MULTIPLE KERNEL ACTIVE LEARNING FOR IMAGE CLASSIFICATION

Jingjing Yang[1,2], Yuanning Li[1,2], Yonghong Tian[3], Lingyu Duan[3], Wen Gao[1,3]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, China
[2]Graduate School, Chinese Academy of Sciences, Beijing, 100039, China
[3]The Institute of Digital Media, School of EE & CS, Peking University, Beijing, 100871, China
{jjyang, ynli}@jdl.ac.cn, {yhtian, lingyu, wgao}@pku.edu.cn

ABSTRACT

Recently, multiple kernel learning (MKL) methods have shown promising performance in image classification. As a sort of supervised learning, training MKL-based classifiers relies on selecting and annotating extensive dataset. In general, we have to manually label large amount of samples to achieve desirable MKL-based classifiers. Moreover, MKL also suffers a great computational cost on kernel computation and parameter optimization. In this paper, we propose a local adaptive active learning (LA-AL) method to reduce the labeling and computational cost by selecting the most informative training samples. LA-AL adopts a top-down (or global-local) strategy for locating and searching informative samples. Uncertain samples are first clustered into groups, and then informative samples are consequently selected via inter-group and intra-group competitions. Experiments over COREL-5K show that the proposed LA-AL method can significantly reduce the demand of sample labeling and have achieved the state-of-the-art performance.

Index Terms— Multiple kernel learning, active learning, image classification

1. INTRODUCTION

Classifying images into a number of predefined categories is an important yet challenging task. To address the well-known "semantic gap" issue, many researchers resort to advanced machine learning techniques for mapping low-level visual features to high-level concepts. Remarkably, multiple kernels learning (MKL) methods, which optimize the classifiers via a linear combination of kernels, have shown prominent advantages in image classification [1, 2].

However, the computational complexity of MKL is very high for two major reasons: 1). Similar to normal kernel-based methods, MKL needs to compute kernel functions for each sample-pair over the training set; 2). MKL needs to optimize the classifier parameters and kernel weights in an alternative manner, thus learning global optimal parameters would incur intensive computation. To speed up the process of learning MKL, many research efforts have been done [3, 4]. In particular, as the size of training set becomes much bigger, the higher complexity of computing kernel matrix would become a bottleneck for efficiently training MKL-based classifiers. Intuitively, we may remove those redundant data and keep more informative samples to control the size of training data while maintaining the classifier's comparable discriminative power. Unfortunately, much fewer works have been made on sample selection for MKL, which could be a crucial step for learning classifier over large dataset.

Active learning is one of widely used methods to reduce the labeling cost in supervised learning tasks. It repeatedly queries the unlabeled samples and selects the most informative samples to label, which aims to reduce the demand for a large quantity of labeled data [5]. If active learning could be incorporated into MKL, kernel matrix can be computed over those selected sample-pairs only, and then classifier parameters and kernel weights can be updated by training over informative samples. Hence, active learning is helpful in reducing the complexity of kernel matrix computation and optimal parameters learning.

In the past decade, a great deal of active learning approaches were developed by using different learning models and sample selection strategies [5, 6, 7]. In image retrieval, a well-known active learning technique is support vector machine (SVM) active learning [6], which learns a SVM classifier from feedback images and employs the classifier to find the most informative but unlabeled images. However, this method is designed to select a single image in each learning round. More recently, some active learning methods are proposed for batch querying at each round [7]. In [7], a semi-supervised SVM batch active learning approach is proposed to take into account the "batch sampling problem" for image retrieval. However, this approach needs to compute kernel matrix among all unlabeled image-pairs and solve quadratic programming to select informative images, thus leading to higher computational complexity.

In this paper, we propose a novel approach called local adaptive active learning (LA-AL) which combines both multiple kernel learning and adaptive data sampling strategy. Our main contributions are summarized as follows:

- We incorporate the advantages of active learning in MKL framework to reduce computational complexity.
- We propose an effective and efficient LA-AL approach to automatically adapting sample selection with local data distribution.

