Which data do we need for training? Domain adaption and learning under label noise

Franz Rottensteiner, Hannover

1. INTRODUCTION

The classification of images and other remote sensing data is a fundamental task to derive semantic information about the objects in the depicted scene automatically. Current research focuses on statistical approaches, in which the knowledge about the objects is given implicitly in the form of training samples that are used to train a classifier. These approaches can be easily adapted to new domains by defining a new representative set of training data. However, this flexibility comes at a cost: the need to generate training data. This article presents two strategies for avoiding the manual generation of new training data.

2. DOMAIN ADAPTATION

Transfer Learning (TL) tries to answer the question whether a classifier trained in the past using a dataset for which a sufficient amount of training data was available can be of any help in the classification of new data even if they have slightly different properties. Pan and Yang (2010) define a *domain* $D = \{\mathscr{F}; P(X)\}$ to consist of a feature space $\mathscr{F}$ and a marginal probability distribution $P(X)$ with $X \in \mathscr{F}$. In TL, we consider different domains, the *source domains* D_S and the *target domain* D_T. Given a domain D, a *task* $T = \{\mathbb{C}; h(\cdot)\}$ consists of a label space $\mathbb{C}$ and a predictive function $h(\cdot)$. This function can be learned from the training data $\{\mathbf{x}_i, C_i\}$, where $\mathbf{x}_i \in X$ and $C_i \in \mathbb{C}$. TL is defined as a procedure that helps to learn the predictive function $h_T(\cdot)$ in D_T using the knowledge in D_T and D_S, where either the domains or the tasks, or both, are *different but related* (Pan & Yang, 2010).
Domain adaptation (DA) is a special sub-category of TL in which different domains are supposed only to differ by the marginal distributions of the features and the posterior class distributions (Bruzzone & Marconcini, 2009). In our application, a domain corresponds to a set of remote sensing data, and the source and target domains (datasets) are different, e.g. due to different lighting conditions. DA allows transferring a classifier trained on a set of remote sensing data where training data are available (D_S) to other scenes (D_T) without having to provide additional training data in D_T. There are different strategies for DA (Pan & Yang, 2010). Methods based on *feature representation transfer* try to find feature representations that allow a simple transfer from the source to the target domain. Methods based on *instance transfer* try to re-use training samples from D_S directly, replacing them by samples from D_T receiving their class labels (*semi-labels*) based on the current state of the classifier, e.g. (Bruzzone & Marconcini, 2009).

2.1. Domain Adaptation by Instance Transfer based on Logistic Regression

Paul et al. (2016) propose a method for DA based on instance transfer for logistic regression. In logistic regression, the class label C_i of a pixel i is determined by maximising the posterior $P(C_i \mid \mathbf{f}_i(\mathbf{x}))$ for C_i given some feature vector $\mathbf{f}_i(\mathbf{x})$ defined for that pixel:

$$P(C_i = C^k | \mathbf{f}_i(\mathbf{x})) = \frac{exp(\mathbf{w}_k^T \cdot \mathbf{f}_i(\mathbf{x}))}{\sum_a exp(\mathbf{w}_a^T \cdot \mathbf{f}_i(\mathbf{x}))}, \quad (1)$$

where C^k is a specific value for C_i, $\mathbf{w}=[\mathbf{w}_1, \ldots, \mathbf{w}_K]$ are the parameters to be determined in training, and K is the number of classes. The parameters $\mathbf{w}$ in Eq. 1 can be determined by maximising

$$p(\mathbf{w}|\overline{TD}) \propto p(\mathbf{w}) \cdot \prod_{i,j}\left[P\left(C_i = C^j|\mathbf{f}_i(\mathbf{x})\right)\right]^{g_i \cdot t_{ij}} \quad (2)$$

given a training dataset $\overline{TD}$ consisting of pairs of feature vectors $\mathbf{f}_i$ ($\mathbf{x}$) and corresponding class labels C_i. In Eq. 2, t_{ij} is a variable indicating if the value of the class label C_i of sample i is C^j or not, and g_i is a weight assigned to sample i. The prior $p(\mathbf{w})$ is essential for regularisation.

