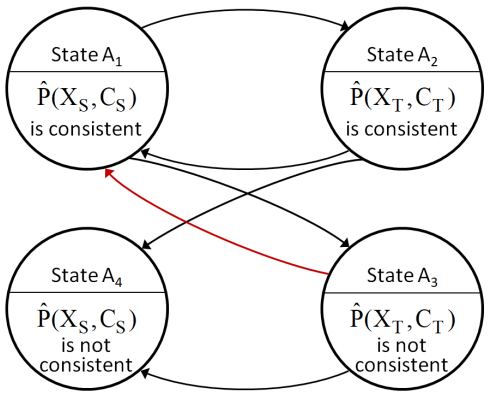


Figure 1: Diagram of circular validation with all possible transitions (adapted from Bruzzone and Marconcini (2010)). The transition from A_3 to A_1 (red arrow) is prohibited according to Bruzzone and Marconcini (2010), thus $P(A_1|A_3) \stackrel{!}{=} 0$.

to state A_1, i.e. $P(A_1|A_3) \stackrel{!}{=} 0$ (red arrow in Figure 1), and if starting from state A_2 the system can return to state A_1, i.e. $P(A_1|A_2) > 0$.

3.3 Maximum Mean Discrepancy

The MMD of Gretton et al. (2012) delivers a generic measure of similarity of the marginal distributions of the source and target domains without applying any knowledge about class labels. It is computed as the distance between the means of the probability distributions in a Reproducing Hilbert Kernel Space (RKHS).

In practice, the MMD is computed empirically using finite number of data samples taken from the investigated distributions. Let us assume $\hat{X}_S$ to be a subset of source data samples $\hat{X}_S \subseteq X_S$, $\hat{X}_S = \{\mathbf{x}_{S,i}\}_{i=1}^{\hat{N}_S}$ and $\hat{X}_T$ to be a subset of target data samples $\hat{X}_T \subseteq X_T$, $\hat{X}_T = \{\mathbf{x}_{T,i}\}_{i=1}^{\hat{N}_T}$, where $\hat{N}_S$ and $\hat{N}_T$ are the number of samples in each subset. Then, for $\hat{X}_S$ and $\hat{X}_T$, drawn independently and identically from X_S and X_T, respectively, the empirical biased MMD can be computed as follows (Gretton et al., 2012):

$$d^2_{MMD}(\hat{X}_S, \hat{X}_T) = \frac{1}{\hat{N}_S^2}\sum_{i=1}^{\hat{N}_S}\sum_{j=1}^{\hat{N}_S} k(\mathbf{x}_{S,i}, \mathbf{x}_{S,j}) - \frac{2}{\hat{N}_S\hat{N}_T}\sum_{i=1}^{\hat{N}_S}\sum_{j=1}^{\hat{N}_T} k(\mathbf{x}_{S,i}, \mathbf{x}_{T,j}) + \frac{1}{\hat{N}_T^2}\sum_{i=1}^{\hat{N}_T}\sum_{j=1}^{\hat{N}_T} k(\mathbf{x}_{T,i}, \mathbf{x}_{T,j}), \tag{2}$$

where $k(\cdot)$ is a kernel used for mapping from the original feature space to the RKHS to enhance the accuracy of linear discriminants in this alternate feature space. In our considerations we use the Gaussian kernel:

$$k_{\text{RBF}}(\mathbf{x}_l, \mathbf{x}_r) = \exp\left(-\frac{\|\mathbf{x}_l - \mathbf{x}_r\|^2}{2\sigma^2}\right). \tag{3}$$

The bandwidth parameter σ is determined as the average distance of samples to their nearest neighbours in feature space.

The calculated MMD value is used as a measure for the distance between the source and target distributions. A threshold τ_{MMD} can be applied to this measure to predict potential cases of negative transfer.

3.4 Modified MMD

We extend the MMD distance measure, proposing a strategy to include the knowledge about the class labels in the source domain to improve the discriminative power of that distance. For that purpose, we use the probabilistic output of a classifier trained on source domain data to change both source and target domain distributions. In particular, we shift samples of both domains in the feature space relative to the decision boundaries of the classifier using the gradient of the probability function obtained for the source domain:

$$\frac{\partial p}{\partial \mathbf{x}} = p_k \cdot \left(\mathbf{w}_k^T - \sum_j \mathbf{w}_j^T \cdot p_j\right)^T \cdot \frac{\partial \Phi(\mathbf{x})}{\partial \mathbf{x}} \tag{4}$$

where p_k and p_j are posteriors according to Eq. (1) for classes k and j, respectively, and k is the class of maximum probability, i.e., the class to which a sample $\mathbf{x}$ belongs according to the decision boundaries. We then calculate new values for $\mathbf{x}_s$ and $\mathbf{x}_t$ according to Eq. (5).

