Transfer Learning Based on the Observation Probability of Each Attribute

Masahiro Suzuki, Haruhiko Sato, Satoshi Oyama, and Masahito Kurihara
Graduate School of Information Science and Technology
Hokkaido University, Sapporo, Hokkaido 060–0815
Email: {masa, haru}@complex.ist.hokudai.ac.jp, {oyama, kurihara}@ist.hokudai.ac.jp

***Abstract*—Machine learning is the basis of important advances in artificial intelligence. Unlike the general methods of machine learning, which use the same tasks for training and testing, the method of transfer learning uses different tasks to learn a new task. Among the various transfer learning algorithms in the literature, we focus on the attribute-based transfer learning. This algorithm realizes transfer learning by introducing attributes and transferring the results of training to another task with the common attributes. However, the existing method does not consider the frequency in which each attribute appears in feature vectors (called the observation probability). In this paper, we present a generative model with the observation probability. By the experiments, we show that the proposed method has achieved a higher accuracy rate than the existing method. Moreover, we see that it makes possible the incremental learning that was impossible in the existing method.**

Keywords*—*transfer learning, attributes, multiclass classification, incremental learning, generative model.

I. Introduction

Machine learning has proven successful in diverse fields of information processing such as image recognition, speech recognition, and natural language processing. In general, machine learning techniques require large datasets to overcome the over-fitting problem. In the real world, you can sometimes solve this problem by taking an approach of obtaining large data samples from the Internet. 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 tasks or source domains) are different from those of the working problem (i.e. target tasks or target domains). 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 methods of transfer learning, we focus on attribute-based transfer learning [3][4] .

The attribute-based transfer learning algorithm exploits the semantic knowledge of the object attributes such as shape, color, and texture. This knowledge is shared by all classes in the source and target tasks. Therefore, this transfer learning approach can learn target tasks even if few or no training samples exist. Since it seems that human beings also recognize unseen objects by transferring object attributes, this approach is intuitive and natural. Moreover, it is much easier to define the relations between attributes and classes than to label huge

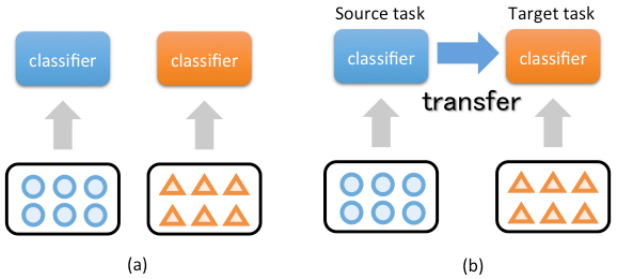

Fig. 1. Traditional machine learning (a) and transfer learning (b)

data.However, the frequency in which each attribute appears in input feature vectors is not considered in the existing method. We represent the frequency as the observation probability. In this study, we assume that the observation probability of each attribute differs from each other, but is common for all classes. Moreover, we develop a generative model and compare it with the existing method. Further, we study the possibility of its applicability to the incremental learning and verify the effectiveness of the proposed method.

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

II. Related work

Subsection II-A of this section describes the existing research on transfer learning. Attribute-based transfer learning, referred to as DAP, 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 the source task to the target task (as shown in Figure 1). Transfer learning was conceptualized long ago and has been called 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” [5] defined transfer learning as the

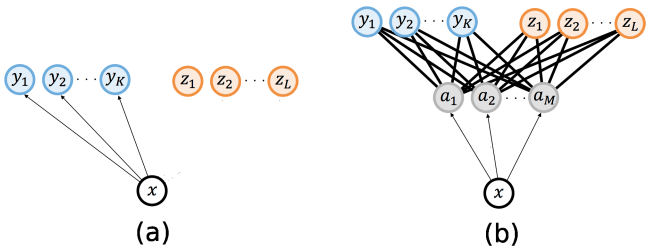

Fig. 2. Traditional machine learning (a) and attribute-based transfer learning (b)

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 [3][4] is a computer vision algorithm that realizes transfer learning. This algorithm, which has been investigated in several studies [6][7][8], is now 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: $X \to Z$ for L discrete classes in the 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 information from the training data $X \times Y$. 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 obtained without collecting many training samples because human beings can easily provide the relations 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 a class y and corresponding attribute values, given by $a_y = (a_1^y, \ldots, a_M^y)$ are known in advance, DAP can construct the classifier that classify input feature vectors into classes in the source task, by simply learning a set of classifiers, each of which determines the probability that the input vector has a certain attribute.

