Ensemble of *k*-labelset classifiers for multi-label image classification

Dapeng Zhang[1], Xi Liu[2]

[1] Institute of Information Science and Engineering, Yanshan University,
Qinhuangdao, 066004, China
daniao@ysu.edu.cn
[2] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing, 100190, China
Liux@ics.ict.ac.cn

Abstract. In the real world, images always have several visual objects instead of only one, which makes it difficult for traditional object recognition methods to deal with them. In this paper, we propose an ensemble method for multi-label image classification. First, we construct an ensemble of *k*-labelset classifiers. A voting technique is then employed to make predictions for images based on the created ensemble of *k*-labelset classifiers. We evaluate our method on Corel dataset and demonstrate the precision, recall and F_1 measure superior to the state-of-the-art methods.

Keywords: Ensemble learning, *k*-labelset classifier, multi-label, image classification, voting

1 Introduction

The images in real world are usually associated with multiple labels. With regard to their semantic understanding, each image will be assigned with one or more labels from a predefined label set. However, most of traditional image classification methods are concerned with learning from a set of images with only one label. It is hard for them to directly handle with the multi-label learning problem. Multi-label image classification is still a challenging research issue.

Many methods have been developed to solve multi-label image classification. In general, they transform the multi-label classification task to a set of independent two-class classification problems. Early work by Boutell et al. [1] built individual binary classifiers for each label to perform multi-label scene classification. The labels of an image are determined by the outputs of these individual classifiers. The solution is theoretically simple and intuitive but it ignores label correlations that exist in the images.

To exploit these correlations, researchers have made modifications to feature representations or existing discriminative methods. Godbole et al. [2] presented a new feature representation by extending the original features with relationships between classes. The new heterogeneous features were used to train a new SVM ensemble. Qi

et al. [3] simultaneously classified labels and modeled the correlations between them by using a unified Correlative Multi-Label Support Vector Machine. In [4], a Max-Margin factorization model was created to learn label classifiers. The regularization term in the model forced label classifiers to share a low dimensional representation, which enabled the algorithm to use correlation between labels for the label prediction task. In multi-label learning, one fundamental method of using label correlations is the label combination method, or label powerset method. The basis of this method is to consider each different subset of labels as a single label to form a single-label binary classifier. We call this classifier as *k*-labelset classifier if the number of the labels in the subset is *k*. Although the label powerset method suffers from high time complexity and few training examples, it is simple and directly takes into account label correlations.

This paper proposes a novel approach to multi-label image classification. It constructs an ensemble of *k*-labelset classifiers and gives the final labels by voting. To avoid label combination complexity and few examples for some *k*-labelsets, we abandon those *k*-labelset classifiers whose training samples are below a specified threshold. Besides, the images are always associated with three labels so we only consider $k = 1$, 2, 3 labelset. For the 1-labelset classifiers which focus on only one label, we use local features such as SIFT while for the other *k*-labelset classifiers, we use global features such as GIST. The final ensemble combination is accomplished by summing the votes of all the *k*-labelset classifiers for each label. Thresholding all the label votes gives the classified labels. Our approach is experimentally tested on the Corel datasets.

2 Ensemble of *k*-labelset classifiers

In this section, we first describe the construction of *k*-labelset binary classifiers and then detail how to make ensemble combination of the generated *k*-labelset classifiers for multi-label image classification. The whole procedure is illustrated in Fig. 1.

Given a collection of training images x_i, $i = 1, \ldots, n$, each example x_i is represented as a vector of *d* dimensions and is annotated by a subset of label set *L*. The image dataset is therefore represented as $(x_1,Y_1), \ldots, (x_n,Y_n)$. The aim is to train a classifier that can predict a subset of labels for each image.

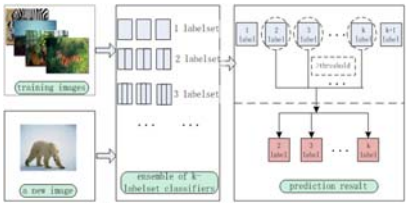

Fig. 1. The workflow of our approach

2.1 *k*-labelset classifiers

Let the image label set $L= \{l_i\}$, i=1, …, $|L|$. A set Y_i with $|Y_i| = k$ is called k-labelset. We use the term L^k to denote the set of all distinct k-labelsets on L and the size of L^k is given by the binomial coefficient: $| L^k | = \binom{|L|}{k}$. For each distinct k-labelset, we regard it as a single label. The number of class values for the single label will be 2^k. To ease the computational complexity, in this paper we only consider the class in which all k labels appear in an image. Therefore the images which contain all the k labels are considered as positive images while the other images are considered negative for this single label. A binary classifier can then be created accordingly and we call it k-labelset classifier.

We iteratively construct an ensemble of k-labelset classifiers. Since most of the images are often possessed with no more than 3 labels, we only create 1-labelset, 2-labelset and 3-labelset classifiers. For k-labelset classifiers (k =1, 2, 3), the algorithm selects a k-labelset Y_i from L^k at each iteration. If the number of the training images which have the selected k-labelset is above the specified threshold, a k-labelset classifier for the k-labelset will be created. Finally, we can obtain an ensemble of k-labelset classifiers and their corresponding k-labelsets. The pseudocode is given in Algorithm 1.