$$\mathbf{x}_{new} = \mathbf{x} + \mathbf{x} \cdot \frac{\partial p}{\partial \mathbf{x}} \cdot \lambda \tag{5}$$

Eq. (5) is applied to all samples independently from the domain. Consequently, we omit the domain index, i.e. a sample $\mathbf{x}$ may be from the source domain ($\mathbf{x}_s$) or the target domain ($\mathbf{x}_t$). The parameter λ modulates the distance by which a sample is shifted, and its sign decides whether the samples are shifted away from ($\lambda > 0$) or closer to the decision boundary ($\lambda < 0$). Greater absolute values of λ lead to larger shifts. Having shifted the samples from both domains according to the classifier in the source domain, the modified MMD (MMD_m) is calculated according to Eq. (2) using the new values $\mathbf{x}_{new}$ for both domains.

The idea behind this adaptation is following: if the original distributions of the data in the two domains are similar, $\mathbf{x}_s$ and $\mathbf{x}_t$ will be changed in a similar way, and the computed distances will hardly change. On the other hand, if the distributions and, thus, the optimal decision boundaries in the two domains are very different, shifting target samples away from the decision boundaries in the source domain will increase the distance. We therefore expect this strategy to amplify the distances between dissimilar distributions. Again, the value of MMD_m value (MMD after modifying the distributions) is used as a distance measure between the source and target distributions, to which a threshold τ_{MMD_m} can be applied for predicting negative transfer.

4. EXPERIMENTS

The experiments are carried out to evaluate the limits and the effectiveness of the presented strategies for negative transfer detection in DA based on LR classifier. We determine the overall accuracy OA_{DAR} in the source domain after the two (forward and reverse) DA processes for the circular validation strategy and the values of MMD and MMD_m values for the two strategies based on domain distances. These values are used as measures for the relatedness of the domains; we expect that negative transfer can be avoided by using the DA results in the target domain only if the values are in an acceptable range, i.e. if OA_{DAR} is larger than a given threshold for the circular validation strategy (cf. Section 3.2) or if MMD or MMD_m are smaller than given thresholds in the other cases (cf. Sections 3.3 and 3.4).

4.1 Test data and experimental setup

Our experimental evaluation is based on the Vaihingen and Potsdam datasets from the ISPRS 2D semantic labelling challenge (Wegner et al., 2016), consisting of multispectral DOP and DSM. The test data show suburban scenes with a total of six object classes, namely *impervious surface*, *building*, *low vegetation*, *tree*, *car* and *clutter/background*. For our experiments we assume *impervious surface* to also include *car* and *clutter/background*, and consequently, we only consider the remaining four classes. As the original datasets have different resolutions and different spectral configurations, we resampled all images to a ground sampling distance of 8 cm and used colour infrared (CIR) composites of all images. Only patches for which a reference is publicly available were used in our experiments. The properties of the datasets are summarized in Table 1.

Dataset	GSD	Channels	Patches	F.	Classes
Vaihingen	8 cm	RGIR	15	5	4
Potsdam	8 cm	RGBIR	23	5	4

Table 1: Dataset properties. F.: Number of features.

All experiments are based on a pixel-wise classification. We use the same feature space for all datasets. Under this constraint, we select the five most discriminative features using *Random Forest* based feature selection (Breiman, 2001) from a pool of spectral, structural and texture features. We settle on the *normalized difference vegetation index* (NDVI), *normalized digital surface model* (nDSM) and the pixelwise red, green & near infrared spectral components.

We test all pairs of patches in each dataset, using one patch as the source domain and the other one as the target domain. We did not mix patches from the Vaihingen and Potsdam datasets due to time restrictions. As the datasets were acquired at different vegetation periods, we expect them to be too different for DA to work, but this still needs to be verified by experiments. Whereas reference labels are available for all patches and, thus, for both the source and the target domains in all tests, the reference labels in the target domain $\mathcal{D}_T$ are only considered for the performance evaluation and not for DA or the prediction of target and source domains. For each pair of source and target domains, we compute all distance metrics: $\mathrm{OA_{DAR}}$, MMD and MMD_m and use them to predict negative transfer. To improve the robustness of the computation of MMD and MMD_m, we determined it 10 times for each pair of source and target image patch, using different bootstrap datasets, each consisting of $\hat{N}_S$ and $\hat{N}_T$ randomly chosen samples from the source and target domains respectively with $\hat{N}_S = \hat{N}_T = 10000$. The values used for the prediction of negative transfer are the averages of the 10 independent bootstrap runs. The value of λ for shifting the samples in both domains for computing MMD_m as described in Section 3.4 is set to $\lambda = 1$.