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

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

Since the probability of attributes for a given input is formulated as $p(a|x) = \prod p(a_m|x)$, the posterior probability $p(z|x)$ 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 $p(a_m) = \frac{1}{K} \sum_{k=1}^{K} m_m^{y_k}$. 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, DAP can estimate a 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 described in the previous section, attribute-based transfer learning can infer a class in the target task by sharing the classifier $p(a_m^z|x)$ of each attribute. However, it does not consider the frequency of attributes appearing in feature vectors. For example, the attribute of "black" frequently appears in feature vectors of the animal images represented as RGB values for pixels. By contrast, the attribute of "hunter" is hardly observed in images. In our previous work [9], we defined this concept and the bias of the attribute value as the predictive ability and considered it by weighting the logarithm of Equation (4) as

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

where $weight_m$ reflects the predict ability of attribute m.

Thereby, we confirmed that the accuracy rate of this method was higher than that of the existing method. However, the problem was that there was not an appropriate mathematical explanation to validate those weights in the equation of MAP estimation as Equation (4). In this study, we develop a new method based on the framework of the probability theory.

At first, we redefine the frequency in which each attribute appearing in feature vectors as the observation probability of the attribute. Unlike the predictive ability in [9], the bias of the attribute value is not considered in the observation probability. Moreover, we propose a generative model that realizes transfer learning by using the observation probability as a prior distribution in the different tasks.

The process of generating feature vectors is assumed to be Algorithm 1, and its graphical representation is illustrated in Figure 3. Since this model represents the generating process of the feature vectors in both source and target tasks, all classes in all tasks are denoted by z_n in this model unlike the existing method. A class z_n generates an attribute c_{mn} according to the relations between classes and corresponding attribute

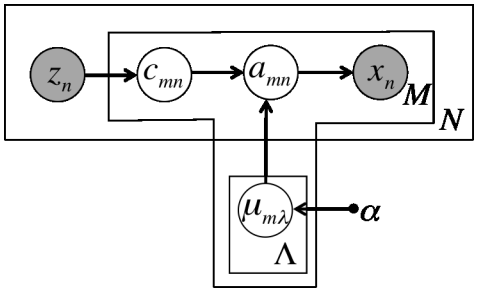

Fig. 3. The graphical model of the proposed generative model

Algorithm 1 The process of the proposed generative model

1: Choose $\mu_{m\lambda} \sim Beta(\alpha)$, where $m \in \{1, ..., M\}$, and $\lambda \in \{1, ..., \Lambda\}$
2: **for** $n \in \{1, ..., N\}$ **do**
3: Choose z_n randomly
4: **for** $m \in \{1, ..., M\}$ **do**
5: Choose c_{mn} from class-attribute matrix
6: Choose $a_{mn} \sim Bern(\mu_{c_{mn}})$, where $\mu_{c_{mn}}$ means $\mu_{m\lambda}$ on condition that $\lambda = c_{mn}$
7: Choose $x_n \sim f(a_{mn})$
8: **end for**
9: **end for**

values which are already known. Whereas an attribute c_{mn} is true value, an attribute a_{mn} is observed value. Therefore, we call c_{mn} a true attribute and a_{mn} an observed attribute. The observed attribute a_{mn} is generated by the discrete Bernoulli distribution the parameter of which is the observation probability and is denoted by $\mu_{m\lambda}$ as

$$Bern(a_{mn}|\mu_{m\lambda}) = \mu_{m\lambda}^{a_{mn}}(1-\mu_{m\lambda})^{1-a_{mn}} \tag{5}$$

where λ is binary and is the value of the true attribute as $\lambda = c_{mn}$.

