MULTI-VIEW MULTI-LABEL ACTIVE LEARNING FOR IMAGE CLASSIFICATION

Xiaoyu Zhang[†‡], *Jian Cheng*[†‡], *Changsheng Xu*[†‡], *Hanqing Lu*[†‡], *Songde Ma*[†]

[†] National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[‡] China-Singapore Institute of Digital Media, 119615, Singapore
{xyzhang, jcheng, csxu, luhq, masd}@nlpr.ia.ac.cn

ABSTRACT

Image classification is an important topic in multimedia analysis, among which multi-label image classification is a very challenging task with respect to the large demand for human annotation of multi-label samples. In this paper, we propose a multi-view multi-label active learning strategy, which integrates the mechanism of active learning and multi-view learning. On one hand we explore the sample and label uncertainties within each view; on the other hand we capture the uncertainty over different views based on multi-view fusion. Then the overall uncertainty along the sample, label and view dimensions are obtained to detect the most informative sample-label pairs. Experimental results demonstrate the effectiveness of the proposed scheme.

Index Terms — Active learning, Multi-view learning, Image classification, Multi-label classification, Multi-view fusion

1. INTRODUCTION

Image classification at the semantic level has emerged as an important topic in multimedia analysis. Existing researches on image classification mainly fall into two scenarios: the *multi-class* and *multi-label* classification [1][2]. In the multi-class setting, each image can only be annotated with a single label. While in most cases, especially for the real-world images, multi-label classification is a better choice, in which one or more labels can be assigned to each image. In this paper, we focus on the multi-label image classification problem. With more labels incorporated, the annotation of multi-label images will become much more time-consuming and labor-intensive compared to multi-class problem. Therefore, effective algorithms are needed to alleviate the burden on human labeling.

Active learning is an effective method for efficient labeling, and has been widely used in image classification. The main idea of active learning is to iteratively select the "most informative" samples to label so that the classification performance can be optimally boosted. Earlier active learning approaches mainly focused on the multi-class setting [1][3]. In recent years, some algorithms have been proposed to deal with the multi-label active learning problems [4][5], which convert multi-label classification to a combination of several multi-class classification problems. However, these approaches handle each label independently, neglecting the correlation embedded in the multiple labels. To solve this problem, a *two-dimensional active learning* (*2DAL*) scheme was proposed [6], which selects sample-label pairs instead of samples in each iteration. By considering the redundancy of multi-label samples along both the sample and label dimensions, 2DAL significantly reduced the requirement for human labeling.

Multi-view learning is another important mechanism which reduces the amount of labeled samples required for learning. It is often applied to problems with multiple views (or representations). In multi-view learning, multiple hypotheses (or classifiers) are trained separately from the same labeled data set, then the agreement (or disagreement) among different learners can be utilized to improve the overall classification performance. The widely used multi-view learning algorithms include co-training [7] and co-EM [8], which utilize the information acquired in one view to train the other. The idea of multi-view learning and active learning can be effectively integrated. As a family of multi-view active learning, co-testing [9] selects the most informative unlabeled samples by detecting the contention samples on which the multi-view predictions disagree.

In this paper, we extend 2DAL to multi-view setting and propose a *Multi-view Multi-label Active Learning* scheme for image classification. We use *intra-view uncertainty* to represent the sample and label uncertainties within each view, and *inter-view uncertainty* to reflect the uncertainty across multiple views. Based on the integrated uncertainty along the three dimensions, the most informative sample-label pairs are selected for annotation. Experimental results demonstrate that by taking advantage of both active learning and multi-view learning the demands for annotation can be effectively reduced.

2. MULTI-VIEW MULTI-LABEL ACTIVE LEARNING

In this section, we introduce the proposed multi-view multi-label active learning scheme (Figure 1) in detail.

