Image Classification by Transfer Learning Based on the Predictive Ability of Each Attribute

Masahiro Suzuki, Haruhiko Sato, Satoshi Oyama, and Masahito Kurihara

***Abstract*—Machine learning is the basis of important advances in artificial intelligence such as image and speech recognition and natural language processing. Unlike general machine learning, which uses the same task for training and testing, transfer learning uses the results trained by other tasks to learn a new task. Among the various transfer learning algorithms have been proposed to date, we focus on attribute-based transfer learning. This algorithm realizes transfer learning by introducing attributes and transferring the results of training. However, the existing algorithm does not consider the extent to which attributes contribute to predicting the target class (called the predictive ability in this study). Here, we weighted each attribute by the extent to which it contributes to the evaluation equation. We confirmed that the accuracy rate of the proposed technique was higher than that of the preceding work.**

***Index Terms*—transfer learning, attributes, multiclass classification, SVM.**

I. Introduction

MACHINE learning is the method by which patterns and knowledges are automatically learned from training data, and it is used to predict unseen test data. Machine learning has proven successful in diverse fields such as image recognition, speech recognition, and natural language processing. Many machine learning techniques require large datasets to overcome the over-fitting problem. In the real world, such large data samples are readily extractable from the Internet, and offer a means of improving the above problem. However, this approach does not work well in machine learning such as supervised learning because Internet-derived samples are almost unlabeled, and their feature spaces or distributions (i.e. source task or source domain) are different from that of the working problem (i.e. target task or target domain). To solve this problem, transfer learning can be applied. In the transfer learning[1][2] framework, source task data are used to train the target task by transferring prior knowledge acquired from the source task to the target task. The difference between traditional machine learning and transfer learning is illustrated in Figure 1. Among the various transfer methods, we focus on attribute-based transfer learning[3].

The transfer learning algorithm exploits the semantic knowledge of the object attributes such as shape, color, and texture. This knowledge is shared by all objects in the source and target tasks. Therefore, we can learn target tasks even if few or no training samples exist. Since human beings appear to also recognize unseen objects by transferring object attributes, this transfer learning approach is intuitive and natural. However, the extent to which the attributes contributed to the precision of the target is not considered in the existing algorithm. We define this property as the predictive ability. In this study, we assumes that the predictive ability of each attribute differs, but is the same for all classes. The weighted attributes are then introduced into a DAP model.

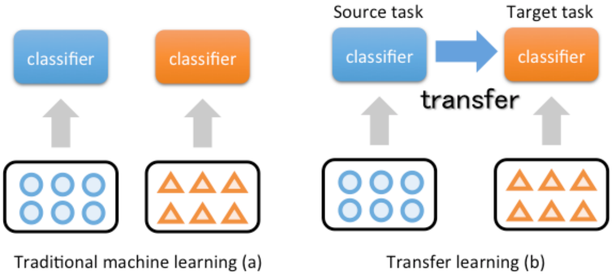

Fig. 1. Traditional machine learning(a) and transfer learning(b)

The remainder of this paper is organized as follows: Section II discusses the related work, and Section III introduces our approach. Experimental results are presented in Section IV. Section V concludes the paper and discusses ideas for future study.

II. Related Studies

Subsection II-A of this section describes the existing research on transfer learning. Attribute-based transfer learning, referred to as the DAP model, is presented in subsection II-B.

A. Transfer learning

Whereas traditional machine learning assumes the same feature space or distribution for both the training and test data, transfer learning allows them to be different. The data of the source task are used to train the target task by transferring prior knowledge acquired from source task to target task (as shown in Figure 1). Transfer learning was conceptualized long ago and has been assigned many names; inductive transfer, domain adaptation, multitask learning, and others.

The term *transfer learning* is used within the broad framework of machine learning; therefore, it eludes a precise definition and discussion. In 2005, the NIPS workshop on “ Inductive Transfer: 10 Years Later ”[4] defined transfer learning as the problem of retaining and applying the knowledge learned in one or more tasks to efficiently develop an effective hypothesis for a new task. A few surveys have been published on transfer learning[1][2].

B. Attribute-based transfer learning

Attribute-based classification is a computer vision algorithm that realizes transfer learning. This algorithm, which

Manuscript received January 8, 2014; revised January 29, 2014.

