# **Multi-View Learning from Imperfect Tagging**

Zhongang Qi, Ming Yang, and Zhongfei (Mark) Zhang Dept. of ISEE, Zhejiang University, China {zhongangqi, cauchym, zhongfei}@zju.edu.cn

# ABSTRACT

In many real-world applications, tagging is imperfect: incomplete, inconsistent, and error-prone. Solutions to this problem will generate societal and technical impacts. In this paper, we investigate this arguably new problem: learning from imperfect tagging. We propose a general and effective learning scheme called the Multiview Imperfect Tagging Learning (MITL) to this problem. The main idea of MITL lies in extracting the information of the imperfectly tagged training dataset from multiple views to differentiate the data points in the role of classification. Further, a novel discriminative classification method is proposed under the framework of MITL, which explicitly makes use of the given multiple labels simultaneously as an additional feature to deliver a more effective classification performance than the existing literature where one label is considered at a time as the classification target while the rest of the given labels are completely ignored at the same time. The proposed methods can not only complete the incomplete tagging but also denoise the noisy tagging through an inductive learning. We apply the general solution to the problem with a more specific context - imperfect image annotation, and evaluate the proposed methods on a standard dataset from the related literature. Experiments show that they are superior to the peer methods on solving the problem of learning from imperfect tagging in cross-media.

#### **Categories and Subject Descriptors**

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Retrieval models*; H.2.8 [Database Management]: Database Applications—*Data mining, Image databases* 

#### **General Terms**

Algorithms, experimentation

# Keywords

Imperfect tagging; multi-label space; image annotation completion, correction and prediction

# 1. INTRODUCTION

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'12, October 29-November 2, 2012, Nara, Japan.

Copyright 2012 ACM 978-1-4503-1089-5/12/10 ...\$10.00.

Zhengyou Zhang Microsoft Research One Microsoft Way, Redmond, WA, USA zhang@microsoft.com



Figure 1: (a) An example of an incompletely tagged image. (b) An example of a noisily tagged image.

Tagging, as one of the content forms contained in multimedia, demonstrates a strong power in helping develop effective solutions to many important real-world problems, especially in the crossmedia field. For example, with an appropriate tagging to images as part of the annotation, one can develop powerful image annotation and/or image retrieval techniques [15]; with an appropriate tagging to movies as part of the reviews, one can develop powerful movie recommendation systems [25]; and with an appropriate tagging to Web-pages as part of the comments, one can develop more powerful search engines [10].

However, with the typically explosive amount of data to be tagged, one can hardly do tagging manually. Consequently, social tagging has become an effective alternative to many such problems [11], also partly due to the ubiquitous access to the Internet. For social tagging, the tagging results can never be expected to be perfect. The imperfect tagging may result in either of or both of the following two possible tagging scenarios: (1) incomplete tagging, and (2) noisy tagging. Incomplete tagging means that the given tags to a document still fail to give a complete description of the content of the document up to all the details. Figure 1(a) shows such an example for the tags as the annotation to an image. The given annotations of this image are buildings, cityscape, sky, and street. However, it is clear from this image that the given annotation tagging is far away from being actually complete. There are many other things missed to be tagged, such as person, tree, water, fountain, and street lamp. Noisy tagging means that the given tags to a document contain errors; in other words, some of the tags are incorrect in describing the content of the document or even completely irrelevant to the document. Figure 1(b) shows such an example of noisy tagging. One of the given tags for the image of Figure 1(b) is *fish*, while it is obvious that the animal in the image of Figure 1(b) is bird instead of fish, and the label *fish* is given incorrectly.

In general an N-class classification problem can always be decomposed into N binary classification problems in one-vs-all (OVA) mode. Figure 2 shows examples of perfect tagging and imperfect tagging in such an OVA binary classification problem. Figure 2(a) is an example of perfect tagging where all the data points are correctly tagged. Figure 2(b) visualizes the OVA binary classification

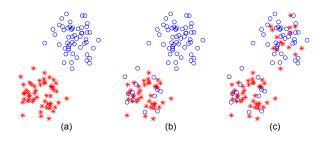


Figure 2: (a) Perfect tagging (The asterisks are tagged as positive ones; the circles are tagged as negative ones). (b) Incomplete tagging (The asterisks are tagged as positive ones; the circles are untagged ones). (c) Noisy tagging (The asterisks are tagged as positive ones; the circles are tagged as negative ones).

with incomplete tags where the asterisks are the given correctly tagged but incomplete data from the positive class, and the circles are the untagged data points possibly from both classes. Figure 2(c) is an example of the OVA binary classification with noise where both the asterisks and the circles contain the data from the positive and the negative classes, and consequently, are cross-tagged in error in the two classes. In fact, incomplete tagging can be considered as a special case of noisy tagging where all the given positive samples are correct.