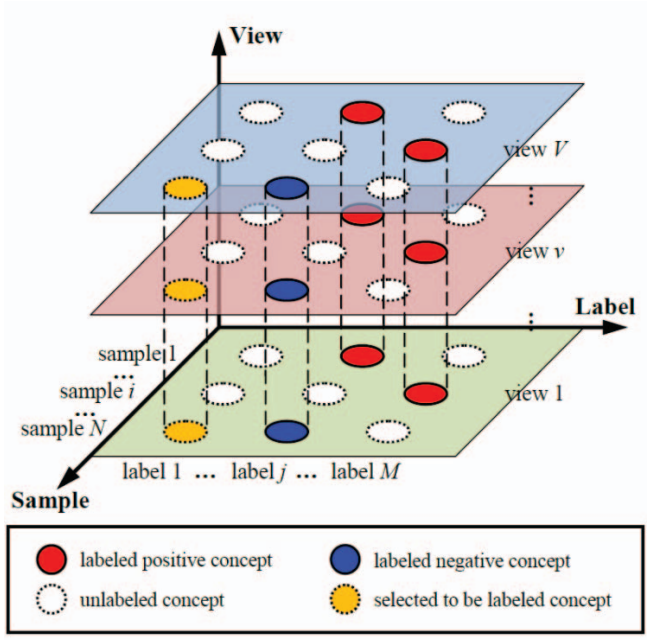

Figure 1. Multi-view multi-label active learning.

2.1. Intra-view Uncertainty

As discussed in [6], there exist redundancies in multi-label samples along different labels as well as different samples. Traditional one-dimensional active learning merely reduces the sample redundancy. In contrast, 2DAL reduces both sample and label redundancies simultaneously by selecting the most informative sample-label pairs.

Now we define some notations. Each sample $\boldsymbol{x}$ in the dataset X is attached with a set of labels $\boldsymbol{y} = \{y_1, \ldots, y_M\}$, where M is the number of labels and each $y_i \in \{+1, -1\}$ ($1 \le i \le M$) indicates whether the corresponding semantic concept occurs (+1) or not (−1). We use $U(\boldsymbol{x})$ to denote the unlabeled part of label set $\boldsymbol{y}$ for sample $\boldsymbol{x}$, and $L(\boldsymbol{x})$ the labeled part.

Based on the *multi-label Bayesian error bound* [6], when the sample $\boldsymbol{x}$ is represented with a single view, the sample-label pair for annotation is selected according to:

$$(\boldsymbol{x}_s^*, y_s^*) = \arg\max_{\boldsymbol{x}_s \in X, y_s \in U(\boldsymbol{x}_s)} \sum_{i=1}^{M} MI\left(y_i; y_s \middle| L(\boldsymbol{x}_s), \boldsymbol{x}_s\right). \quad (1)$$

This selection strategy reflects the uncertainty along both the sample and label dimension. Detailed discussion can be found in [6].

In the multi-view setting, each sample is split into $\boldsymbol{x} = \boldsymbol{x}^{(1)} \cup \ldots \cup \boldsymbol{x}^{(V)}$, where V is the number of different views. For each view v ($1 \le v \le V$), we can define the *intra-view uncertainty* to measure the sample and label uncertainties directly according to (1):

$$UC_{intra}(\boldsymbol{x}_s^{(v)}, y_s) = \sum_{i=1}^{M} MI\left(y_i; y_s \middle| L(\boldsymbol{x}_s^{(v)}), \boldsymbol{x}_s^{(v)}\right). \quad (2)$$

Based on (2), each view can determine the most informative sample-label pairs for annotation separately. However, as illustrated in Figure 1, when a sample-label pair is labeled, it is labeled simultaneously for all the views. As a result, given the limited number of sample-label pairs to be labeled, we should make sure that the selected sample-label pairs are informative for all the views. The intra-view uncertainty of a sample-label pair for all views can be defined as the minimum:

$$UC_{intra}(\boldsymbol{x}_s, y_s) = \min_{1 \le v \le V} UC_{intra}(\boldsymbol{x}_s^{(v)}, y_s), \quad (3)$$

or mean:

$$UC_{intra}(\boldsymbol{x}_s, y_s) = \frac{1}{V} \sum_{v=1}^{V} UC_{intra}(\boldsymbol{x}_s^{(v)}, y_s), \quad (4)$$

of the intra-view uncertainty of each single view.