M. Suzuki, H. Sato, S. Oyama, and M. Kurihara are with the Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan 060-0814.
E-mail: masa@complex.ist.hokudai.ac.jp,
haru@complex.ist.hokudai.ac.jp,
oyama@ist.hokudai.ac.jp, kurihara@ist.hokudai.ac.jp

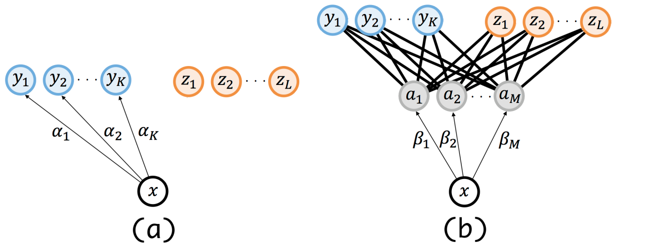

Fig. 2. Traditional machine learning(a) and attribute-based transfer learning(b)

has been investigated in several studies[5][6][7], is here called attribute-based transfer learning to emphasize its transfer learning property.

Let $(x_1, l_1), \ldots, (x_n, l_n) \subset X \times Y$ be training data samples, where X is an arbitrary feature space and $Y = \{y_1, \ldots, y_K\}$ consists of K discrete classes in the source task. Our goal is to learn a classifier: $X \rightarrow Z$ for target task $Z = \{z_1, \ldots, z_L\}$ that is different from Y.

Traditional machine learning requires training samples on $X \times Z$ to solve this problem. However, collecting new training samples for all classes is a difficult task, and we would prefer to exploit $X \times Y$. Attribute-based transfer learning is based on attributes, which constitute high-level semantic knowledge. In addition, each attribute is binary and shared among all classes. Therefore, information about each class can be obtain without collecting many samples and training because human beings can easily provide the relationships between attributes and classes.

This method, called Direct attribute prediction (DAP), is illustrated in Figure 2(b). Compared with traditional machine learning (Figure 2(a)), DAP introduces a middle layer consisting of attributes $A = \{a_1, \ldots, a_M\}$. If the relations between class y and corresponding attribute values, given by $a_y = (a_1^y, \ldots, a_M^y)$ are known in advance, we can learn the classifier: $X \rightarrow Y$ by learning the classifier β: between input features and correspond attributes.

The test data used in the test stage are samples belonging to the target task Z. Moreover, the relations between z and attribute values, denoted $a_z = (a_1^z, \ldots, a_M^z)$, are assumed to be known. Since the posterior of the test class z given the sample x can be expressed as $p(z|x)$, we can estimate the best output class from all test classes of the target task using MAP prediction:

$$\arg\max_z p(z|x) \tag{1}$$

Since the probability of attributes for a given input is formulated as $p(a|x) = \prod p(a_m|x)$, the posterior $p(z|x)$ can be calculated as follows:

$$p(z|x) = \sum_{a \in \{0,1\}^M} p(z|a)p(a|x) = \frac{p(z)}{p(a^z)} \prod_{m=1}^{M} p(a_m^z|x) \tag{2}$$

In Equation (2), the factor $p(a^z)$ is assumed as a factorial distribution $p(a^z) = \prod p(a_m)$ and is calculated by $p(a_m) = \frac{1}{K}\sum_{k=1}^{K} a_m^{y_k}$. Furthermore, $p(a_m^z)$ has already been learned as classifier β and the factor $p(z)$ can be ignored because all classes have the same prior probability. Therefore, we can estimate class z as follows:

$$\arg\max_z \prod_{m=1}^{M} \frac{p(a_m^z|x)}{p(a_m^z)} \tag{3}$$

III. Proposed method

As stated above, attribute-based transfer learning can recognize a new class by transferring the information $p(a_m^z|x)$. However, the extent of contribution of the attributes in predicting the target class is not specified in this approach. We define this new property as the predictive ability of the attribute.

When the predictive ability of the attribute is low, the influence of this attribute should be made small because the attribute makes a small but non-negligible contribution to class prediction. Furthermore, we assume that the predictive ability of each attribute is the same for all classes. Therefore, we propose to weight each attribute based on its predictive ability in Equation (3). Since weighting of Equation (3) is inconvenient, we apply weighting to the logarithm of Equation (3). Expressing the weight of attribute m as $weight_m$, the proposed equation is becomes:

$$\arg\max_z \sum_{m=1}^{M} weight_m log \frac{p(a_m^z|x)}{p(a_m^z)} \tag{4}$$

This equation is equivalent to

$$\arg\max_z \prod_{m=1}^{M} \left\{ \frac{p(a_m^z|x)}{p(a_m^z)} \right\}^{weight_m} \tag{5}$$

In other words, if the predictive ability of an attribute is higher, this attribute is regarded as more important and makes a higher contribution to the equation.

We assume that the predictive ability is mainly affected by two factors; first, whether the attribute is easily learned from the input data; second, the bias of the attribute value. In the training stage, because bias in attribute values indicates imbalanced data, the classifier of each attribute β_m cannot be properly learned. Therefore the influence of biased attributes should be reduced.