Algorithm 1: GenEnsembleofK-labelsetClassifier

1. Input: training set D = $\{(x_i, Y_i), \ldots ,(x_n, Y_n)\}$, set of labels L, minimum training sample number T_n.
2. Output: an ensemble of k-labelset classifiers $h_{k,i}$ and corresponding to k-labelsets Y_i
3. For k =1, 2, 3 do
4. $R \leftarrow L^k$
5. For i = 1 to $|L^k|$ do
6. $Y_i \leftarrow$ a k-labelset selected from R
7. If the number of training images which contain Y_i is larger than T_n
8. Train a k-labelset classifier $h_{k,i}$: $x \rightarrow P(Y_i) \in \{0,1\}$
9. End
10. $R \leftarrow R \setminus \{ Y_i \}$
11. End

The minimum training sample number T_n is a user specified parameter. It is set to be a fixed number such as 20 or a fixed ratio of the total training samples. This can avoid the cases in which there exist few samples for some k-labelset classifiers and also many k-labelset classifiers will thus not be created, which greatly reduce the computation. The few training number of some k-labelsets also means that the k labels

have little correlation, and we can hypothesize that our approach will manage to model label correlations by using the minimum training sample threshold. It is also easily seen that we will suffer imbalanced data problem during the training of a *k*-labelset classifier because of the relatively small number of positive images for the *k*-labelset. To tackle with this, the negative images for the *k*-labelset will be sampled by a rate (i.e. 1:2) so as to train a good *k*-labelset classifier. Furthermore, for 1-labelset classifier, we will use the local image features and for 2-labelset and 3-labelset classifiers, we will use the global images features because 1-labelset classifiers focus on one local label while 2 or 3-labeset classifiers consider 2 or 3 labels as a whole.

2.2 Ensemble voting

The labels of an image will be predicted based on the voting outputs of the obtained ensemble of the *k*-labelset classifiers. Given a new image, each *k*-labelset classifier $h_{k,i}$ provides binary decisions for each label in the corresponding *k*-labelset Y_i. Sum the binary decisions of all the *k*-labelset classifiers for each label and then calculate the average decision. A label will be assigned to the image if the label's average decision is larger than a user-specified threshold t. Also we can directly give top n (i.e. 3) labels with high average precision for the image. Algorithm 2 illustrates the detailed procedure of the ensemble voting prediction. The *mk* in the algorithm represents the number of *k*-labelset classifiers.

Algorithm 2: EnsembleVoting_Prediction

1. Input:a new image x, an ensemble of *k*-labelset classifiers and the corresponding set of *k*-labelsets, set of labels L, the threshold for label prediction t.
2. Output: multi-label image classification vector Result
3. Initialize: Votesj=0, Sumj=0, j=1, …, $|L|$
4. For k= 1 to 3 do
5. For i = 1 to mk do
6. For each label l_j in the corresponding *k*-labelset of $h_{k,i}$
7. Votesj = Votesj + $h_{k,i}$ (x, l_j);
8. Sumj = Sumj + 1;
9. End
10. End
11. End
12. For j=1 to $|L|$ do
13 Avgj = Sumj / Votesj;
14 if Avgj > t then
15 Resultj = 1;

16 else Resultj = 0;

17. End

3 Experiment

3.1 Experiment Setup

In this work, we use bag-of-local feature descriptors as image representations for 1-labelset classifiers and use a global image feature gist for 2, 3-labelset classifiers. In the BoW framework, a set of local feature descriptors are extracted from each image. All the local descriptors are then clustered and the prototype of each cluster is treated as a "visual word". Assigning each local feature to its nearest visual word and counting the occurrence number, we can represent an image by a histogram of visual words. The cluster number is set 500 in the experiment. For each image, 12*12 pixel local patches over a grid with spacing of 6 pixels are extracted and the local patches are described by a 200-dimensional texton histogram descriptor [5], which encodes both texture and color information. As for image gist [6], it is a holistic image representation that describes global structure of an entire image. We compute the gist by use of Oliva and Torralba's implemention [7]. Linear kernel based support vector machines (SVM) are employed to train the *k*-labelset classifiers and we will use SVMlight [8] to implement these SVMs.