In order to evaluate the properties of our DA method, we also generate three classification results for the target domain and compare them to the reference to determine three different values for the overall accuracy (OA). The first value, $\mathrm{OA_{ST}}$, is obtained by applying the base classifier (LR) trained on the source domain data directly to the target data without DA. The second one, $\mathrm{OA_{TT}}$ is the OA of a classifier trained and evaluated on the target set. We consider it to represent the optimal accuracy that can be achieved. Finally, $\mathrm{OA_{DA}}$ is the OA that is achieved when applying DA from the source to the target domain using the method described in Section 3.1. Thus, the difference $\Delta\mathrm{OA_{DA}} = \mathrm{OA_{DA}} - \mathrm{OA_{ST}}$ is the change in OA due to DA; negative transfer is characterised by $\Delta\mathrm{OA_{DA}} < 0$. The differences $\Delta\mathrm{OA_{ST}} = \mathrm{OA_{TT}} - \mathrm{OA_{ST}}$ and $\Delta\mathrm{OA_{DT}} = \mathrm{OA_{TT}} - \mathrm{OA_{DA}}$ measure the loss in OA due to the non-availability of training data in the target domain, if the source classifier is applied to the target domain directly ($\Delta\mathrm{OA_{ST}}$) or after DA ($\Delta\mathrm{OA_{DT}}$). The measures reported in this paper are computed on a per-pixel basis for each pair of source and target domains independently from each other; we also compute average values.

In order to evaluate the suitability of the three similarity measures $M \in \{\mathrm{OA_{DAR}}, \mathrm{MMD}, \mathrm{MMD}_m\}$ for predicting negative transfer, we start with a regression analysis to assess the degree to which the performance of DA as measured by $\Delta\mathrm{OA_{DT}}$ and the measure M are correlated; we expect a measure having a higher correlation to be a better predictor for $\Delta\mathrm{OA_{DT}}$ and, thus, for negative transfer. After that, we test the capabilities of the metrics for predicting negative transfer by applying thresholds to these metrics. After predicting the occurrence of negative transfer using a given threshold, we compare the predictions to the reference (given by the sign of $\Delta\mathrm{OA_{DT}}$). We present *Receiver Operating Characteristic* (ROC) curves for all metrics, obtained by varying the corresponding thresholds. The ROC curves show the true positive rates (TPR) of the prediction (*sensitivity*) as a function of the false positive rates (FPR = *1-specificity*) for various threshold values. They help to evaluate the accuracy of the predictions of negative transfer.

Finally, we asses the impact of the prediction of negative transfer on the DA performance. Again, this is done for each metric $M \in \{\mathrm{OA_{DAR}}, \mathrm{MMD}, \mathrm{MMD}_m\}$. The corresponding thresholds τ_M are obtained as the optimal ones according to the ROC curves. We do this to show the potential of that prediction under optimal conditions (because the ROC curves are based on labels in the target domain); the sensitivity of the results to the threshold selection is part of our future work. The optimal threshold value τ_M for a metric M is determined from the ROC point having the shortest distance to the best possible case, i.e. the point where TPR = 1 and FPR = 0. Comparing M to τ_M, we predict negative transfer. If the test indicates that negative transfer is to be expected, DA is rejected and the classifier trained using source domain data is applied to the target domain; otherwise, our DA method is applied to obtain and final classifier in the target domain. Consequently, after the comparison of the results with the reference, we can obtain new values $\Delta\overline{\mathrm{OA}}_{\mathrm{DA}}$ showing the improvement of OA due to the prediction of negative transfer. For circular validation strategy, $\Delta\overline{\mathrm{OA}}_{\mathrm{DA}}$ becomes

$$\Delta\overline{\mathrm{OA}}_{\mathrm{DA}} = \begin{cases} \mathrm{OA_{DA}} - \mathrm{OA_{ST}} & \text{if } \mathrm{OA_{DAR}} > \tau_{\mathrm{OA_{DAR}}} \\ 0 & \text{otherwise} \end{cases}, \quad (6)$$

whereas for $M \in \{\mathrm{MMD}, \mathrm{MMD}_m\}$. we have

$$\Delta\overline{\mathrm{OA}}_{\mathrm{DA}} = \begin{cases} \mathrm{OA_{DA}} - \mathrm{OA_{ST}} & \text{if } M < \tau_M \\ 0 & \text{otherwise} \end{cases}. \quad (7)$$

Of course, if negative transfer is predicted, the improvement is 0. The reason for different direction of the inequality signs in Eq. (6) and (7) is that we expect lower $\mathrm{OA_{DAR}}$ values indicating the negative transfer in case of the circular validation strategy and greater values for one of the MMD metrics indicating the increasing dissimilarity between domains.