It is easy to understand that the more tags the data are given, the more information the tag space contains. However, most of the discriminative methods for the multi-label classification problem only consider the multi-label space as the classification target, and fail to make use of the information contained in the multi-label space effectively. Especially in the OVA mode, these classification methods only consider one tag at a time as the classification target while at the same time completely ignore the rest of the tags. In this paper, we explicitly consider all the given tags simultaneous as an additional feature which further helps improve the classification performance. Figure 3 shows exemplar images with multiple tags. The images in the first row in Figure 3, which are all tagged as fish, always have the accompanied tags of water, coral, and ocean, while the images in the second row in Figure 3, which are all tagged as bird, always have the accompanied tags of sky, cloud, grass, and tree. Obviously, these accompanied tags can be utilized as an additional feature to help better distinguish images tagged as *fish* from images tagged as bird.

This paper addresses the arguably new problem of learning from imperfect tagging not only to complete the incomplete tagging but also to denoise the noisy tagging through an inductive learning. Even though there is recent literature to address the incomplete tagging and noisy classification separately, combining them as a single, overall problem and thus developing a synergistic solution to address them together in a single framework to this problem are still significant and challenging. On the other hand, an effective solution to this problem shall benefit not only the research community but also the society in substantially improving the quality of the techniques and technologies that rely on social tagging, generating great societal and technical impacts. Further, this paper proposes a new way to make use of the given multiple tags simultaneously as an additional feature to deliver a more effective classification in solving a general multi-label classification problem.

While it is a challenge for a classifier to learn from imperfect tagging, if a multi-view based technique is ready to apply to this problem, an effective classification framework can be constructed to take advantage of the imperfect tagging. The multiple views to be used in the proposed classification framework are similar to the various records obtained by different historians independently on the



Figure 3: Exemplar images with multiple tags.

same historical event. Though there exist overlaps in these records, the most valuable benefit lies in the non-overlap parts, which help contribute systematically to recovering the whole event or even correcting the sporadic subjectivity existing in the individual records obtained by different historians. Likewise, the classification can be completed or corrected under the multiple classifiers' results systematically in the multi-view based classification framework.

In this paper, we develop a general and synergistic solution which we call Multi-view Imperfect Tagging Learning (MITL) to the problem of learning from imperfect tagging. The main idea of this work lies in extracting the information of the imperfectly tagged training dataset from multiple views to differentiate the data points in the role of classification. Further, a novel discriminative classification method (MSS-2K) which takes advantage of the information contained in the multi-label space is proposed under the framework of MITL. We apply the general solution to the problem with a more specific context — imperfect image annotation, and demonstrate through extensive evaluations using real data that the proposed methods perform well in comparison with the peer methods as an effective and promising solution to the problem of learning from imperfect tagging in cross-media.

# 2. RELATED WORK

We begin by reviewing the literature in the incomplete tagging and noisy classification areas, and then review the related work on multi-view learning. Finally, we introduce the literature in the field of image annotation and retrieval for we apply MITL to the problem of learning from imperfect tagging with a more specific context imperfect image annotation.

Most of the existing classification methods, based on either supervised learning or semi-supervised learning, assume that the tagged training data are perfect, which means that there are no incomplete tags or noisy tags contained in the training set. There has been work of incomplete tagging learning and noisy data learning separately in the literature. Partially supervised learning [14] studies the problem where the training set contains a small set of positive examples and a large set of unlabeled examples within both positive and negative instances, as shown in Figure 2(b). In [8, 14, 17], instances which are with the highest confidence to be negative are extracted as the labeled negative examples used to train the classifier based on supervised or semi-supervised learning. Compared with the partially supervised learning of the binary classification problem, Sun et al. [24] propose the Weak Label Learning method which considers the correlation between the labels to solve the weak label problem as a type of multi-label learning. Qi et al. [22] consider the partially annotated image learning as a multi-label problem and solve it by a statistical model.