Consequently, by maximizing the uncertainty in (3) or (4), we can select the sample-label pair for annotation, which corresponds to *min-max* strategy:

$$\begin{aligned}(\boldsymbol{x}_s^*, y_s^*) &= \arg\max_{\boldsymbol{x}_s \in X, y_s \in U(\boldsymbol{x}_s)} UC_{intra}(\boldsymbol{x}_s, y_s) \\ &= \arg\max_{\boldsymbol{x}_s \in X, y_s \in U(\boldsymbol{x}_s)} \left[\min_{1 \le v \le V} UC_{intra}(\boldsymbol{x}_s^{(v)}, y_s)\right],\end{aligned} \quad (5)$$

or *mean-max* strategy:

$$\begin{aligned}(\boldsymbol{x}_s^*, y_s^*) &= \arg\max_{\boldsymbol{x}_s \in X, y_s \in U(\boldsymbol{x}_s)} UC_{intra}(\boldsymbol{x}_s, y_s) \\ &= \arg\max_{\boldsymbol{x}_s \in X, y_s \in U(\boldsymbol{x}_s)} \left[\frac{1}{V} \sum_{v=1}^{V} UC_{intra}(\boldsymbol{x}_s^{(v)}, y_s)\right],\end{aligned} \quad (6)$$

respectively.

The sample-label pair selection strategy of (5) or (6) is a straightforward extension of (1) to the multi-view setting.

2.2. Inter-view Uncertainty

The intra-view uncertainty reflects the uncertainty of a sample-label pair within each view. However, it does not take into account the inherent correlation among the multiple views. Essentially, the sample-label pairs do have uncertainty over different views. In this paper, we further explore the uncertainty of a sample-label pair along the view dimension. The main idea originates from multi-view active learning, in which predictions from the multiple views are used collaboratively to obtain the multi-view uncertainty and detect the most informative data accordingly.

We will first discuss the multi-view fusion strategies for the predictions from multiple views, and then present the *inter-view uncertainty* based on the multi-view prediction.

2.2.1. Multi-view fusion

For a sample-label pair, prediction is made from each single view. These multi-view predictions vary in both the predicted label and the confidence. Thus, we need to combine the multiple predictions optimally to obtain the overall prediction.

After the distribution $P^{(v)}(y|\boldsymbol{x}^{(v)})$ has been trained for each view v (v = 1,…, V). The prediction function (or classifier) for view v can be simply computed as:

$$F^{(v)}(\boldsymbol{x}, y) = P^{(v)}\left(y = 1 \middle| \boldsymbol{x}^{(v)}\right) - P^{(v)}\left(y = -1 \middle| \boldsymbol{x}^{(v)}\right). \quad (7)$$

Then the prediction for a sample-label pair ($\boldsymbol{x}_s$, y_s) in view v can be made as:

$$\hat{y}_s^{(v)} = \operatorname{sgn}\left(F^{(v)}(\boldsymbol{x}_s, y_s)\right), \quad (8)$$

which is equivalent to adopting the label with larger posterior probability.

Given the predictions from multiple views, various fusion strategies can be adopted, among which *weighted-sum* is the most popular one:

$$F(\boldsymbol{x}, y) = \sum_{v=1}^{V} \mu_v F^{(v)}(\boldsymbol{x}, y). \quad (9)$$

where μ_v is the weight for the classifier of view v. The weighted-sum strategy is a linear fusion with respect to the multiple classifiers. As discussed in [10], the weighted-sum fusion is effective for linear-models. However, it will fail to capture the interrelations of the more complex non-linear models.

In this paper, we adopt the *super-kernel fusion* [10] to obtain the overall prediction function:

After we have trained V models for V views, we create a new training set Z for multi-view fusion. We pass each training sample-label pair ($\boldsymbol{x}_t$, y_t) to each of the V models, and obtain a V-dimensional new training feature vector $\boldsymbol{z}$:

$$\boldsymbol{z} = \left[F^{(1)}(\boldsymbol{x}_t, y_t) \ \ldots \ F^{(V)}(\boldsymbol{x}_t, y_t)\right]^T. \quad (10)$$

As a result, we obtain a training set Z consisting of N new training instances, where N is the number of sample-label pairs in the original training set.