4.2 Evaluation

4.2.1 Domain adaptation: Figure 2 shows results of DA without applying any strategy for negative transfer prediction. It shows percentile plots of the cumulative distribution over $\Delta\mathrm{OA_{DA}}$ for all patches in the two datasets. We achieve positive transfer using our DA approach in 37% of all cases for Vaihingen and in 28% for Potsdam. These values are the complements to 100% of the intersection points of the curves with the vertical axis. In a similar way, we find that for 89% and 76% of all pairs of Vaihingen and Potsdam patches, respectively, the negative transfer

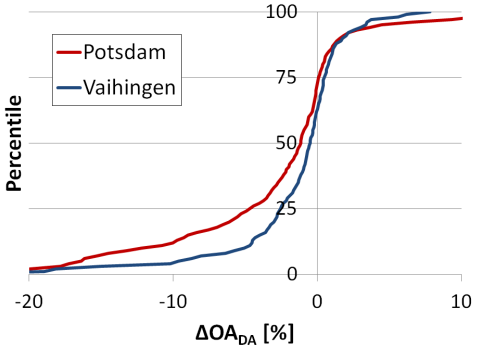
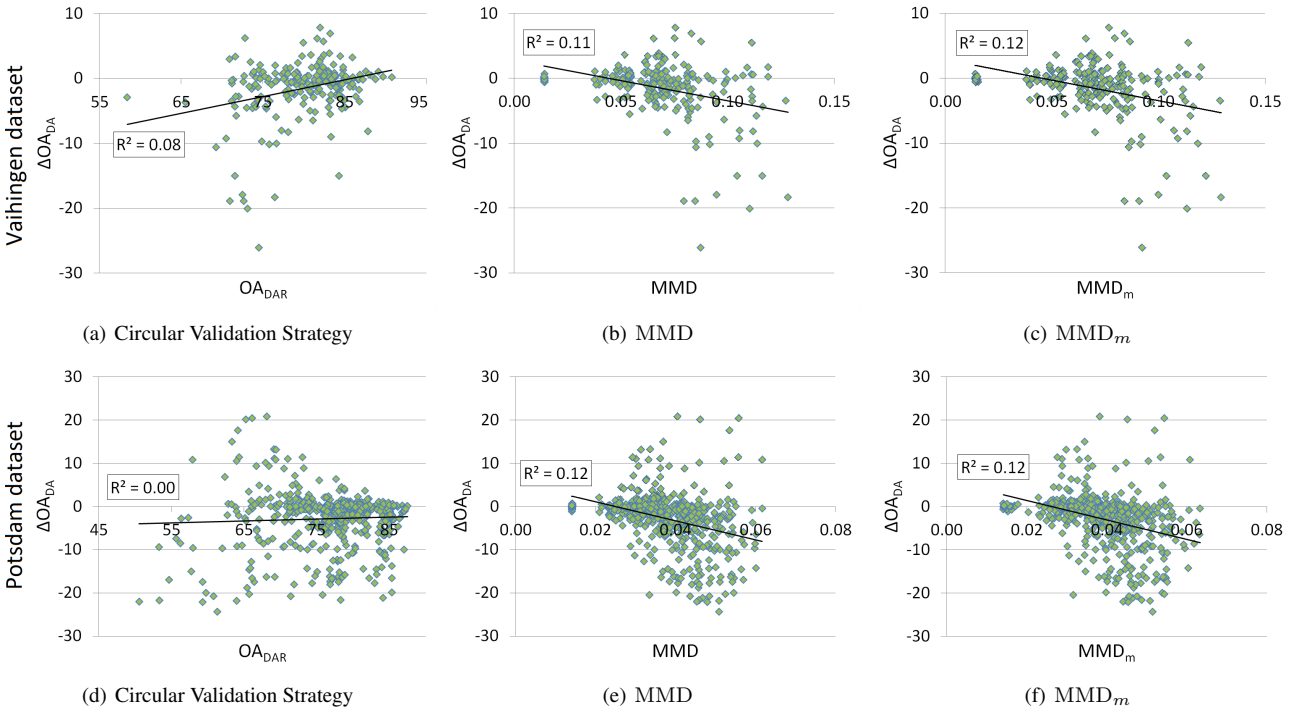


Figure 2: Percentile plots of the DA performance, measured by $\Delta\mathrm{OA_{DA}}$, for the Vaihingen and Potsdam datasets.