The noise among the data is categorized into two types: attribute noise and class noise. In this paper, we focus on eliminating the bad effect raised by the class noise. There are a number of denoising methods for classification; they can be further classified into two categories: filtered preprocessing of the data and robust design of the algorithms. In the former category, filtered preprocessing is developed to remove the noise from the training set as much as possible [27, 32]. For the latter category, robust algorithms are designed to reduce the impact of the noise in the classification [16, 19, 26]. Lin and Wang [16] propose the fuzzy SVM to classify the noisy data by assigning a fuzzy membership to the cost of a target function. Liu and Zheng [19] propose the soft SVM, which is capable of dealing with binary as well as real-valued class memberships.

Data represented by multiple views exist in many real-world applications. It has been shown extensively in the literature that leveraging the redundancy among the multiple views can improve the learning performance [6, 13, 23]. Farquhar et al. [6] incorporate the correlation between two views into a single classifier called SVM-2K. Rosenberg et al. [23] plug the multi-view kernel into the standard kernel methods that are based on data-dependent regularization in reproducing kernel Hilbert spaces. However, all the above multi-view approaches use datasets of perfectly tagged instances as the training set, and ignore the noisy and incomplete tags of the training instances in the real-world scenarios.

The image annotation and retrieval problem has been cast as a multi-label learning problem in the literature with various approaches proposed [5]. One category of these approaches is to introduce a set of bridge variables to learn the joint distribution between the image features and the semantic labels by generative models. Barnard et al. [1] propose the mixture of multi-modal Latent Dirichlet Allocation model using the latent variables as the bridge. Blei and Jordan [2] propose the correspondence Latent Dirichlet Allocation model under which each label shares a latent variable with a randomly selected image feature descriptor in one image. In [7, 12], the annotated images in the training set are used as the connection variables to compute the joint probability between the labels and the image regions.

Different from the above approaches, another paradigm for image annotation is to reduce the problem to a set of binary classification problems solved by discriminative methods where each semantic label is considered as an independent class. Boutell et al. [3] introduce several transformation methods that map the multilabel learning problem into the single-label classification problem. A more complicated situation beyond the multi-label learning problem is when an image is represented as a bag of instances, and belongs to a bag of classes. Hence, the original annotation problem becomes a multi-instance and multi-label learning problem. Zhou and Zhang [30] solve this multi-instance and multi-label learning problem by mapping it into a single-instance and multi-label learning problem. The representative technique for this category of approaches is the classification technique such as the Support Vector Machine (SVM), which demonstrates a strong discrimination power [9, 21, 29]. Yang et al. [29] propose an asymmetrical support vector for region-based image annotation. Qi and Han [21] propose an automatic image annotation system, which integrates the multiple instance learning (MIL)-based and global-feature-based SVMs for annotation. However, most of these discriminative methods for image annotation only consider the multi-label space as the classification target, and fail to make use of the given multiple labels simultaneously as an additional feature.

There has been work on image tag refinement in the literature. Wang et al. [28] address the tag optimization problem given the tagging labels of images obtained by a classifier; Zhu et al. [31] propose a classification method to address the noisy tagging problem; neither of the methods explicitly tackles the incomplete tagging problem. Liu et al. [18] propose a solution that denoises the noisy tagging and enriches the incomplete tagging, while their enrichment can only add synonyms and hypernyms, and needs the availability and assistance from WordNet. In addition, the classification solutions in [18] and [31] are not inductive in learning.

Compared with the existing work which is either only able to handle the incomplete tagging learning problem and the noisy data learning problem separately, or is transductive learning only, our work tackles the imperfectly tagged data synergistically not only to make the incomplete tagging to be more complete, but also to make the noisy tagging to be more noise-free through an inductive learning. By reducing the image annotation problem to a set of binary classification problems, the proposed methods are successfully applied to solving the problem of learning from imperfectly annotated images. Further, our work explicitly considers all the given labels simultaneously as an additional feature. Hence, the information contained in the multi-label space can be taken as an advantage to improve the performance of the classification.