The training dataset $\overline{TD^0}$ is initialized by training samples from the source domain D_S, where class labels are assumed to be available. Consequently, the parameters $\mathbf{w}$ are determined by maximising Eq. 2, using $g_i = 1$ for all training samples and a zero mean Gaussian prior for $p(\mathbf{w})$.

In DA, the classifier trained using only source data is gradually adapted to the distribution of the data in D_T. In iteration t, a new training dataset $\overline{TD^t}$ is generated. Firstly, ρ source samples are removed from $\overline{TD^{t-1}}$, starting with samples on the wrong side of the decision boundary and continuing with samples on the correct side, in both cases ordering the samples by their distances from the decision boundary. Secondly, ρ samples from D_T are added to $\overline{TD^t}$, receiving their semi-labels from the current state of the classifier. The strategy for the selection of these samples is crucial for success; we select the target samples having the smallest average distance from their k nearest neighbours in $\overline{TD^{t-1}}$. Having thus defined a new training dataset $\overline{TD^t}$, the weights g_i of the samples in $\overline{TD^t}$ are determined so that samples that are close to the current decision boundary receive a lower weight than more distant ones. Finally, new values for the parameters $\mathbf{w}^t$ are determined by maximising Eq. 2, using the values $\mathbf{w}^{t-1}$ as the mean of the Gaussian prior. This procedure is repeated until no source samples are contained in $\overline{TD^t}$ (Paul et al., 2016).

In their experiments, Paul et al. (2016) considered pairs of image patches of the Vaihingen dataset (Wegner et al., 2014) as pairs of source and target domains. Applying logistic regression trained on the source domain to the target domain without DA resulted in a loss in overall accuracy (OA) larger than 5%. Using DA resulted in a positive transfer in 22 of the 36 test cases (61%); the average improvement in OA for these pairs due to DA was 4.7%. In the remaining cases, the prerequisites of the domains to be related was not fulfilled.

2.2. Source Selection and Domain Adaptation

Here we assume that there is a database of labelled images that can be used as source domains to train a classifier for a new (target) image. It is the goal of source selection to find the most appropriate source image for training a classifier to analyse a specific target image. To measure the similarity of domains, Vogt et al. (2017) propose a *supervised domain distance* d_{SDA}:

$$d_{SDA}(\overline{TD}_S, \overline{TD}_T) = \epsilon(h_s, \overline{TD}_S) + 2 \cdot d_{LMMD}(\overline{TD}_T, \overline{TD}_S). \quad (3)$$

In Eq. 3, $d_{SDA}(\overline{TD}_S, \overline{TD}_T)$ is a distance between the distributions of the data in the source and target domains and $\epsilon(h_s, \overline{TD}_S)$ is the classification error of the source task. The term $d_{LMMD}(\overline{TD}_T, \overline{TD}_S)$ is a linear kernel-based estimation of the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012). Vogt et al. (2017) also propose the *unsupervised domain distance* d_{UDA}:

$$d_{UDA}(\overline{TD}_S, \overline{TD}_T) = 2 \cdot d_{LMMD}(\overline{TD}_T, \overline{TD}_S), \quad (4)$$

which can be determined without training labels in the source domain. The optimal source $\bar{S}$ can be determined by minimizing either d_{SDA} or d_{SDA}. As the optimal source may still be relatively dissimilar from the target domain, Vogt et al. (2017) additionally suggest using a synthetic source S_π that is the linear combination of multiple sources and whose features are distributed according to:

$$p_{S_\pi}(\mathbf{x}) = \sum_s \pi_s \cdot p_{S^s}(\mathbf{x}), \quad (5)$$

where $p_{S^s}(\mathbf{x})$ is the distribution of the features $\mathbf{x}$ for source s and $\pi_s > 0$ is the corresponding weight with $\sum_s \pi_s = 1$. The MMD distance can also be determined for the synthetic domain S_π, and Vogt et al. (2017) propose a fast greedy scheme for determining the weights π_s.