Then, we train a super-classifier out of the new training set Z for each corresponding label, which is a traditional single-label learning problem and can be solved by many single-label data classification algorithms. In this paper, we employ SVM to train the super-classifier for its effectiveness. The kernel function and the corresponding parameters are carefully chosen via cross validation.

Finally, the overall prediction function F of a sample-label pair can be written as the fusion of the multi-view prediction models:

$$F(\boldsymbol{x}, y) = f\left(F^{(1)}(\boldsymbol{x}, y), \ldots, F^{(V)}(\boldsymbol{x}, y)\right). \quad (11)$$

2.2.2. Inter-view uncertainty

Once the overall prediction function of all the views is obtained, the prediction for a sample-label pair ($\boldsymbol{x}_s$, y_s) can be defined as:

$$\hat{y}_s = \operatorname{sgn}\left(F(\boldsymbol{x}_s, y_s)\right). \quad (12)$$

The larger the absolute value of F, the more confident it is on the prediction.

From multi-view active learning point of view, the sample-label pair with lower prediction confidence is more uncertain, and thus more informative. Therefore, we define the inter-view uncertainty as follows:

$$UC_{inter}(\boldsymbol{x}_s, y_s) = \frac{1}{\operatorname{abs}\left(F(\boldsymbol{x}_s, y_s)\right) + \varepsilon}, \quad (13)$$

where ε is a small number preventing divide by zero.

Compared with the intra-view uncertainty that relies on all possible labels of y_s, the inter-view uncertainty is closely related to the specific prediction of y_s, and thus provides additional information for sample-label pair selection.

2.3. Overall Uncertainty

The intra-view uncertainty represents the uncertainty of a sample-label pair along the sample and label dimensions, while the inter-view uncertainty reflects the uncertainty along the view dimension. In order to measure the uncertainty along all the three dimensions, we define the overall uncertainty as:

$$UC(\boldsymbol{x}_s, y_s) = UC_{intra}(\boldsymbol{x}_s, y_s) + \lambda UC_{inter}(\boldsymbol{x}_s, y_s). \quad (14)$$

where λ is a tuning parameter to balance the intra-view and inter-view uncertainty.

Consequently, the most informative sample-label pair for annotation can be selected according to:

$$(\boldsymbol{x}_s^*, y_s^*) = \underset{\boldsymbol{x}_s \in X, y_s \in U(\boldsymbol{x}_s)}{\arg\max} \ UC(\boldsymbol{x}_s, y_s). \quad (15)$$

It is worth noting that although the selection strategy is based on the uncertainty along three dimensions, we still select sample-label pairs within the sample-label dimension. The view dimension serves as a complimentary guidance for selecting informative sample-label pairs. As a result, the proposed strategy can be seen as a multi-view 2DAL strategy. When the multiple views are treated as an entire view, (15) is reduced to (1), from which we can see 2DAL is a special single-view case of the proposed strategy.

3. EXPERIMENTS

In this section, we evaluate the proposed strategy on a real-world image database, which contains 2000 images obtained from the web and Coral CDs. Six classes are defined based on high-level semantics including beach, sunset, mountain, urban, field and indoor. Each image belongs to one or more of the six classes. There are totally 2000×6 = 12000 sample-label pairs in the multi-label database.

We employ color and texture features to represent the images. The color features consist of 125-dimensional color histogram and 6-dimensional color moment (mean and variance) in RGB color space. The texture features are extracted using 3-level discrete wavelet transformation, and the mean and variance averaging on each of 10 sub-bands form a 20-dimensional vector.

The following four strategies are compared in our experiments:

(s1) The proposed multi-view multi-label active learning (multi-view 2DAL) strategy, in which we treat the color and texture features as two views, and select informative sample-label pairs according to (15).

(s2) Two-dimensional active learning (2DAL) strategy, in which we treat the color and texture features as an entire feature, and select informative sample-label pairs according to (1).

(s3) One-dimensional active learning (1DAL) strategy, in which we only take into account the uncertainty along the sample dimension, and select informative samples rather than sample-label pairs.

(s4) Non-active learning (random) strategy, in which the sample-label pairs are selected randomly.