## **3. MODEL FORMULATION**

We assume the existence of different views to a dataset, which is typically true in many real-world problems such as image annotation where we have different views to an image dataset from image texture, color, shape, correlated text information, etc. We claim that leveraging the redundancy among the multiple views can improve the performance when learning from the dataset with imperfect tags. It is easy to understand that an instance which is classified into the same category from most of the multiple views is a valuable representation of this category in the role of classification. On the other hand, an instance which is classified into different categories from most of the multiple views is a useless representation of any category in the role of classification. Due to the existence of imperfect tags, different weights should be given to different training instances from the combination of multiple views to reflect their confidence of classification. Without loss of generality, we focus on two views, and develop an iterative learning approach to update the imperfect tagging with a better tagging by leveraging the complementary information from the two views. Further, a novel discriminative classifier based on two views is proposed, which takes advantage of the information contained in the multi-label space to improve the performance when learning from imperfect tagging.

#### **3.1 Data Representations and Symbols**

We denote a training dataset as  $\mathcal{I}$ . Two views of the dataset  $\mathcal{I}$ are denoted as  $\mathcal{F}^{(a)}$  and  $\mathcal{F}^{(b)}$ , respectively. Ideally, the two views are conditionally independent. Each instance  $I_i \in \mathcal{I}$  is tagged with various tags. The whole tag vocabulary for  $\mathcal{I}$  forms the Sdimensional multi-label space  $\mathcal{T}$ . When one tag  $T_r$   $(1 \leq r \leq S)$ is chosen as the classification target, the other tags can form the additional feature space of tags, denoted as  $\mathcal{L}$ . Obviously,  $dim(\mathcal{L})$ is S - 1. For each  $I_i \in \mathcal{I}$ , there exist feature descriptors  $\mathbf{x}_i^{(a)}$ and  $\mathbf{x}_i^{(b)}$  in  $\mathcal{F}^{(a)}$  and  $\mathcal{F}^{(b)}$ , respectively. Let an S-dimensional vector  $\mathbf{d}_i = (d_{i,1}, d_{i,2}, \ldots, d_{i,S})'$  be the tag representation for  $I_i$ , where  $d_{i,r} \in \{0,1\}, 1 \leq r \leq S$  represents the occurrence of the  $r^{th}$  tag  $T_r$  for  $I_i$ . For each tag  $T_r$ ,  $1 \leq r \leq S$ , an imperfectly tagged training set of  $D_r$  data points is denoted as  $\mathscr{D}_r =$  $\{I_i\}_{i=1}^{D_r} = \{\mathbf{x}_i^{(a)}, \mathbf{x}_i^{(b)}, \mathbf{d}_i\}_{i=1}^{D_r}$ . For each  $I_i$  in  $\mathscr{D}_r$ , we denote  $y_{i,r}$ as the weight of  $I_i, f_{i,r}^{(a)}$  and  $\mathcal{F}^{(b)}$ , respectively. Further,  $p_{i,r}^{(a)}$  and  $\mathcal{F}^{(b)}$ , respectively (see Section 3.4).

#### **3.2** The Overall Framework of MITL

We now apply a learning scheme to the imperfectly tagged training dataset by iteratively updating the weight  $y_{i,r}$  of each data point

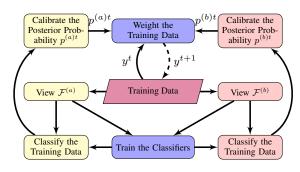


Figure 4: The flow chart of the MITL approach.

 $I_i$  in the training set  $\mathscr{D}_r$  from views  $\mathcal{F}^{(a)}$  and  $\mathcal{F}^{(b)}$ . The flow chart of MITL is shown in Figure 4. First, we extract feature vectors  $\mathbf{x}_i^{(a)}$  and  $\mathbf{x}_i^{(b)}$  in views  $\mathcal{F}^{(a)}$  and  $\mathcal{F}^{(b)}$ , respectively. Second, the feature vectors and the weights of the training data are used to train classifiers. Third, the training data are re-classified using the trained classifiers, and the returned  $f_{i,r}^{(a)}$  and  $f_{i,r}^{(b)}$  are transformed to the posterior probability  $p_{i,r}^{(a)}$  and  $p_{i,r}^{(b)}$  in  $\mathcal{F}^{(a)}$  and  $\mathcal{F}^{(b)}$ , respectively. Finally, we update the weights of the training data which are the prior of the next iteration. Through updating the weight  $y_{i,r}$ from the combination of the two views, the model progressively identifies the missing tags and the incorrect tags, and ultimately makes the classification in each view more precise and complete. Although different views are used jointly during the training process, each view can be utilized independently when predicting tags of any new, untagged instance after the training.

