Multiview Semi-Supervised Ranking for Automatic Image Annotation

Ali Fakeri-Tabrizi[*] Massih-Reza Amini[†] Patrick Gallinari[*]

[*]Université Pierre et Marie Curie
Laboratoire d'Informatique de Paris 6
4, place jussieu
75252 Paris, France
First.Last@lip6.fr

[†]Université Joseph Fourier
Laboratoire d'Informatique de Grenoble
Centre Equation 4
38041 Grenoble, France
Massih-Reza.Amini@imag.fr

ABSTRACT

Most photo sharing sites give their users the opportunity to manually label images. The labels collected that way are usually very incomplete due to the size of the image collections: most images are not labeled according to all the categories they belong to, and, conversely, many class have relatively few representative examples. Automated image systems that can deal with small amounts of labeled examples and unbalanced classes are thus necessary to better organize and annotate images. In this work, we propose a multiview semi-supervised bipartite ranking model which allows to leverage the information contained in unlabeled sets of images in order to improve the prediction performance, using multiple descriptions, or views of images. For each topic class, our approach first learns as many view-specific rankers as available views using the labeled data only. These rankers are then improved iteratively by adding pseudo-labeled pairs of examples on which all view-specific rankers agree over the ranking of examples within these pairs. We report on experiments carried out on the NUS-WIDE dataset, which show that the multiview ranking process improves predictive performances when a small number of labeled examples is available specially for unbalanced classes. We show also that our approach achieves significant improvements over a state-of-the art semi-supervised multiview classification model.

Categories and Subject Descriptors

I.2 [**Artificial intelligence**]: Learning; I.4 [**Image processing and computer vision**]: General—*Image processing software*

General Terms

Theory, Algorithms, Experimentation

Keywords

Multiview semi-supervised ranking, Image Annotation

MM'13, October 21–25, 2013, Barcelona, Spain.
Copyright 2013 ACM 978-1-4503-2404-5/13/10 ...$15.00..

1. INTRODUCTION

Image annotation is the task of assigning labels to images. Each label describes an entity, a content or even a part of a given information contained in an image. In some photo sharing sites, users have the opportunity to manually assign labels to images; these labels are further used to browse or search the image collection. Labeling a huge collection of images is usually a difficult and a time consuming task, and the automatisation of image annotation has motivated the development of many different methods in the past years.

The intrinsic properties of image collections that have to be considered when developing automatic annotation tools are (a) in many categories, there are generally a very limited number of labeled examples, together with a very large number of additional unlabeled images, (b) images are naturally represented by several distinct features sets, or *views*, such as RGB color histograms or *textual* features; such as the bag-of-words representation of the surrounding texts of images, and, (c) classes are imbalanced such that the number of examples in one class could be very few, or in the opposite dominates the other classes. Recently, multi-view semi-supervised learning techniques for classification have been developed to address this kind of problem [11]. Some approaches use the labeled data to train several view-specific classifiers, and rely on the intuition that these predictors should have similar predictions on the unlabeled set in order to assign pseudo-labels to these examples and to train new view-specific classifiers using the initial labeled training set and the newly pseudo-labeled unlabeled examples [1].

In this work we propose a new multiview bi-partite ranking framework for this task. Our approach to semi-supervised multiview ranking follows the same intuition than before. Given V score functions $(h_1, \ldots, h_V)$ independently trained on each view, we define a notion of global agreement between them as the expectation, over random pairs of images $(\mathbf{x}, \mathbf{x}')$, that two score functions $(h_v, h_{v'})$ predict the same relative ordering. We hence describe a learning process in which view-specific ranking functions should achieve high ranking performance on the labeled training set, while minimizing a disagreement measure between each other on the unlabeled dataset. The overall motivation of using a bi-partite ranking framework rather than classification one here, is that it has recently been shown that in the supervised case, the associated classifier of a ranking function can better handle class imbalancements [5]. Our experimental results show that our approach achieves significant improvements over a state-of-the art semi-supervised multiview classification model.

2. SEMI-SUPERVISED MULTIVIEW RANKING FOR IMAGE ANNOTATIONS

We suppose that we have n labeled examples $\mathcal{Z} = (\mathbf{x}^i, y^i)_{i=1}^n$ composed of pairs of (observation,target), where $y^i \in \{-1, 1\}$ is the relevance associated to the observation x^i. In a semi-supervised learning setting we also assume to have access to a set of unlabeled examples $\mathcal{U} = \{(\mathbf{x}^{n+i}) | i = 1, \ldots, m\}$. In the single view setting, the aim of learning is to find a scoring function h, so that relevant observations (i.e. those with $y = +1$) are assigned higher scores than irrelevant ones, and the ranking criterion to be optimized to achieve this goal is usually the Area Under the ROC Curve (`AUC`).