(a) Circular Validation Strategy (b) MMD (c) MMD_m

(d) Circular Validation Strategy (e) MMD (f) MMD_m

Figure 3: Results of regression analysis for the Vaihingen dataset (a)-(c) and Potsdam the dataset (d)-(f).

(loss of classification accuracy in the target domain due to DA) is below 5%. Average values for the loss due to the non-availability of training data in the target domain and the impact of DA are shown in Table 2. The average loss $\Delta\mathrm{OA}_{\mathrm{ST}}$ is $4.45\,\%$ for Vaihingen and $6.28\,\%$ for Potsdam dataset. After DA, the average loss ($\Delta\mathrm{OA}_{\mathrm{DT}}$) is slightly larger, namely $5.98\,\%$ for Vaihingen and $9.11\,\%$ for Potsdam. Consequently the average change $\Delta\mathrm{OA}_{\mathrm{DA}}$ in OA due to DA is negative, namely $-1.53\,\%$ and $-2.83\,\%$ for the Vaihingen and Potsdam datasets, respectively. This indicates that on average, there is a negative transfer which, as indicated by Figure 2, is relatively small in most cases but may become larger than $20\,\%$ in OA. This highlights the importance of investigating strategies for avoiding negative transfer.

Dataset	$\Delta\mathrm{OA}_{\mathrm{ST}}$	$\Delta\mathrm{OA}_{\mathrm{DT}}$	$\Delta\mathrm{OA}_{\mathrm{DA}}$
Vaihingen	$4.5\,\%$	$6.0\,\%$	$-1.5\,\%$
Potsdam	$6.3\,\%$	$9.1\,\%$	$-2.8\,\%$

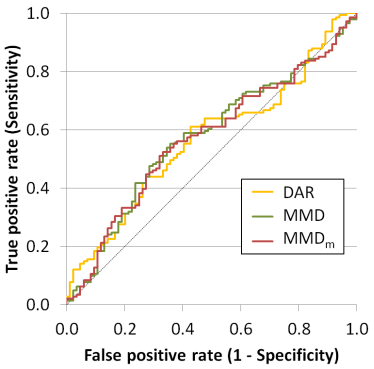
Table 2: Average values of $\Delta\mathrm{OA}_{\mathrm{DA}}$, $\Delta\mathrm{OA}_{\mathrm{ST}}$, $\Delta\mathrm{OA}_{\mathrm{DT}}$ as defined in Section 4.1 over the Vaihingen and Potsdam datasets.

4.2.2 Regression analysis: Here, we evaluate and compare the performance of the $\mathrm{OA}_{\mathrm{DAR}}$, MMD and the modified MMD_m distance metrics by regression analysis. We expect a measure having a higher correlation to be a better predictor for $\Delta\mathrm{OA}_{\mathrm{DA}}$ and, thus, for negative transfer. The results for our datasets are presented in Figure 3. Our experiments show no significant correlation between the performance of DA as measured by $\Delta\mathrm{OA}_{\mathrm{DA}}$ and either of the distance metrics. The highest correlation with $\Delta\mathrm{OA}_{\mathrm{DA}}$ is achieved using MMD and the modified MMD_m, but only with a coefficient $R^2 = 0.12$, i.e. only 12% of the dispersion can be explained by the regression model. The corresponding empirical correlation coefficient, which is the root of R^2, is about 35%. However, based on inclination angles of the regression lines (cf. Figure 3), the probability for positive transfer increases, for larger values of $\mathrm{OA}_{\mathrm{DAR}}$ and smaller values of MMD and the modified MMD_m, which is as expected. The reason for the worse result using the circular validation strategy ($R^2 = 0.08$ for Vaihingen and 0 for Potsdam) can be attributed to the violation of the transition assumption from the state A_3 to state A_1 (cf. Figure 1). Indeed, it was possible using our DA approach move back to the state A_1 even if the knowledge transfer was unsuccessful, i.e., we observed $P(A_1|A_3) > 0$.

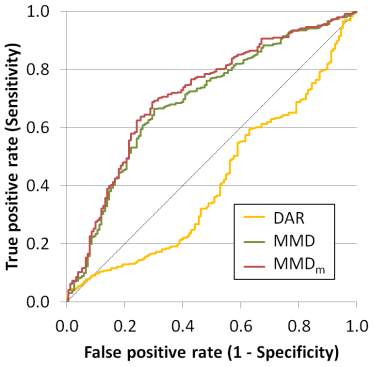
4.2.3 Negative transfer prediction: The ROC curves generated in the way described in Section 4.1 are shown in Figure 4. The figure shows that the two MMD-based metrics perform similarly, whereas the circular validation performs slightly worse in Vaihingen and considerably worse in Potsdam. In Vaihingen, the best results are in the order of 60% in TPR at about 45% FPR, which shows that it is difficult to predict negative transfer for that dataset. In Potsdam, the results are somewhat better for the MMD-based metrics, with a sligth advantage of MMD_m over MMD. Here, the best results are in the order of 70% in TPR at about 30% FPR.