In principle, any classification method which uses the weights of the training data as the prior can be used in the framework of MITL. In this paper, we have developed the Multi-label Soft Support Vector Machine in two views (called MSS-2K) as the core classifier to facilitate incorporating the information contained in the multi-label space, and use the Sigmoid model to calibrate the posterior probability of the classifier. The methods to weight the training data in Figure 4 are different between dealing with the incorrect tags and dealing with the missing tags. The algorithm of the MITL approach is described in Algorithm 1. We discuss the details of the different components of the MITL approach in the following subsections. For the reference purpose, we call the MITL with MSS-2K as the core classifier as MITL in the rest of the paper.

#### 3.3 Classification: MSS-2K

In the two views learning, we assume that the data in the target dataset are of the following formulation:  $I_i = {\mathbf{x}_i^{(a)}, \mathbf{x}_i^{(b)}, l_i}_{i=1}^n$ , where  $\mathbf{x}_i^{(a)}$  ( $\mathbf{x}_i^{(b)}$ ) is a feature vector of  $I_i$  in view  $\mathcal{F}^{(a)}$  ( $\mathcal{F}^{(b)}$ ), and  $l_i \in {-1, +1}$  is a class label of  $I_i$ . The classic SVM-based classification is achieved through minimizing the distance between support planes constrained by the plane equations of the support vectors. In the two views learning, additional constraints are introduced to maximize the similarity between the classification results of the same instances in the two views. These two views constraints are presented as follows:

$$\forall_{i=1}^{n}: |\mathbf{w}^{(a)T}\mathbf{x}_{i}^{(a)} + \hat{b}^{(a)} - \mathbf{w}^{(b)T}\mathbf{x}_{i}^{(b)} - \hat{b}^{(b)}| \le \eta_{i}, \ \eta_{i} \ge 0$$
(1)

where  $\mathbf{w}^{(a)}$ ,  $\hat{b}^{(a)}$  ( $\mathbf{w}^{(b)}$ ,  $\hat{b}^{(b)}$ ) are the weight and bias of the SVM in  $\mathcal{F}^{(a)}$  ( $\mathcal{F}^{(b)}$ ), respectively.

The SVM-based methods solve the binary classification problem by finding the division plane to separate the instances of the two classes. In the OVA mode, these methods only utilize one tag of the data at a time, and ignore the other tags the data contain at the same time. It is easy to understand that the more the number of the tags,

Al	gorithm 1: MITL algorithm
1 I	nitialize $y_{i,r}^0$
2	foreach $T_r$ , $1 \le r \le S$ do
3 4	foreach $I_i \in \mathscr{D}_r$ do
4	foreach $I_i \in \mathscr{D}_r$ do $y_{i,r}^0 = 2 \cdot d_{i,r} - 1;$
5 L	et t = 0;
6 W	<b>hile</b> $t < the maximum number of iterations do$
7	foreach $T_r, \ 1 \leq r \leq S$ do
8	<b>Select</b> two-thirds instances of $\mathscr{D}_r$ randomly
9	to form the set $\mathscr{K}_r^t$ , $\mathscr{S}_r^t = \mathscr{D}_r - \mathscr{K}_r^t$ ;
10	<b>Train</b> MSS-2K using $y_{i,r}^t$ , $\mathbf{x}_i^{(a)}$ , $\mathbf{x}_i^{(b)}$ , and $\mathbf{d}_i$ of each
11	instance $I_i \in \mathscr{K}_r^t$ ;
12	foreach $I_i \in \mathscr{D}_r$ do
13	<b>Compute</b> the returned values $f_{i,r}^{(z)t}$ of MSS-2K
14	$\int  \text{in } \mathcal{F}^{(z)}, z = a, b;$
15	<b>Train</b> the Sigmoid models using $f_{i,r}^{(z)t}$ and $y_{i,r}^t$ of
16	each instance $I_i \in \mathscr{S}_r^t$ in $\mathcal{F}^{(z)}$ , $z = a, b$ ;
17	foreach $I_i \in \mathscr{D}_r$ do
18	<b>Calibrate</b> $p_{i,r}^{(z)t+1}$ through the Sigmoid models
19	in $\mathcal{F}^{(z)}, z = a, b;$
20	Update $y_{i,r}^{t+1}$ using $p_{i,r}^{(a)t+1}$ , $p_{i,r}^{(b)t+1}$ , and $y_{i,r}^{t}$ ;
21	Let $t = t + 1$ ;