In the multiview setting, an observation (e.g. an image described by different feature sets), $\mathbf{x} = (x_1, ..., x_V)$ is represented in several vector spaces, $\mathcal{X}_v, v \in \{1 \ldots V\}$, such that each vector x_v (an image descriptor) can be used to build a predictor. Following the framework of [9] for multiview classification or regression, we can define the objective of multiview ranking as jointly learning *view-specific* scoring functions $h_v : \mathcal{X}_v \to \mathbb{R}$ (in our case, h_v only considers the v-th descriptor of images) so that their average risk is small, where the joint learning of these view-specific predictors consists in constraining them to agree with each other (i.e. have similar predictions). In this framework, the agreement between predictors is measured without knowing the labels of the observations, and therefore it can be estimated and optimized from the pool of unlabeled data. Since constraining the view-specific predictors to have a low disagreement reduces the function space, we can expect better generalization guarantees using semi-supervised multiview learning than using supervised learning.

A measure of disagreement, D, between two scoring functions h_v and $h_{v'}$ can be defines as the probability, over any two random observations, that they do not predict the same ordering:

$$D(h_v, h_{v'}) = \mathbb{P}\big((h_v(X) - h_v(X'))(h_{v'}(X) - h_{v'}(X')) \leq 0\big)$$

which can be estimated on the *unlabeled* data set U by $\widehat{D}_U(h_v, h_{v'})$ defined as :

$$\frac{1}{m(m-1)} \sum_{i \neq j} \mathbb{1}_{(h_v(x_v^{n+i}) - h_v(x_v^{n+j})) \times (h_{v'}(x_v^{n+i}) - h_{v'}(x_v^{n+j})) \leq 0}$$

Where, $\mathbb{1}_.$ is the indicator function. This notion of disagreement and its empirical counterpart can then be extended to more than two views by taking the average disagreement between scoring functions for any pair of views:

$$\widehat{D}_U(h_1, \ldots, h_V) = \frac{2}{V(V-1)} \sum_{v < v'} \widehat{D}_U(h_v, h_{v'}) \ . \qquad (1)$$

To avoid the costly computation of the disagreement at each iteration for all unlabeled data, we consider an estimation to use the pairs sampled at the current iteration. Our semi-supervised multiview process follows existing iterative pseudo-labeling methods for classification, but it has two properties which reduce the overall time and space complexity to $\tilde{O}(n + m)$. The first one is a reduction from the pseudo-labeled pairs to bipartite ranking in order to use efficient learning to rank algorithms. The second one is a random sampling of pairs at each iteration rather than considering all possible pairs of unlabeled examples.

Following iterative pseudo-labeling methods for classification (here, applied to pairs of inputs) we form unlabeled pairs of examples and decide whether or not they contain information that should be added to the training set based on a measure of confidence of the pseudo-label. We hence select only the pairs for which all the view-specific rankers agree on the relative ordering. This requirement of unanimity may be too restrictive when there are many views, but we observed that it works very well in practice.

Algorithm 1: Semi-supervised Multiview Ranking

```
Input:
▷ supervised bipartite ranking algorithm: $\mathcal{A}$;
▷ size of the random pairs sample: $S$;
▷ labeled $\mathcal{Z}$, and unlabeled $\mathcal{U}$ multiview training data;
Initialize:
for each view, train $h_v^{(0)}$ on $Z$ with $\mathcal{A}$.
$t \leftarrow 0$;
repeat
  for $s = 1..S$ do
    $(i, j) = \text{sample}\big(\{(k, \ell) \in \{1, ..., m\}^2, k \neq \ell\}\big)$
    if $\forall v, h_v^{(t)}(x_v^{n+i}) > h_v^{(t)}(x_v^{n+j})$ then
      $Z \leftarrow Z \cup \{(x^{n+i}, +1), (x^{n+j}, -1)\}$
    else if $\forall v, h_v^{(t)}(x_v^{n+i}) < h_v^{(t)}(x_v^{n+j})$ then
      $Z \leftarrow Z \cup \{(x^{n+i}, -1), (x^{n+j}, +1)\}$
    end if
  end for
  $t \leftarrow t + 1$;
  for each view, train $h_v^{(t)}$ on $Z$ with $\mathcal{A}$;
until $\hat{D}_U\big(h_1^{(t)}, ..., h_V^{(t)}\big) \geq \hat{D}_U\big(h_1^{(t-1)}, ..., h_V^{(t-1)}\big)$
Output: $\forall v \in \{1, .., V\}, h_v^{(t)}$;
```