- We have achieved significant improvements over recently reported results [7, 8] on the COREL-5K dataset.

The remainder of this paper is organized as following. Section 2 gives an overview of LA-AL. Section 3 presents our multiple kernel based active learning. Experiment results are presented in Section 4. Section 5 concludes the paper.

2. FRAMEWORK OVERVIEW

In this section, we brief the framework of our proposed LA-AL. The main processes are illustrated in Fig.1. Firstly, a preliminary MKL-based classifier is trained on an initial labeled dataset. Secondly, the images from the unlabeled data pool are filtered by the learnt classifier, leaving uncertain samples for further selection. Thirdly, a hierarchical sample selection is conducted, and a batch of informative samples is selected and labeled by user. Fourthly, classifier goes on updating on the labeled batch data to refine its discriminative power. After the four steps above, a complete round is finished and active learner jumps to the next step and continues.

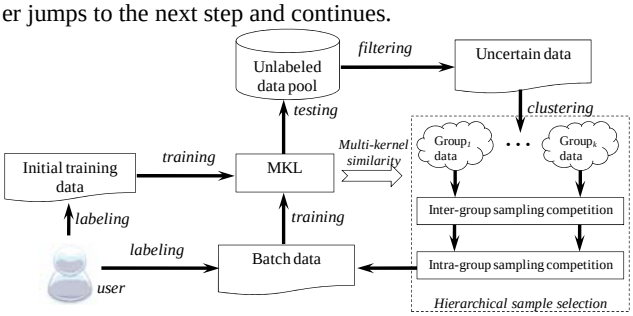

Figure 1. Local adaptive active learning framework

In LA-AL, a key process is the hierarchical sample selection, in which informative samples are selected in a top-down manner. At the beginning, all uncertain samples are firstly clustered into groups. At the level of inter-group, different groups compete to be selected as many samples as possible by leveraging their representativeness and informativeness. At the level of intra-group, competition is conducted within a group to select samples by dynamic bin-width histogram which adapts with the local data distribution. Consequently, finding informative samples in the LA-AL is a global to local, coarse to fine locating and searching process.

3. MULTIPLE KERNEL ACTIVE LEARNING

3.1. Learning an MKL-based classifier

Given a labeled training set $L = \{(x_1, y_1), ..., (x_{N_l}, y_{N_l})\}$ and an unlabeled set $U = \{(x_1, y_1), ..., (x_{N_u}, y_{N_u})\}$, where x is the image sample and $y \in \{-1,1\}$, MKL takes advantage of a convex combination of kernels [3] as follows:

$$K(x_i, x_j) = \sum_{m=1}^{M} \beta_m K_m(x_i, x_j), \text{with } \sum_{m=1}^{M} \beta_m = 1 \text{ and } \beta_m \geq 0, \quad (1)$$

where $K(x_i, x_j)$ measures the similarity between x_i and x_j, $K_m(\cdot,\cdot)$ is a kernel function which satisfies the Mercer's condition [3], M is the total number of kernels, and $\{\beta_m\}_{m=1}^{M}$ are the kernel weights. In MKL, $K_m(\cdot,\cdot)$ may employ various kernel functions. In this paper, we outline the decision function of MKL for binary classification:

$$f(x) = \sum_{i=1}^{N_l} \alpha_i y_i \sum_{m=1}^{M} \beta_m K_m(x_i, x) + b, \quad (2)$$

where $\{(x_i, y_i) \in L\}_{i=1}^{N_l}$ denotes training samples, $\{\alpha_i\}_{i=1}^{N_l}$ and b are the coefficients of the classifier. We refer the readers to [3] for the details of MKL.

The coefficients $\{\alpha_i\}_{i=1}^{N_l}$ and the kernel weights $\{\beta_m\}_{m=1}^{M}$ can be learnt by solving a joint optimization problem as follows.