the more the information contained in the tag space. Hence, the multi-label space can not only be considered as the classification target, but also be considered as an additional feature to improve the performance of the classification. However, most of the SVM-based methods fail to utilize the given multiple tags simultaneously.

In the OVA mode, when one tag  $T_r$  is chosen as the classification target, the other tags can form the additional feature space of the tags  $\mathcal{L}$ . It is a reasonable assumption that the similarity of the classification is inversely proportional to the distance of the instances in  $\mathcal{L}$ . Closer instances in  $\mathcal{L}$  have a higher similarity in the classification. We denote the feature vector of  $I_i$  in  $\mathcal{L}$  as  $L_i$ , where  $L_i = (d_{i,1}, \ldots, d_{i,r-1}, d_{i,r+1}, \ldots, d_{i,S})'$ . The neighborhood of  $L_i$  in  $\mathcal{L}$  (including  $L_i$  itself) is denoted as  $\mathscr{N}(L_i)$ . The information of the multi-label space can then be added into the two views learning by introducing the following constraints:

$$\begin{aligned} &\forall_{i=1}^{n} \ and \ \forall j \in \mathscr{N}_{i}: \\ &|\mathbf{w}^{(a)T}\mathbf{x}_{i}^{(a)} + \hat{b}^{(a)} - \mathbf{w}^{(b)T}\mathbf{x}_{j}^{(b)} - \hat{b}^{(b)}| \leq \eta_{ij}^{(ab)}, \ \eta_{ij}^{(ab)} \geq 0 \quad (2) \\ &|\mathbf{w}^{(b)T}\mathbf{x}_{i}^{(b)} + \hat{b}^{(b)} - \mathbf{w}^{(a)T}\mathbf{x}_{j}^{(a)} - \hat{b}^{(a)}| \leq \eta_{ij}^{(ba)}, \ \eta_{ij}^{(ba)} \geq 0 \quad (3) \end{aligned}$$

where  $\mathcal{N}_i \triangleq \{j | L_j \in \mathcal{N}(L_i)\}$ . Combining these two views constraints (2) and (3) with the typical SVM constraints and allowing different regularization constants, we obtain the following optimization:

$$\begin{split} \min_{\mathbf{v}^{(a)},\mathbf{w}^{(b)}} &\frac{1}{2} \|\mathbf{w}^{(a)}\|^{2} + \frac{1}{2} \|\mathbf{w}^{(b)}\|^{2} + C^{(a)} \sum_{i=1}^{n} \xi_{i}^{(a)} + C^{(b)} \sum_{i=1}^{n} \xi_{i}^{(b)} \\ &+ \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_{i}} C_{ij} (\eta_{ij}^{(ab)} + \eta_{ij}^{(ba)}) \\ C_{ij} &= \begin{cases} C & i = j \\ C^{*}/e^{dis(L_{i}, L_{j})} & i \neq j \quad (C^{*} < C) \end{cases} \quad (4) \\ s.t. \quad \forall_{i=1}^{n} : \quad l_{i} (\mathbf{w}^{(a)T} \mathbf{x}_{i}^{(a)} + \hat{b}^{(a)}) \geq 1 - \xi_{i}^{(a)}, \quad \xi_{i}^{(a)} \geq 0 \\ l_{i} (\mathbf{w}^{(b)T} \mathbf{x}_{i}^{(b)} + \hat{b}^{(b)}) \geq 1 - \xi_{i}^{(b)}, \quad \xi_{i}^{(b)} \geq 0 \end{cases} \\ \forall_{i=1}^{n} \quad and \quad \forall j \in \mathcal{N}_{i} : \\ |\mathbf{w}^{(a)T} \mathbf{x}_{i}^{(a)} + \hat{b}^{(a)} - \mathbf{w}^{(b)T} \mathbf{x}_{j}^{(b)} - \hat{b}^{(b)}| \leq \eta_{ij}^{(ab)}, \quad \eta_{ij}^{(ab)} \geq 0 \\ |\mathbf{w}^{(b)T} \mathbf{x}_{i}^{(b)} + \hat{b}^{(b)} - \mathbf{w}^{(a)T} \mathbf{x}_{j}^{(a)} - \hat{b}^{(a)}| \leq \eta_{ij}^{(ba)}, \quad \eta_{ij}^{(ba)} \geq 0 \end{split}$$