In terms of these factors, we re-express Equation (5) as

$$\arg\max_z \sum_{m=1}^{M} v_m w_m log \frac{p(a_m^z|x)}{p(a_m^z)} \tag{6}$$

where the weights v_m and w_m reflect the effects of the above-described first and second factors, respectively.

The first factor can be represented by the accuracy of each attribute. Therefore, the weight v_m is estimated as

$$v_m = E_X[\delta(p_m, a_m)] \tag{7}$$

where

$$x = y : \delta(x, y) = 1 \tag{8}$$

$$x \neq y : \delta(x, y) = 0 \tag{9}$$

p_m is the predicted value of attribute a_m.

In assigning w_m, we assume that the prediction value p_m is a better measure than the bias of the attribute value a_m because its bias is larger than that of a_m. Furthermore, the weight w_m should be minimized when the mean of the attribute value is 0 or 1, and maximized when the mean is 0.5. Therefore, the weight w_m is estimated as

$$w_m = 1 - 2|E_Y[p_m] - 0.5| \tag{10}$$

However, this equation is non-differentiable when $E_Y[p_m]$ is 0.5. To express w_m in a differentiable form, we define its entropy function as

$$\begin{aligned} w_m = & -E_Y[p_m]logE_Y[p_m] \\ & -\{1 - E_Y[p_m]\}log\{1 - E_Y[p_m]\} \end{aligned} \tag{11}$$

where $0log0$ is taken to be 0.

In this way, we have defined the weights v_m and w_m.

IV. Experimental results

Experiments were conducted on the " Animals with Attributes " data set[8]. This data set includes 30,475 images from 50 animal classes. The classes are defined by 85 attributes. The relations between classes and attributes are labeled by humans and presented in a 50×85 matrix. In the experiment, we selected 40 classes as the source task and the remaining classes as the target task.

In the " Animals with Attributes " dataset, each image is extracted by six types of features. We selected feature types SURF and RGB color histograms because these features yield the first and second highest accuracy rate, respectively, in the nearest neighbor algorithm[9].

Since the number of feature types is greater than one, we use Multiple Kernel Learnig(MKL)-SVM. The probability estimates from SVM are obtained by Platt scaling[10].

To verify the performance of our proposed method, we separately experiment on the weights v_m and w_m.

First, we estimated the weight v_m by two approaches. The first approach uses the accuracy of the test data to confirm that the accuracy rate is truly improved by the weighting. However, this approach is inappropriate in practice. The second approach estimates the accuracy by training 90% and testing the remaining 10% of the training data. We expect that both approaches will yield the same result.

Figure 3 shows the result of using the weight v_m. The vertical axis indicates the accuracy of the classification, and the horizontal axis denotes the number of training and test images in each class. The result is the average of six experimental runs. While the performance of the first approach is better than the existing approach, the second approach is not outperformed.

Next, we experiment on the weight w_m. This weight is expressed by two equations which are respectively evaluated to obtain a comparison. We also estimate the prediction value of weight w_m by training and testing 90% and 10% of the training data, respectively.

Figure 4 shows the result of using the weight w_m. In this experiment, both equations outperform the existing method, although Equation (10) is higher accuracy than Equation (11).