$\underset{\beta}{\text{Max}}\underset{\alpha}{\text{Min}}\ J$, where

$$J = \frac{1}{2}\sum_{i=1}^{N_l}\sum_{j=1}^{N_l} \alpha_i \alpha_j y_i y_j \left(\sum_{m=1}^{M} \beta_m K_m(x_i, x_j)\right) - \sum_{i=1}^{N_l} \alpha_i, \quad (3)$$

$$\text{s.t.} \sum_{i=1}^{N_l} \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C,\ \sum_{m=1}^{M} \beta_m = 1,\ \beta_m \geq 0,$$

where C is regularization parameter and we set C= 1000 by cross-validation. Usually, training a MKL-based classifier not only needs to solve the max-min problem above, but also to compute kernel functions for each image-pair in the training set L. Hence, removing redundant training samples can efficiently speed up the learning process of MKL.

3.2. Local Adaptive Active Learning

In the LA-AL, informative samples are selected from a top-down procedure, involving global grouping, inter-group and intra-group competitions.

3.2.1. Filtering and Grouping Uncertain Samples

The key idea of our LA-AL approach is to treat uncertain samples as the candidate informative samples since such samples may offer more information to the learner. As proven in [7], the most informative samples should be selected from unlabeled samples close to the learnt decision boundary. Hence, we define the *uncertain samples* as those unlabeled samples near to the decision boundary, according to the score *f(x)* of the MKL-based classifier:

$$Unc(x) = \begin{cases} 1 & \text{if } T^- \leq f(x) \leq T^+ \\ 0 & \text{otherwise} \end{cases}, \quad (4)$$

where T^- and T^+ are the negative and positive bounds, respectively. These two bounds can be user-defined or estimated using a heuristic method.

Then the uncertain samples are further clustered into groups, so that similar samples are merged into one group for sampling competition. We use k-means for clustering in this paper, while other clustering methods can be applied.

3.2.2. *Inter-Group Sampling Competition*

In this sub-section, we present the inter-group sampling strategy aiming at allocating different sampling numbers for groups. Two criteria are considered here: *representativeness* and *informativeness*.

Intuitively, groups with larger sample quantity are likely to be allocated with more selected samples. Hence, we define the representativeness measure of group g as follows.

$$Rep(g) \propto N_g^{unc}, \tag{5}$$

where N_g^{unc} is the number of the uncertain sample in group g. Entropy is employed to measure the informativeness of each group as follows:

$$Info(g) \propto -\sum_{b_j^{equ} \in g} P(b_j^{equ} \mid g) \cdot \log P(b_j^{equ} \mid g), \tag{6}$$

where b_j^{equ} is the j*th* entry of an equal bin-width histogram which represents the multi-kernel similarity (see Eqn.(1)) distribution for the samples within g. Fig.2.(a) is an illustration of an equal bin-width histogram for the multi-kernel similarity distribution.

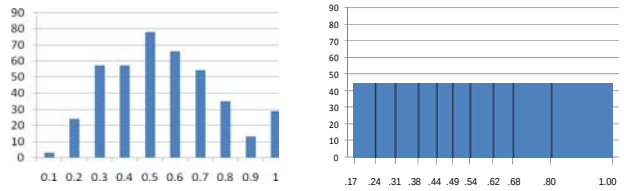

Figure 2. Two cases of multi-kernel similarity histograms: (a). Equal bin-width; (b). Dynamic bin-width. X-axis stands for the multi-kernel similarity between samples and its group center, y-axis denotes the sample quantity for different bins.

To seek a trade-off between representativeness and informativeness, we calculate the sampling number of each group as a linear combination of them:

$$N_g^s = N^s(\gamma \cdot Rep(g) + (1-\gamma) Info(g)), \tag{7}$$

where N^s is the predefined sampling number in the current round, N_g^s is the number of samples to be selected in group g , and γ is the parameter for adjusting the importance of each criteria. γ can be estimated by cross-validation .