We selected optimal thresholds for the prediction of negative transfer from the ROC curves and considered that prediction in the DA process in the way described in Section 4.1. Figure 5 shows percentile plots of the cumulative distribution of the effect of DA with negative transfer prediction on the OA of the classification ($\Delta\overline{\mathrm{OA}}_{\mathrm{DA}}$) for the three metrics and also compares them to the results achieved without negative transfer prediction ($\Delta\mathrm{OA}_{\mathrm{DA}}$). Using negative transfer prediction, we obtain better DA results. The percentage of cases without negative transfer ($\Delta\mathrm{OA}_{\mathrm{DA}} \geq 0$) is about 75% of all cases for Vaihingen and Potsdam dataset, compared to 37% and 28% of the original DA results, respectively (cf. Section 4.2.1). For 25% of all cases we achieve an improvement of $\Delta\overline{\mathrm{OA}}_{\mathrm{DA}}$ of about 2.3% or larger in OA for Vaihingen and even of about 4.0% or larger in OA for Potsdam dataset. Table 3 presents the average values over the whole datasets along the optimal thresholds used for the prediction of negative transfer. The average values presented in Table 3 show an improvement in OA of 1.4% and 2.8% for Vaihingen and Potsdam, respectively, due to the prediction of negative transfer. Comparing the three metrics, Table 3 indicates a slightly better performance of MMD_m compared to MMD and a considerably better one of

both MMD-based methods compared to circular validation on the Potsdam dataset. This is also confirmed by Figure 5b, showing a relatively large percentage of cases with negative transfer for Potsdam.



(a)



(b)

Figure 4: ROC curves for the prediction of negative transfer for Vaihingen (a) and Potsdam (b). DAR: Circular validation.

Vaihingen

	τ_M	$\Delta\mathrm{OA}_{\mathrm{DA}}$	$\Delta\overline{\mathrm{OA}}_{\mathrm{DA}}$	$\Delta\overline{\mathrm{OA}}_{\mathrm{DA}} - \Delta\mathrm{OA}_{\mathrm{DA}}$
DAR	81.50	-1.5%	-0.2%	1.3%
MMD	0.073	-1.5%	-0.1%	1.4%
MMD$_m$	0.070	-1.5%	-0.1%	1.4%

Potsdam

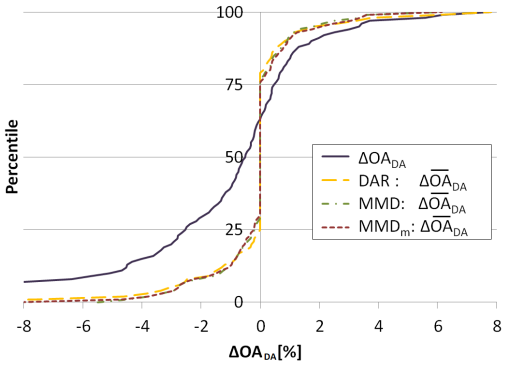
	τ_M	$\Delta\mathrm{OA}_{\mathrm{DA}}$	$\Delta\overline{\mathrm{OA}}_{\mathrm{DA}}$	$\Delta\overline{\mathrm{OA}}_{\mathrm{DA}} - \Delta\mathrm{OA}_{\mathrm{DA}}$
DAR	78.40	-2.8%	-1.2%	1.6%
MMD	0.037	-2.8%	-0.1%	2.7%
MMD$_m$	0.036	-2.8%	-0.0%	2.8%

Table 3: Domain adaptation results over the Vaihingen and Potsdam dataset after applying threshold τ_M for negative transfer prediction. $\Delta\mathrm{OA}_{\mathrm{DA}}$: average loss in OA before applying the strategy. $\Delta\overline{\mathrm{OA}}_{\mathrm{DA}}$: loss in OA after negative transfer prediction. $\Delta\overline{\mathrm{OA}}_{\mathrm{DA}} - \Delta\mathrm{OA}_{\mathrm{DA}}$: improvement in OA.