We use the example-based multi-label classification evaluation measures in [9] as our experimental evaluation metric. Let D be a multi-label image test set, which consists of $|D|$ multi-label images (x_i, Y_i), $i = 1, \ldots, |D|$, $Y_i \subseteq L$. Y_i is the true set of image labels for image i in D and let Z_i be the predicted set of labels for image i. The precision, recall and F_1 measures are calculated respectively as follows:

$$Precision(D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap Z_i|}{|Z_i|} \tag{1}$$

$$Recall(D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap Z_i|}{|Y_i|} \tag{2}$$

$$F_1(D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{2|Y_i \cap Z_i|}{|Z_i| + |Y_i|} \tag{3}$$

3.2 Evaluation on Corel dataset

The Corel dataset [10] has been extensively used as a standard benchmark dataset for annotation prediction tasks and we will evaluate our multi-label image classification approach on it. In the dataset, there are altogether 5000 images in 50 different sets (CDs). Its vocabulary size is 374 and all the images are associated with 1~5 labels. We will conduct the experiments on a randomly selected subset of the dataset, which contains 25 labels such as "bear", "forest", "mountain", and "sky" and about 4000

images. 80 percent of the new image set is kept as training set, and the left images are made as test set.

T_n in Alg.1 affects our method a lot. Too large will lead to a small number of *k*-labelset classifiers while too small will cause the few samples problem. We experiment with several t_n (15, 20, 25, 30, 35), the number of *k*-labelset classifiers for each t_n is shown in Table 1.

Table 1. The number of *k*-labelset classifiers for different T_n

t_n	15	20	25	30	35
*m*1	25	25	25	25	25
*m*2	89	83	71	60	53
*m*3	78	63	48	33	25
avgL	17.48	15.20	12.44	9.76	8.24

The 2-labelset classifiers provide classification for 2 labels and the 3-labelset classifiers provide classification for 3 labels. We define the metric $avgL = (m1+2*m2+3*m3)/|L|$, $|L|=25$, which measures the average number of classifiers for each label. From Table 1, we can see $t_n = 25$ is a reasonable value under which *avgL* is adequate for ensemble voting for each label, and the number of 2,3-labelset classifiers is large enough for capturing the label correlations and also the 25 least number of training sample make the *k*-labelset classifiers sufficient for training.

The different value of the threshold t in Alg.2 will lead to different precision, recall and F_1 measure. Note that when calculating these measures the ground-truth labels for test images are confined to the 25 labels. Table 2 gives the results of $t = 0.35$, 0.40, and 0.45 respectively. From the perspective of F_1 measure, $t = 0.40$ presents the best results.

Table 2. Comparison of the performances of our method with other approaches

Method	Precision	Recall	F_1-measure
Trans[10]	0.06	0.04	0.05
CRM[11]	0.16	0.19	0.17
Independent SVMs[4]	0.22	0.25	0.23
Ours, t=0.35	0.49	0.57	0.527
Ours, t=0.40	0.52	0.54	0.530
Ours, t=0.45	0.56	0.43	0.486

To further demonstrate the effectiveness of our method, we compare it with several word annotation prediction models including Translation Model (Trans) [10], Continuous Relevance Model (CRM) [11], and independent SVMs [4] on the PicSOM features. Although these models use all 4500 training examples, we think that they are partly comparative. As shown in Table 2, our approach gives the best performance.

Fig.2. shows four images which are labeled by our method. The predicted labels are black bold under each image while the ground-true labels are shown in blue italic font. The results are quite satisfying on the whole.

Fig. 2. Some examples of the labeling results

4 Conclusions and future work

This paper presents a novel ensemble method for multi-label image classification. Through a simple ensemble voting of a set of *k*-labelset classifiers, our method can predict labels for any image. The measures evaluated on Corel dataset demonstrate the effectiveness of the method. The use of 1,2,3-labelset classifiers makes it natural to capture label correlations and that the *k*-labelset classifiers with few training images are discarded ensures the computational efficiency. For the future, we intend to consider using unlabeled or partly labeled images to improve the performance since fully labeled images are insufficient and hard to obtain. One possible means is to extend the *k*-labelset classifiers to semi-supervised learning classifiers.

References

1. M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," In Pattern Recognition, 37:1757-1771, 2004.
2. S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," In Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2004.
3. G. J. Qi, X. S. Hua, Y. Rui, J. Tang, T. Mei, and H. J. Zhang, "Correlative multi-label video annotation," In ACM International Conference on Multimedia, pp. 17-26, 2007.
4. Nicolas Loeff and Ali Farhadi, "Scene discovery by matrix factorization," In Proc. ECCV, pp. 451-464, 2008.
5. T. Li and I. S. Kweon, "A semantic region descriptor for local feature based image classification," In ICASSP, 2008.
6. A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," In IJCV, 42(3):145-175,2001.
7. A. Oliva and A. Torralba. Spatial envelop, http://people.csail. mit.edu/torralba/code/spatialenvelop/, 2001.
8. T. Joachims. SVMlight, "Learning to classify text using support vector machines - methods, theory, and algorithms," Kluwer, Dordrecht, the Netherlands, 2002.
9. Grigorios Tsoumakas and Ioannis Vlahavas, "Random k-labelsets: An ensemble Method for Multilabel classification," In ECML, pp. 406-417, 2007.

10. P. Duygulu, K. Barnard, J. d. Freitas, and D. Forsyth, “Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary,” In ECCV, pp. 97-112, 2002.
11. V. Lavrenko, R. Manmatha, J. Jeon, “A model for learning the semantics of pictures,” In NIPS, 2003.