3.2.3. *Intra-Group Sampling Competition*

In order to investigate the distribution of samples within a group, we employ a dynamic bin-width histogram (DBH), which contains an equal number of samples for all bins. An illustration of DBH for a group's sample distribution is shown in Fig.2. (b). Compared with equal bin-width histogram, DBH automatically adapts to the sample distribution. If more samples are distributed densely, bins tend to have narrow bin-width. No matter how close the samples are, their similarity will be displayed by the bin's ordinate on *x*-axis, while fixed bin-width histogram fails to deal with this variation.

Within a group, different bins of the DBH compete to be allocated with different sampling numbers. Criterion of *diversity* is considered. For a bin with large bin-width, samples with diverse appearances are scattered in a sparse area. Then the selected samples should cover as much diversity as possible. Hence, we allocate the number of samples for each bin as follows:

$$N_{b_j^{dyn}}^s = N_g^s Div(b_j^{dyn}), \tag{8}$$

where $Div(b_j^{dyn}) \propto width(b_j^{dyn})$, b_j^{dyn} is the j*th* entry of the DBH.

By the above inter-group and intra-group competition, we can locate where to sample data. Then we sort the candidate samples within the located bin via two criteria: sample's representativeness within the bin, and its diversity with the labeled set L. They are defined as follows:

$$Rep(x) = \sum_{x_k \in b_j^{dyn}} Sim(x, x_k) / N_{b_j^{dyn}}^s, \tag{9}$$

$$Div(x) = 1 - \max_{x_k \in L} Sim(x, x_k), \tag{10}$$

where $Sim(\cdot,\cdot)$ is the multi-kernel similarity learnt from MKL. Then the query function can be linear combination of two criteria to search informative samples, i.e.

$$q(x) = \lambda Rep(x) + (1-\lambda) Div(x), \tag{11}$$

where λ represents the balance of the two criteria and can be estimated by cross-validation. Finally, samples with the highest scores *q(x)* are selected for user labeling.

4. EXPERIMENTS

Our experiments are carried out over the COREL-5K image dataset. This dataset is composed of 50 categories and each containing 100 images culled from the COREL image CDs. We compare our proposed method (MKL-LA-AL) with representative learning methods including unbiased active Learning (UAL) [8], SVM active learning (SAL) [6], transductive SVM active learning (TSVM-SAL) [9], semi-supervised active learning (SVM-SSAL) [7], and MKL with random sampling (MKL-Ran).

4.1. Features and Kernels

SIFT [10] and Dense Color-SIFT [11] are employed as local descriptors to represent an image. And k-means is used to quantize these descriptors to obtain codebooks with a size of k (say, 400).

We implement Spatial Pyramid Kernels (SPK) [11] and Proximity Distribution Kernels (PDK) [12]. For SPK, an image is divided into cells and the features from the spatially corresponding cells are matched between images. For PDK, matching is done between local feature distributions of the K-nearest neighbors.

4.2. Performance Evaluation

In our experiment, 2.5k images from different categories are mixed up to form the unlabeled data pool for active learning. The remaining 2.5k images are used for testing. Initially, 5 images are selected in each category for labeling and training. At the stage of active learning, we query and label 250 images (on average 5 positive samples for each of the 50 categories) at each round in total 6 rounds. For fair comparison with other works, we follow the experimental setting in [8, 9]. The average precision (AP) of the top 20 returned images is utilized as the evaluation metric.

In Fig.3, we compare the performances of MKL-LA-AL and other methods including UAL, SAL, TSVM-SAL, SVM-SSAL, and MKL-Ran. As shown in the figure, two MKL based learning methods obtain significant improvement (about 30%) over other methods. This may be attributed to more discriminative power of MKL for image classification, where multiple features and kernel functions are combined. By taking advantage of LA-AL, our MKL-based method achieves the highest performance in all rounds.