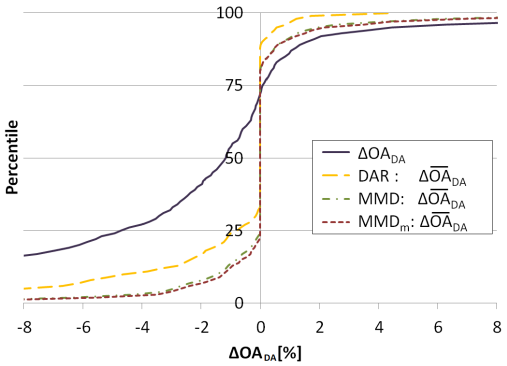
Finally, we compare the effects of negative transfer prediction strategy on the OA after DA compared to a classifier trained on target samples ($\mathrm{OA}_{\mathrm{TT}}$; cf. Section 4.1). We define the difference $\Delta\overline{\mathrm{OA}}_{\mathrm{DT}}$ to the target classifier according to Eq. (8) for circular validation and according to Eq. (9) for MMD or MMD$_m$:

$$\Delta\overline{\mathrm{OA}}_{\mathrm{DT}} = \begin{cases} \mathrm{OA}_{\mathrm{TT}} - \mathrm{OA}_{\mathrm{DA}} & \text{if } M > \tau_M \\ \mathrm{OA}_{\mathrm{TT}} - \mathrm{OA}_{\mathrm{ST}} & \text{otherwise} \end{cases} \quad (8)$$

$$\Delta\overline{\mathrm{OA}}_{\mathrm{DT}} = \begin{cases} \mathrm{OA}_{\mathrm{TT}} - \mathrm{OA}_{\mathrm{DA}} & \text{if } M < \tau_M \\ \mathrm{OA}_{\mathrm{TT}} - \mathrm{OA}_{\mathrm{ST}} & \text{otherwise} \end{cases} . \quad (9)$$



(a)



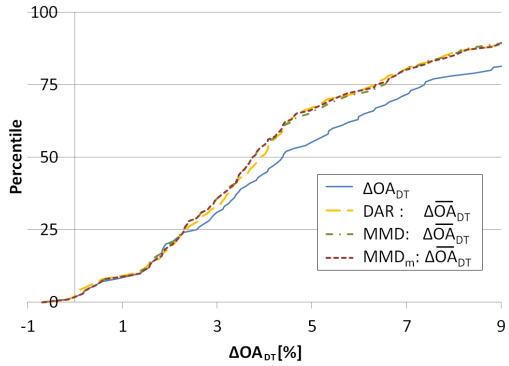
(b)

Figure 5: Percentile plots of $\Delta\overline{\mathrm{OA}}_{\mathrm{DA}}$ after the negative transfer prediction using three metrics for Vaihingen (a) and Potsdam (b) and of the results without that prediction ($\Delta\mathrm{OA}_{\mathrm{DA}}$).

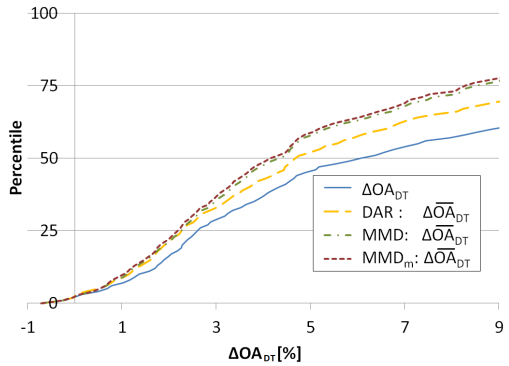
Figure 6 shows percentile plots of the distribution of $\Delta\overline{\mathrm{OA}}_{\mathrm{DT}}$ for the two datasets and compares them to those of DA without prediction of negative transfer. For 50% of all cases we achieve a decrease of $\Delta\overline{\mathrm{OA}}_{\mathrm{DT}}$ of about 0.6% or larger for Vaihingen and of about 1.9% or larger in OA for Potsdam; the results, thus, are nearer to the optimal case by that margin due to negative transfer prediction. Again, the poorer performance of circular validation on the Potsdam data is confirmed.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented results of different strategies for negative transfer prediction for domain adaptation based on logistic regression. We made sure that none of the strategies requires labels in the target domain. We proposed a modified MMD metric MMD$_m$ to incorporate the knowledge about the class posterior given the data. Our results confirm the suitability of the proposed extension compared to the original MMD metric and to a metric derived from circular validation. While it possible to reduce the amount of negative transfer and the mean loss in $\Delta\mathrm{OA}_{\mathrm{DA}}$ over the whole dataset applying this metrics, we did not observe a direct relationship and thus a dependance between one of the metrics and positive or negative transfer. The performance of negative transfer prediction was better for MMD and for MMD$_m$ than for circular validation, though in general it was far from perfect. The problems of MMD and MMD$_m$ in predicting negative transfer are attributed to different posterior class distributions of the source and target domain. The reason for the poorer performance of circular validation compared to (Bruzzone and Marconcini, 2010) may be related to the use of a weaker classifier which might lead to a violation of some fundamental assumptions of the method.