3. ALGORITHM

The pseudo-code of our proposed approach is given in Algorithm 1. In an initialization step, each view-specific ranker is trained independently on the labeled training set. Then, the algorithm iteratively re-trains one ranker per view on increasing training sets composed of the initial labeled examples, and additional pseudo-labeled examples. The iterative procedure is repeated until the disagreement does not decrease after re-training.

Considering our notion of empirical disagreement Eq. (1), it is then natural to define a notion of pseudo-labeling on *pairs* of unlabeled observations: a pair $(\mathbf{x}^{n+i}, \mathbf{x}^{n+j})$ would be labeled $+1$ if the various view-rankers agree on, and -1 if they agree on the inverse relative ordering. After pseudo-labeling, we would then obtain a training set with pseudo-pairwise preferences (instead of pseudo labels in $\{-1, 1\}$). From a computational point of view, however, this procedure would be extremely costly for two reasons. First, it would require a pass over all pairs of unlabeled inputs at each round. Since there are about m^2 pairs, this is too large by an order of magnitude. Secondly, the pairs of unlabeled inputs selected to be added in the training set do not have the structure of a proper bipartite ranking. The underlying supervised learning algorithm should then be an algorithm that can deal with arbitrary pairwise preferences, which have $\Omega(\ell^2)$ space and time complexity (ℓ is the number of objects in the training set). By contrast, efficient algorithms for bipartite ranking like `SVM`^{multi} [6] run in time $\tilde{O}(\ell)$ and require $O(\ell)$ space.

4. EXPERIMENTS AND RESULTS

In this work, we use the NUS-WIDE-LITE image collection which is a subset of the original image collection NUS-WIDE provided by the National University of Singapour [4]. This smaller dataset is composed of 28,807 images for training and 28,808 images for testing. The images in these two sets belong to one or more of 81 classes. This collection provides six visual features already extracted. Three of them are color based features: 1) a color histogram, 2) a color auto-correlogram and 3) block-wise color moments. Using the texture-base technique, two other feature types have been extracted: 4) an edge direction histogram and 5) a wavelet texture. The last set of visual features in this collection is 6) a bag of visual words, obtained by using the SIFT method [7]. All these features are included in the dataset distribution and are based on the visual content. In addition, we extracted a new set of features, based on textual data. We extracted these textual feature automatically from user tags associated to each image, and we took into account a bag-of-words representation with a tf.idf term weighting after filtering out rare words [8]. In [2], it has been shown that the more views are independent the more we can expect higher generalization performance. Therefore, in order to enhance the independence between different feature sets, we concatenate features within the same classes of visual descriptors resulting in four different views ($V = 4$): 1) concatenation of color-based features, 2) concatenation of texture-based features, 3): the bag of visual words and 4): textual feature. Each set of these 4 new features is considered as a *view* in our multiview approach.

To simulate a semi-supervised learning scenario, we divide the training data into two subsets: a subset will be used as the *labeled set* in which we consider that the labels are known; for the remaining training images, the labels are hidden and this subset will be used as the *unlabeled set*. The split is performed randomly, and each experiment is repeated 50 times, over 50 different subsamplings of the training data. We tried four different size for the *labeled set*: 50, 100, 200, 1000 (Resp. 28757, 28707, 28607, 27807 for *unlabeled set*); In each split, the proportions of the class are kept similar to what was observed in the training data set (stratified sampling). The test collection is left unchanged, and all reported results apply to this test collection.

To study the behavior of our model in presence of class imbalancement, we focus on 22 classes of NUS-WIDE-LITE upon which 7 classes have a positive percentage rate greater than 15% and 15 other classes are highly imbalanced having less than 15% positive examples.

In our experiments, we compared our multi view ranking strategy with 4 other strategies: three supervised learning baselines which consists in two ranking and one classification models and one semi-supervised multiview classification model proposed in [1].