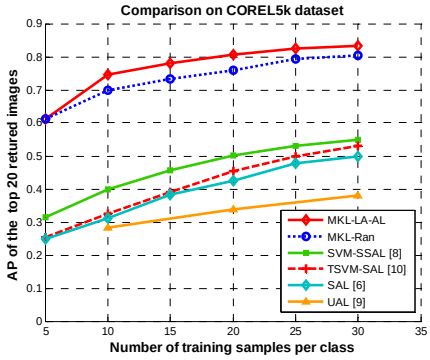

Fig.3. AP of top 20 returned images using MKL-LA-AL and other methods

To further investigate the effectiveness of LA-AL, we list the APs of MKL-Ran and MKL-LA-AL in Table 1. From the table, we observe that MKL-LA-AL achieves different degrees of improvements in different rounds, which can be attributed to the benefit of MKL-LA-AL in selecting more informative samples than the random selection in each round. When the number of training samples per class reaches up to 20, MKL-LA-AL achieves AP 80.7%, which is comparable with the highest AP 80.3% of MKL-Ran in all rounds.

Table 1. The AP of top 20 returned images

# Labels	5	10	15	20	25	30
MKL-Ran	0.612	0.700	0.732	0.758	0.793	0.803
MKL-LA-AL	0.612	0.745	0.779	0.807	0.824	0.832

In summary, MKL-LA-AL has shown great advantages in effectiveness and efficiency. In term of effectiveness, compared with the latest active learning methods, our method achieves the best results on the COREL-5K dataset. For efficiency, compared with MKL using randomly sampling, the proposed method not only achieves higher performance, but also keeps comparable performance with less training samples and lower computation complexity.

5. CONCLUSION

In this paper, we propose a local adaptive active learning (LA-AL) approach for multiple kernel learning. LA-AL has suggested a top-down locating and searching process, which select informative samples from grouped uncertain samples via both inter-group and intra-group competitions. To evaluate the performance of LA-AL, experiments have been conducted on COREL-5K dataset. Extensive comparison results show that our proposed MKL-LA-AL not only outperforms the latest active learning methods and MKL with random sampling, but also achieves an equal performance using less labeled data compared with MKL.

6. ACKNOWLEDGMENTS

The work is supported by grants from Chinese NSF under contract No. 60605020 and No. 90820003, National Hi-Tech R&D Program (863) of China under contract 2006AA010105, and National Basic Research Program of China under contract No. 2009CB320906. This work was partially supported by the research fund from NLPR, Institute of Automation, Chinese Academy of Sciences, and partially supported by the research award of Microsoft Research Asia Internet Services Theme.

7. REFERENCES

[1] M. Varma, and D. Ray, "Learning The Discriminative Power-Invariance Trade-Off," *in Proc.ICCV* , 2007.

[2] J. J. Yang, Y. Li, Y. Tian, L. Duan and W. Gao, "A New Multiple Kernel Approach for Visual Concept Learning," *in Proc.MMM*, 2009.

[3] S.Sonnenburg, G. Raetsch, C.Schaefer, and B.Scholkopf, "Large scale multiple kernel learning," *Journal of Machine Learning Research*, pp.1531–1565, 2006.

[4] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet, " More efficiency in multiple kernel learning," *in Proc. ICML*, 2007.

[5] S. Tong, and D. Koller, "Support vector machine active learning with applications to text classification," *The Journal of Machine Learning Research*, vol. 2, pp. 45 – 66, 2002.

[6] S. Tong, and E. Chang, "Support vector machine active learning for image retrieval," in *Proc. ACM Multimedia*, 2001.

[7] S.C. Hoi, R. Jin, J. Zhu, and M.R. Lyu, "Semi-supervised SVM batch mode active learning for image retrieval," *in Proc.CVPR* 2008.

[8] B. Geng, L. J. Yang, J. Z. Zheng, C. Xu, X.S. Hua, "Unbiased active learning for image retrieval," *in Proc. ICME*, 2008.

[9] L. Wang, K. L. Chan, and Z. Zhang, " Boots trapping SVM active learning by incorporating unlabelled images for image retrieval," *in Proc. CVPR*, 2003.

[10] D. Lowe, "Object recognition from local scale-invariant features," *in Proc.ICCV*, 1999.

[11] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," *in Proc. CVPR*, 2006.

[12] L. Haibin, and S. Soatto, "Proximity Distribution Kernels for Geometric Context in Category Recognition," *in Proc.ICCV*, 2007.