In the experimental evaluation, Vogt et al. (2017) first determine the overall accuracy of a classifier trained on the target domain and then measure the difference in overall accuracy *ΔOA* that is obtained when a classifier trained on a selected source is applied to the target domain without additional domain adaptation. The results show that single source selection according to the distances in Eqs. 3 and 4 outperforms random source selection by a large margin, and synthetic source selection performs even better, with a slight advantage of d_{SDA} over d_{UDA}. Applying domain adaptation results in a small improvement if a synthetic source is used (Vogt et al., 2017).

3. LEARNING UNDER LABEL NOISE

In order to use existing maps for training a classifier to be applied to new data, the training procedure needs to be able to cope with errors in the training labels due to temporal changes (*label noise*; Frénay & Verleysen, 2014). In remote sensing, this problem is often dealt with by detecting and eliminating wrong training samples (*data cleansing*). An alternative is to use probabilistic methods for training under label noise that also estimate the parameters of a noise model. One such approach is the label noise tolerant logistic regression (Bootkrajang & Kabán, 2012) which is also used in (Maas et al., 2016). Eq. 1 is applied to classify each pixel *i* in the image to be classified, but training cannot be based on maximising the probability in Eq. 2, because the true class labels C_i of the training samples are unknown. We only know the observed labels $\underline{C_i}$ that are extracted from the existing map, and thus, we can determine the parameters **w** of the classifier (Eq. 1) by maximising

$$p(\mathbf{w}|\overline{TD}) \propto p(\mathbf{w}) \cdot \prod_{i,j}\left[P\left(\underline{C_i} = C^j|\mathbf{f}_i(\mathbf{x})\right)\right]^{t_{ij}}, \quad (5)$$

where t_{ij} indicates whether the observed label $\underline{C_i}$ is C^j, and $P\left(\underline{C_i} = C^j|\mathbf{f}_i(\mathbf{x})\right)$ is substituted by

$$P\left(\underline{C_i} = C^j|\mathbf{f}_i(\mathbf{x})\right) = \sum_a P\left(\underline{C_i} = C^j|C_i = C^a\right) \cdot P(C_i = C^a|\mathbf{f}_i(\mathbf{x})). \quad (6)$$

In Eq. 6, $P(C_i = C^a|\mathbf{f}_i(\mathbf{x}))$ is the required classifier related to the true labels and parameterised by **w** (Eq. 1.). Eq. 6 also introduces the parameters of a *probabilistic noise model*, namely the transition probabilities $P\left(\underline{C_i} = C^j|C_i = C^a\right)$ that describe how likely an observed label is to take the value C^j if the true label is C^a. Bootkrajang and Kabán (2012) present a method to determine the parameters **w** and $P\left(\underline{C_i} = C^j|C_i = C^a\right)$ in an iterative procedure similar to expectation maximisation.

Maas et al. (2016) could show that this training procedure can deal with a large amount of random noise, resulting in a loss in overall accuracy (OA) of only 4% even if 50% of the training labels were wrong. It also improved the OA compared to standard logistic regression by 1-2% in a more realistic scenario where the wrong labels form clusters corresponding to larger changes in the map.