- **Supervised Ranking.** A supervised view-specific bipartite ranking model trained on each view separately (`R_SUP`). Reported results for for `R_SUP` are the average results of each of the view-specific rankers.

- **Classification.** Supervised view-specific SVMs trained on each view separately (denoted as `C_SUP`). Results for `C_SUP` correspond to the average performance of the view-specific classifiers which are optimized over the F_1 measure [6].

- **Concatenation.** A supervised Ranking SVM working on the concatenation of all of the views (`R_Concat`).

- **Semi-supervised multiview classification learning.** The semi-supervised multiview classification algorithm (`C_SemSUP`) proposed in [1]. As baseline classifiers, we use SVM optimized over the F_1 measure in order to be less affected by the imbalance effects [6].

- **Semi-supervised multiview ranking learning.** Our approach (`R_SemSUP`). At test time, we sum the scores of all views-specific rankers and rerank the scores according to the the summed scores.

In all of our experiments, we used linear kernels and we set the regularization parameter of $\mathtt{SVM}^{multi}$, C, to 1. We also tried to find the parameter C by cross-validation but for small labeled training sets, we considered, this strategy leads easily to overfitting. The performance is reported in terms of Accuracy and AUC of the view-specific scoring functions over the four views mentioned above and different classes.

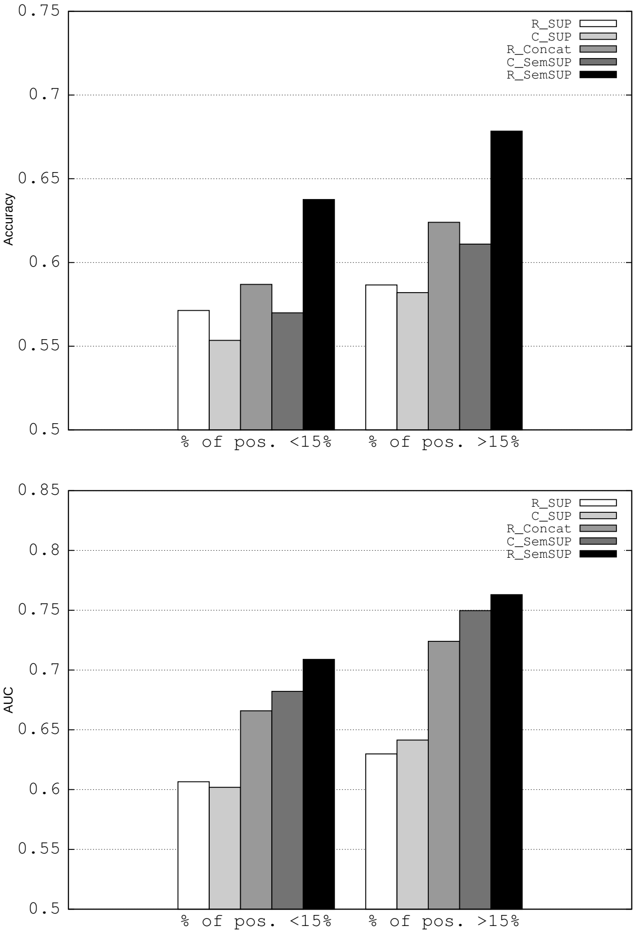

Figure 1: Accuracy (top) and AUC (bottom) performance of all models for 50 labeled examples and two cases where the percentage of positive class is below (left in each figure) and above (right in each figure) 15%.