(a)



(b)

Figure 6: Percentile plots of $\Delta\overline{\mathrm{OA}}_{\mathrm{DT}}$ after the negative transfer prediction using three metrics for Vaihingen (a) and Potsdam (b) and of the results without that prediction ($\Delta\mathrm{OA}_{\mathrm{DA}}$) (the higher the better).

In our future work we will evaluate the sensitivity of negative transfer prediction to the selection of threshold and carry out experiments where we mix images from different datasets, which should further highlight the limitations of the methods. Another step is the comparison of our approach to a cluster-based method or to an inductive setting where a small amount of labelled data is available. We expect this will improve the classification accuracy and the predictive ability of negative transfer detection strategies.

ACKNOWLEDGEMENTS

This work was supported by the German Science Foundation (DFG) under grant HE 1822/30-1. The Vaihingen and Potsdam data were provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) (Cramer, 2010): http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html.

References

Bishop, C. M., 2006. Pattern Recognition and Machine Learning. 1^{st} edn, Springer, New York (NY), USA.

Breiman, L., 2001. Random forests. Machine Learning 45(1), pp. 5–32.

Bruzzone, L. and Marconcini, M., 2010. Domain adaptation problems: A dasvm classification technique and a circular validation strategy. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(5), pp. 770–787.

Cramer, M., 2010. The DGPF test on digital aerial camera evaluation – overview and test design. Photogrammetrie Fernerkundung Geoinformation 2(2010), pp. 73–82.

Csurka, G., 2017. A comprehensive survey on domain adaptation for visual applications. In: G. Csurka (ed.), Domain Adaptation in Computer Vision Applications, Springer International Publishing, Cham, pp. 1–35.

Dai, W., Xue, G.-R., Yang, Q. and Yu, Y., 2007. Transferring naive bayes classifiers for text classification. In: Proceedings of the 22^{nd} AAAI Conference on Artificial Intelligence, pp. 540–545.

Eaton, E. and desJardins, M., 2011. Selective transfer between learning tasks using task-based boosting. In: 25^{th} AAAI Conference on Artificial Intelligence, pp. 337–342.

Eaton, E., desJardins, M. and Lane, T., 2008. Modeling transfer relationships between learning tasks for improved inductive transfer. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases Part I, pp. 317–332.

Ge, L., Gao, J., Ngo, H., Li, K. and Zhang, A., 2014. On handling negative transfer and imbalanced distributions in multiple source transfer learning. Statistical Analysis and Data Mining 7(4), pp. 254–271.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A., 2012. A kernel two-sample test. Journal of Machine Learning Research 13(2012), pp. 723–773.

Pan, S. J. and Yang, Q., 2010. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22(10), pp. 1345–1359.

Paul, A., Rottensteiner, F. and Heipke, C., 2016. Iterative reweighted instance transfer for domain adaptation. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences III-3, pp. 339–346.

Persello, C. and Bruzzone, L., 2016. Kernel-based domain-invariant feature selection in hyperspectral images for transfer learning. IEEE Transactions on Geoscience and Remote Sensing 54(5), pp. 2615–2626.

Rosenstein, M. T., Marx, Z., Kaelbling, L. P. and Dietterich, T. G., 2005. To transfer or not to transfer. In: NIPS05 Workshop, Inductive Transfer: 10 Years Later.

Seah, C. W., Ong, Y. S. and Tsang, I. W., 2013. Combating negative transfer from predictive distribution differences. IEEE Transactions on Cybernetics 43(4), pp. 1153–1165.

Thrun, S. and Pratt, L., 1998. Learning to learn: Introduction and overview. In: S. Thrun and L. Pratt (eds), Learning to Learn, Kluwer Academic Publishers, Boston, MA (USA), pp. 3–17.

Vogt, K., Paul, A., Ostermann, J., Rottensteiner, F. and Heipke, C., 2018. Unsupervised source selection for domain adaptation. Photogrammetric Engineering & Remote Sensing. In print.

Wegner, J. D., Rottensteiner, F., Gerke, M. and Sohn, G., 2016. The ISPRS 2D Labelling Challenge. http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html. Accessed 05/04/2018.

Yao, Y. and Doretto, G., 2010. Boosting for transfer learning with multiple sources. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1855–1862.