Moreover, its conjugate prior is the beta distribution with parameters α as

$$Beta(\mu_{m\lambda}|\alpha,\alpha) = \frac{\mu_{m\lambda}^{\alpha-1}(1-\mu_{m\lambda})^{\alpha-1}}{B(\alpha,\alpha)} \tag{6}$$

where $B(\alpha,\alpha)$ is the beta function.

In the stage of the source task, the test feature space is given as X_{source}, and the MAP estimator for $\hat{\mu}_{m\lambda}$ is given as

$$\hat{\mu}_{m\lambda} = \arg\max_{\mu_{m\lambda}} p(\mu_{m\lambda}|X_{\text{source}}) \tag{7}$$

In order to calculate Equation (7), we maximize the logarithm of $p(\mu_{m\lambda}|X_{\text{source}})$ as

$$\frac{\partial}{\partial \mu_{m\lambda}} \log p(\mu_{m\lambda}|X_{\text{source}}) = 0 \tag{8}$$

Equation (8) can be calculated as follows:

$$\hat{\mu}_{m\lambda} = \frac{\sum_{n:c_{mn}=\lambda} p(a_{mn}=1|x_n) + \alpha - 1}{N_{m\lambda} + 2(\alpha-1)} \tag{9}$$

where the factor $p(a_{mn}|x_n)$ means the confidence value which is estimated by training and testing the input data in the source task. Moreover, $N_{m\lambda}$ means the amount of the test data which satisfy the condition that $c_m = \lambda$.

In order to calculate Equation (9) from Equation (8), we use Jensen's inequality as in the derivation of EM algorithm.

In the stage of the target task, the proposed method estimates $p(a_{mn}|x_n)$ by training and testing like the stage of the source task. The joint distribution of Figure 3 can be written as

$$\begin{aligned} &p(X, A, C, Z; \mu) \\ &= \prod_J P(z_n) \prod_m p(c_{mo}|z_n) p(a_{mn}|c_{mn}, \mu_{m\lambda}) p(x_n|a_{mn}) \end{aligned} \tag{10}$$

Hence,

$$p(X=x, Z=z) = p(z) \prod_m \sum_{a_m} p(a_m|\mu^z_{c_m}) p(x|a_m) \tag{11}$$

Therefore,

$$p(z|x) = \frac{p(x,y)}{p(x)} \propto \prod_m \sum_{a_m} \frac{p(a_m|\mu^z_{c_m}) p(a_m|x)}{p(a_m)} \tag{12}$$

In order to estimate the best output class z in the target task, we use MAP estimation such as Equation (1). According to Equation (12), the proposed method can estimate z as

$$\arg\max_u p(z|x) = \arg\max_z \prod_m \sum_a \frac{p(a_m|\mu_{c^z_m}) p(a_m|x)}{p(a_m)} \tag{13}$$

where c^z_m means the value of the true attribute corresponding to a class z, and $\mu_{c^z_m}$ means $\mu_{m\lambda}$ on condition that $\lambda = c_{mn}$.

Moreover, $p(a_m)$ is estimated as

$$\begin{aligned} p(a_m) &= \sum_n p(x, a_m) = \sum_n p(a_m|x) p(x) \\ &= \frac{1}{N_{\text{target}}} \sum_n p(a_m|x) \end{aligned} \tag{14}$$

where N_{target} means the amount of the test data in the target task.