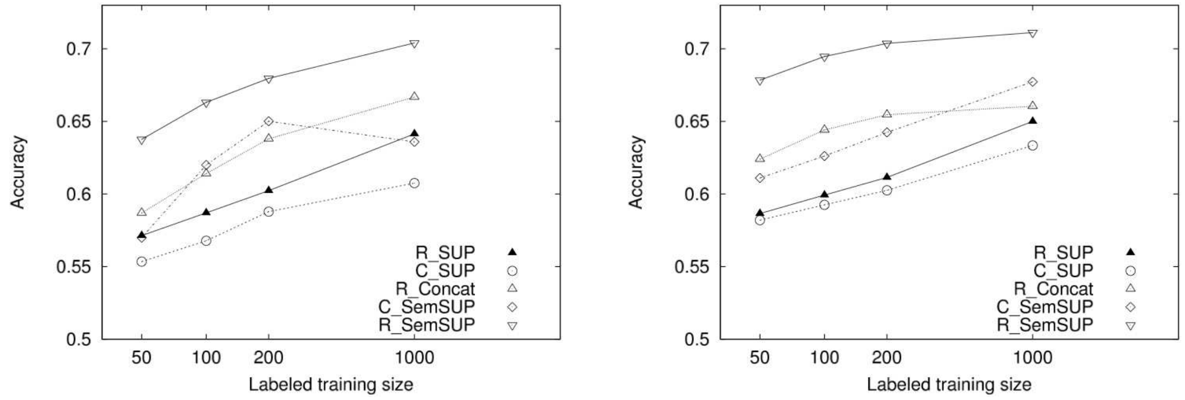

Figure 2: Accuracy of different strategies with respect to four different labeled training size and when the percentage of positive class is below (left) and above (right) 15%, on the NUS-WIDE-LITE dataset. Performance are averaged over 22 classes and 50 random splits.

The histograms in figure 1 illustrate the average of Accuracy and AUC measures over all classes for different strategies. We separated the performance for the cases where the percentage of positives was below or above 15%, in order to see the impact of this imbalancement over the classification and ranking strategies. We see that `R_SemSUP` clearly outperforms the three supervised baselines `R_SUP`, `C_SUP` and `R_Concat`, as well as the semi-supervised classification strategy `C_SemSUP`, specially when the imbalancement is more pronounced. These results also suggest that supervised bi-partite ranking over the concatenation of all of the views may sometimes outperform a semi-supervised multiview classification strategy which optimizes a F_1 value.

Figure 2, shows the accuracy performance of five strategies with respect to different labaeled training size and when the percentage of positive class is below and above 15%. These results are averaged over 50 random labeled/unlabeled splits of the training data and all the classes. From these results, we see that when the percentage of the positive class is below 15%, `C_SemSUP` overcomes in some extent the class imbalancement, one the case where the number of labeled examples increase, but not always, while `R_SemSUP` is not so affected by class imbalancement and that for every number of labeled examples we considered in our experiments.

5. CONCLUSION

In this paper, we proposed a mutiview semi-supervised ranking algorithm for image annotation task in the general case where images are naturally described with heterogeneous visual feature sets, as well as multimodal (visual/text) feature representations, and where classes are mostly imbalanced. The proposed algorithm takes advantage of the existing different views of examples in order to overcome the lack of labeled examples by reducing the disagreement of different view-specific rankers on the unlabaled training set. Experiments conducted on the NUS-WIDE dataset show that our strategy can better tackle the class imbalancement problem than a state-of-the-art multiview semi-supervised classification method, in the most interesting case where the number of labeled training examples is low. Finally, the approach is generic in the sense that it can be applied to many other multi-modal image applications [3, 10].

6. REFERENCES

[1] M. R. Amini, N. Usunier, and C. Goutte. Learning from multiple partially observed views - an application to multilingual text categorization. In *Neural Information Processing Systems (NIPS)*, pages 28–36, 2009.

[2] A. Blum and T. M. Mitchell. Combining labeled and unlabeled data with co-training. In *Conference on Learning Theory (COLT)*, pages 92–100, 1998.

[3] X. Cai, F. Nie, H. Huang, and F. Kamangar. Heterogeneous image feature integration via multi-modal spectral clustering. In *CVPR*, 2011.

[4] T. S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Z. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Conference On Image And Video Retrieva (CIVR)*, 2009.

[5] A. Fakeri-Tabrizi, S. Tollari, N. Usunier, and P. Gallinari. Improving image annotation in imbalanced classification problems with ranking svm. In *CLEF*, pages 291–294, 2010.

[6] T. Joachims. A support vector method for multivariate measures. In *International Conference on Machine Learning (ICML)*, 2005.

[7] D. G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Visio (ICCV)*, pages 1150–1157, 1999.

[8] G. Salton. *Introduction to modern information retrieval.* McGraw-Hill, 1986.

[9] K. Sridharan and S. M. Kakade. An information theoretic framework for multiview learning. In *COLT*, pages 403–414, 2008.

[10] H. Wang, F. Nie, H. Huang, S. L. Risacher, A. J. Saykin, and L. Shen. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics*, 2012.

[11] X. Zhang, J. Cheng, C. Xu, H. Lu, and S. Ma. Multi-view multi-label active learning for image classiffication. In *ICME*, pages 258–261, 2009.