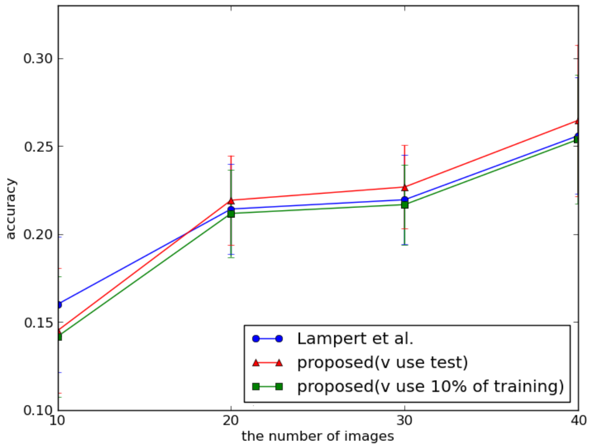

Fig. 3. Empirical evaluation using the weight v_m

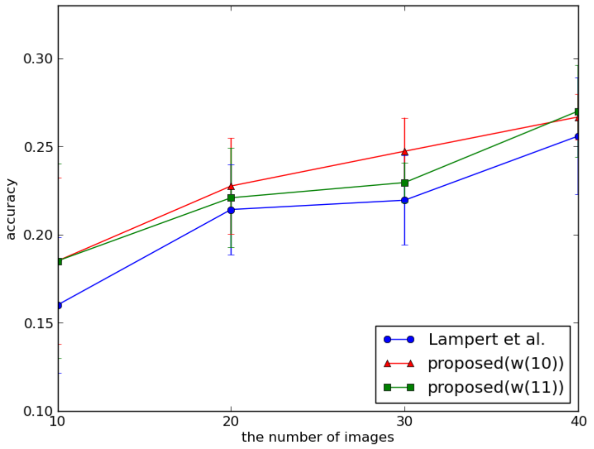

Fig. 4. Empirical evaluation using the weight w_m

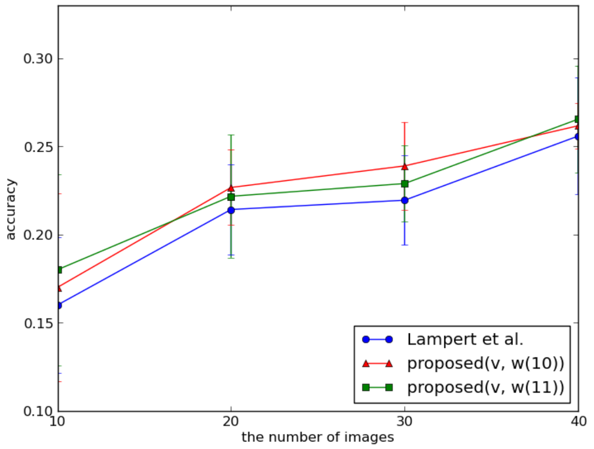

Fig. 5. Empirical evaluation using both weights v_m and w_m

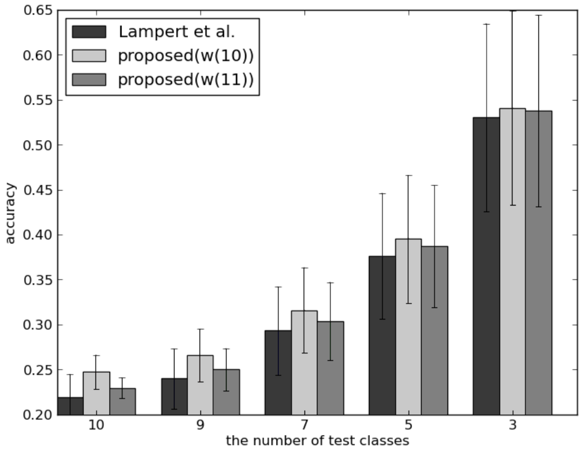

Fig. 6. The accuracy rates sorted by number of classes in the target task: each class has 30 images

Finally, we introduce both weights v_m and w_m. The results are plotted in Figure 5 in which proposed method also outperforms the existing method. However, since the accuracy of this result is similar to that of Figure 4, the weight v_m negligibly enhances the performance.

We then altered the number of classes in the target task, and evaluated the performance. In this experiment, we use only the weight w_m. Moreover, the number of images in each class is fixed at 30 and the results are the average values of six runs.

These results are plotted in Figure 6. The vertical axis indicates the classification accuracy, and the horizontal axis shows the number of classes in the target task. Again, our method outperforms the existing method.

V. Conclusion

This paper considers the predictive ability of attributes in attribute-based transfer learning, and improve performance is confirmed. However, a few weighting schemes are ineffective. Moreover, the experiment was restricted to zero-shot learning in which all classes in the target task are not present in the training set. In future work, we will experiment on other data sets and other situations, such as classes in the target task containing few samples. Furthermore, we aspire to improve the predictive ability of the existing approach by factors other than the weight.

References

[1] 神嶌敏弘. 転移学習. 人工知能学会誌, vol.25, no.4, pp.572-580 (2010).

[2] S.J. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, Vol. 22, No. 10, pp. 1345–1359, 2010.

[3] C.H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 951–958. IEEE, 2009.

[4] *Inductive Transfer: 10 Years Later - NIPS 2005 Workshop*. http://iitrl.acadiau.ca/itws05/.

[5] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR*, 2009.

[6] Xiaodong Yu and Yiannis Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *Proceedings of the 11th European Conference on Computer Vision: Part V*, ECCV'10, pp. 127–140, Berlin, Heidelberg, 2010. Springer-Verlag.

[7] Dhruv Kumar Mahajan, Sundararajan Sellamanickam, and Vinod Nair. A joint learning framework for attribute models and object descriptions. In Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc J. Van Gool, editors, *ICCV*, pp. 1227–1234. IEEE, 2011.

[8] Stefan Harmeling Jens Weidmann Christoph H. Lampert, Hannes Nickisch. animals with attributes a dataset for attribute based classification. http://attributes.kyb.tuebingen.mpg.de.

[9] Sandra Ebert, Diane Larlus, and Bernt Schiele. Extracting structures in image collections for object recognition. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *ECCV (1)*, Vol. 6311 of *Lecture Notes in Computer Science*, pp. 720–733. Springer, 2010.

[10] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.