In the existing method, transfer learning was realized by sharing the classifier. Therefore, it was required to train only in the source task and test only in the target task. In contrast, the proposed method uses the observation probability to transfer

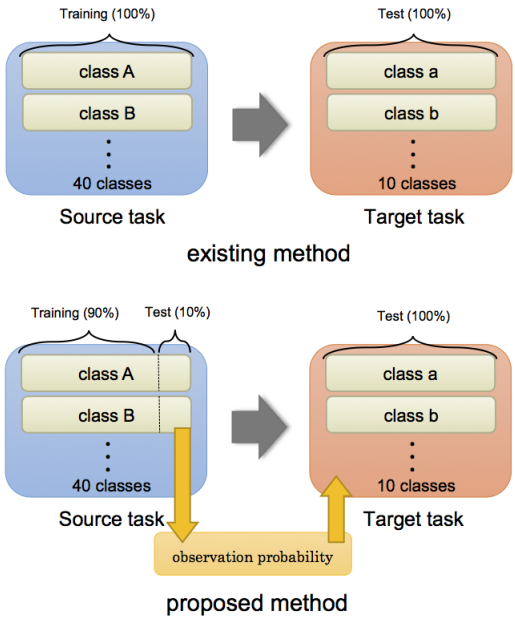

Fig. 4. Outline of the experiment 1

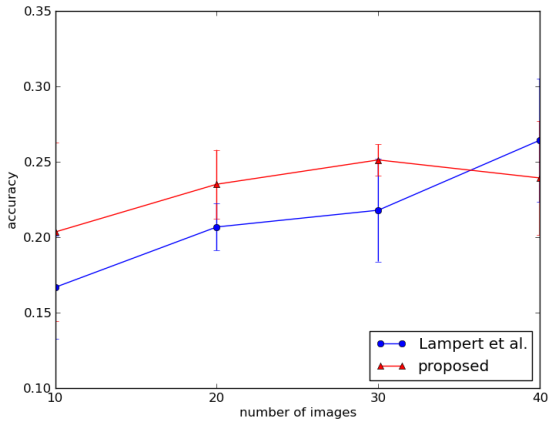

Fig. 5. Empirical evaluation of the experiment 1

the knowledge between different tasks. Hence, it can do both the training and test in each task. This means that incremental learning is possible in the proposed method.

There is a study of attribute-based transfer learning by the generative model as [7]. However, this study differs from our study in that input feature vectors must be codewords.

IV. Experimental Results

Experiments were conducted on the "Animals with Attributes" dataset[1]. This dataset 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 represented in a 50×85 matrix. In the experiments, we selected 40 classes as the source task and the remaining classes as the target task.

[1]http://attributes.kyb.tuebingen.mpg.de

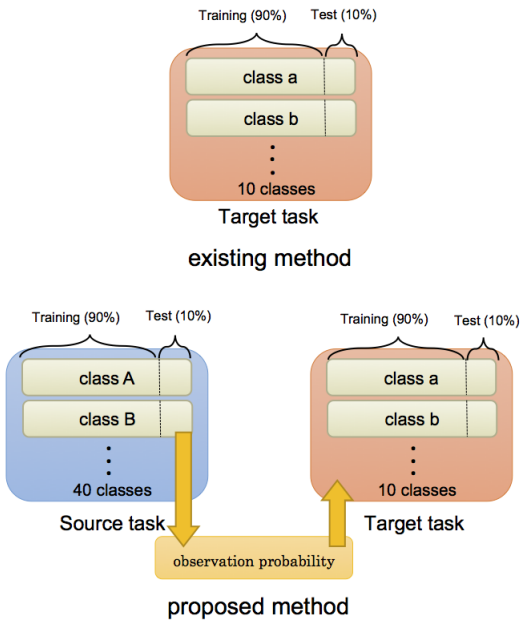

Fig. 6. Outline of the experiment 2

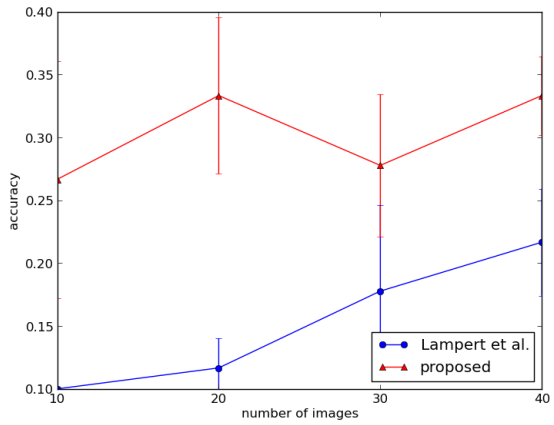

Fig. 7. Empirical evaluation of the experiment 2

In the "Animals with Attributes" dataset, each image is associated with 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 [10].