Maas et al. (2017) proposed to use the observed class labels not only for training, but also for classification. They built a contextual classifier based on Conditional Random Fields (CRF; Kumar & Hebert, 2006). CRF try to maximise the joint posterior of all class labels (collected in a vector **C**) given the observations. In Maas et al. (2017), this joint posterior is factorised according to Eq. 7:

$$P\left(\mathbf{C}|\mathbf{x}, \underline{\mathbf{C}}\right) \propto \prod_i \varphi(C_i, \mathbf{x}) \prod_{i,j} \psi\left(C_i, C_j, \mathbf{x}\right) \prod_i \gamma\left(C_i, \underline{C_i}\right). \quad (7)$$

In Eq. 7, the subscript *i* denotes a particular pixel and (i, j) denotes a pair of neighbouring pixels on the image lattice. The vectors **x** and $\underline{\mathbf{C}}$ denote the observed image data and class labels from the map, respectively. The three terms in Eq. 7 are the association potential $\varphi(C_i, \mathbf{x})$, the interaction potential $\psi\left(C_i, C_j, \mathbf{x}\right)$ and the temporal association potential $\gamma\left(C_i, \underline{C_i}\right)$. The association potential is based on Eq. 1, trained using the label noise tolerant procedure outlined previously. For the interaction potential, we use a model for data-dependent smoothing (Kumar & Hebert, 2006). The new temporal association potential $\gamma\left(C_i, \underline{C_i}\right)$ links the observed and the true class labels. It is determined according to $\gamma\left(C_i, \underline{C_i}\right) = \left[P\left(C_i|\underline{C_i}\right)\right]^{\theta_i}$, where $P\left(C_i|\underline{C_i}\right)$ can be determined from the transition probabilities $P\left(\underline{C_i}|C_i\right)$ (Eq. 6) and θ_i is a site-specific weight that is modified in an

iterative classification process in which pixels inside compact clusters of potential change receive lower weights than pixels that are likely not to have changed (Maas et al., 2017).
Experiments have shown that considering the temporal association potential increases overall accuracy by more than 5% compared to (Maas et al., 2016).

4. CONCLUSIONS

This paper has presented some promising strategies for reducing the requirements for manually annotated training data. Domain adaptation can help to adapt a classifier trained using a source domain to a target domain without additional training data. Leveraging existing maps, classification can be carried out without manually annotated training data, and the map can also provide useful prior information to improve the classification accuracy.

5. ACKNOWLEDGEMENTS

The Vaihingen dataset was provided by the German Society of Photogrammetry, Remote Sensing and Geoinformation (DGPF) (Cramer, 2010). The support of Deutsche Forschungsgemeinschaft (DFG) under grants HE-1822/30-1 and HE-1822/35-1 is gratefully acknowledged.

6. REFERENCES

Bootkrajang, J., Kabán, A., 2012. Label-noise robust logistic regression and its applications. *Joint European Conf. on Machine Learning and Knowledge Discovery in Databases*, pp. 143–158.

Bruzzone, L., Marconcini, M., 2009: Toward the automatic updating of land-cover maps by a domain-adaptation SVM classifier and a circular validation strategy. *IEEE Transactions on Geoscience and Remote Sensing* 47(4): 1108–1122.

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

Frénay, B., Verleysen, M., 2014. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems* 25(5):845–869.

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

Kumar S., Hebert, M., 2006: Discriminative Random Fields. *International Journal of Computer Vision* 68(2):179–201.

Maas, A., Rottensteiner, F., Heipke, C., 2016. Using label noise robust logistic regression for automated updating of topographic geospatial databases. *ISPRS Annals* III-7, pp. 133–140.

Maas, A., Rottensteiner, F., Heipke, C., 2017. Classification under label noise using outdated maps *ISPRS Annals* IV-1/W1, pp. 215–222.

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

Paul, A., Rottensteiner, F., Heipke, C., 2016: Iterative re-weighted instance transfer for domain adaptation. *ISPRS Annals* III-3, 339-346.

Vogt, K., Paul, A., Ostermann, J., Rottensteiner, F., Heipke, C., 2017: Boosted unsupervised multi-source selection for domain adaptation. *ISPRS Annals* IV-1/W1, 229-236.

Wegner, J. D., Rottensteiner, F., Gerke, M., Sohn, G., 2016. The ISPRS 2D labelling challenge. http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html (visited 10/08/2017).