We use 200 images (with all the 6 labels annotated) as the initial training set. In each iteration, we select the same number of sample-label pairs for annotation, i.e. 60 sample-label pairs or equivalently 10 images for 1DAL.

The average $F1$ score over the 6 labels is computed to evaluate the classification performance: $F1 = 2pr/(p+r)$, where p and r are precision and recall respectively.

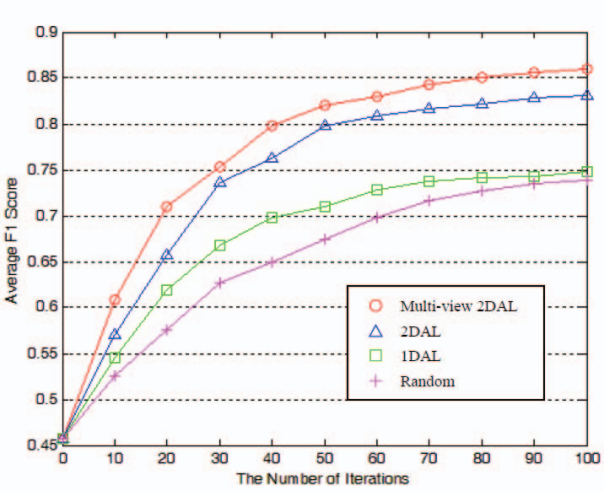

Figure 2. The image classification performance of four strategies.

Figure 2 illustrates the classification performance of the four strategies. As shown in Figure 2, all the three active learning strategies (s1, s2, and s3) outperform the non-active learning strategy (s4), which demonstrates the effectiveness of active learning for image classification. Among active learning, the two-dimensional selection of sample-label pairs (s1 and s2) achieves better performance over the one-dimensional selection of samples (s3), which proves the existence of redundancy along the label dimension. Of all the strategies, the proposed strategy (s1) has the best performance for all iterations, which indicates that the integration of multi-view learning and active learning can further improve the classification performance.

4. CONCLUSION

In this paper, we have proposed a multi-view multi-label active learning scheme, which integrates the mechanism of multi-view learning and active learning for multi-label image classification. We explore the sample and label uncertainties within each view, and meanwhile capture the uncertainty over different views based on multi-view fusion. The most informative sample-label pairs are consequently selected by maximizing the overall uncertainty along the sample, label and view dimensions. Experiments on the real-world multi-label image dataset have demonstrated the effectiveness of the proposed scheme compared with other learning strategies.

5. ACKNOWLEDGMENT

This work is supported by the State Key Program of National Natural Science of China (Grant No. 60835002), and the National Natural Science Foundation of China (Grant No. 60605004).

6. REFERENCES

[1] R. Yan, J. Yang, and A. Hauptmann, "Automatically labeling video data using multi-class active learning", in *IEEE International Conference on Computer Vision*, pp. 516–523, 2003.

[2] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification", *Pattern Recognition*, Elsevier, 37(9), pp. 1757–1771, 2004.

[3] S. Tong and E. Chang, "Support vector machine active learning for image retrieval", in *ACM International Conference on Multimedia*, pp. 107–118, 2001.

[4] X. Li, L. Wang, and E. Sung, "Multi-label SVM active learning for image classification", in *IEEE International Conference on Image Processing*, pp. 2207–2210, 2004.

[5] K. Brinker, "On active learning in multi-label classification", *From Data and Information Analysis to Knowledge Engineering*, Springer, pp. 206–213, 2006.

[6] G.J. Qi, X.S. Hua, Y. Rui, J. Tang, and H.J. Zhang, "Two-dimensional active learning for image classification", in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[7] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training", in *Annual Conference on Computational Learning Theory*, pp. 92–100, 1998.

[8] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training", in *International Conference on Information and Knowledge Management*, pp. 86–93, 2000.

[9] I. Muslea, S. Minton, and C.A. Knoblock, "Selective sampling with redundant views", in *National Conference on Artificial Intelligence*, pp. 621–626, 2000.

[10] Y. Wu, E.Y. Chang, K.C.C. Chang, and J.R. Smith, "Optimal multimodal fusion for multimedia data analysis", in *ACM International Conference on Multimedia*, pp. 572–579, 2004.