Since the number of feature types is greater than one, we used Multiple Kernel Learning (MKL) -SVM. The probability estimates from SVM are obtained by Platt-scaling [11]. Moreover, we implemented programs in Python, and used the SHOGUN machine learning toolbox[2].

We conducted experiments with two different settings, experiment 1 and experiment 2.

[2]http://www.shogun-toolbox.org

A. Experiment 1

We compared the accuracy rate of the existing method with the proposed method. We tackled the problem of zero-shot learning in which no classes of the target task are presented in the training set. Therefore, these methods used the data of the source task in training and inferred a class in the target task. In the proposed method, 10% of the training data were used in order to estimate the observation probability. The outline of this method is illustrated in Figure 4.

Figure 5 shows the result of zero-shot learning. 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 three experimental runs. In this experiment, our method almost outperformed the existing method. However, when the number of images was large such as 40, our method did not outperform the existing method.

B. Experiment 2

Next, we conducted experiments on training in the target task. While the existing method cannot transfer the knowledge in such a situation, our method can transfer the knowledge of the source task by using the observation probability (called incremental learning). Therefore, this experiment was carried out to confirm whether incremental learning by using the observation probability is effective. Figure 6 shows the outline of this experiment. As in this figure, these methods used 90% of the target task data in training and inferred the classes of the remaining 10% of the data. In addition, the method for estimating the observation probability was same as experiment 1.

Figure 7 is the result of this experiment. The horizontal axis indicates the sum of the number of the training and test images of each class. For example, if the number of images is ten, the number of training and test images of each class is nine and one respectively. This figure shows that the performance of our approach was better than the existing approach. Hence, we confirmed that the incremental learning worked well.

V. Conclusion

In this study, we proposed the observation probability of attributes in attribute-based transfer learning. Further, we confirmed the following two observations: (1) The accuracy rate of attribute-based transfer learning was improved by introducing the observation probability in most cases. However, when the number of images was large, our method did not outperform the existing method. (2) We confirmed that the incremental learning by using the observation probability was effective.

However, there are some remaining issues. First, in the proposed graphical model, we assumed that probabilities $p(x|a_m)$ are independent to each other. In practice, it is hard to imagine that input feature vectors are generated in such a process. Therefore, we should devise a probabilistic model that is more reflective of the real data generation process. Next, the proposed method performed MAP estimation for $\mu_{m\lambda}$ and class z_n. Hence, these values may have been trapped at a local optimum solution. Therefore, we will estimate them by using Bayesian estimation. Further, we must prepare the class-attribute matrix before we tackle the transfer learning problem. In future work, we plan to develop a method to reduce the burden to develop this matrix.

References

[1] T. Kamishima. Transfer learning. *Journal of Japanese Society for Artificial Intelligence*, Vol. 25, No. 4, pp. 572–580, 2010.

[2] S.J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 951–958, 2009.

[4] C.H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, Vol. 36, No. 3, pp. 453–465, 2014.

[5] Inductive transfer: 10 years later. In *NIPS 2005 Workshop*, 2005.

[6] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1778 – 1785, 2009.

[7] X. Yu and Y. Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Vol. V, pp. 127–140, 2010.

[8] D. Mahajan, S. Sellamanickam, and V. Nair. A joint learning framework for attribute models and object descriptions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1227–1234, 2011.

[9] M. Suzuki, H. Sato, S. Oyama, and M. Kurihara. Image classification by transfer learning based on the predictive ability of each attribute. In *Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS)*, Vol. 1, pp. 75–78, 2014.

[10] S. Ebert, D. Larlus, and B. Schiele. Extracting structures in image collections for object recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Vol. I, pp. 720–733, 2010.

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