where  $dis(L_i, L_j)$  represents the distance between  $L_i$  and  $L_j$  in  $\mathcal{L}$ .

In order to reduce the optimization complexity, we need to approximate the optimization by reducing the constraint (3). We show that constraint (3) can be approximately obtained from constraint (2), and vice versa. For  $\forall_{i=1}^{n}, \forall j \in \mathcal{N}_{i}$ , we obtain the following constraint from constraint (2):

$$\begin{cases} |w^{(b)T}x_i^{(b)} + \hat{b}^{(b)} - w^{(a)T}x_j^{(a)} - \hat{b}^{(a)}| \le \eta_{ji}^{(ab)} \text{ if } i \in \mathscr{N}_j \\ |w^{(b)T}x_i^{(b)} + \hat{b}^{(b)} - w^{(a)T}x_j^{(a)} - \hat{b}^{(a)}| \le \tilde{\eta}_{ij} \text{ if } i \notin \mathscr{N}_j \end{cases}$$
(5)

where  $\tilde{\eta}_{ij} = \eta_{ij}^{(ab)} + \eta_{ii}^{(ab)} + \eta_{jj}^{(ab)}$ . When  $i \in \mathcal{N}_j$ ,  $\eta_{ij}^{(ba)} = \eta_{ji}^{(ab)}$ , and (5) coincides with constraint (3) strictly; when  $i \notin \mathcal{N}_j$ , for  $\eta_{ii}^{(ab)}(\eta_{jj}^{(ab)}) < \eta_{ij}^{(ab)}$ , (5) coincides with (3) approximately with a little larger constraint variable. Hence, we only select constraint (2) as the multi-label constraint in order to decrease the dimensionality of the parameters, as well as the computational complexity.

Further, in order to use the various weights of the training data in the role of classification, the range of the class label  $l_i$  is relaxed from  $\{-1, +1\}$  to an interval [-1, 1], which means that  $l_i$  can take any real value in [-1, 1] other than only two values in  $\{-1, 1\}$ . Consequently, MSS-2K is appropriate to be used in MITL by utilizing  $y_i$ , the weight of each instance  $I_i$  in the training set, as the class label. In MSS-2K,  $y_i$  is decomposed into two parts: the sign and the absolute value, i.e.,  $y_i = \text{sgn}(y_i) \cdot |y_i|$ , where  $\text{sgn}(y_i)$  is the indicator of the classification, and  $|y_i|$  reflects the confidence of the classification. Since  $\text{sgn}(y_i) \in \{-1, +1\}$ , the constraint in MSS-2K is analogous to the typical SVM constraints:

$$\operatorname{sgn}(y_i) \cdot (\mathbf{w}^T \mathbf{x}_i + \hat{b}) \ge |y_i| - \xi_i, \ \xi_i \ge 0.$$
(6)

Hence, the optimization is reduced to

$$\min_{\mathbf{w}^{(a)},\mathbf{w}^{(b)}} \frac{1}{2} \|\mathbf{w}^{(a)}\|^{2} + \frac{1}{2} \|\mathbf{w}^{(b)}\|^{2} + C^{(a)} \sum_{i=1}^{n} |y_{i}| \xi_{i}^{(a)} \\
+ C^{(b)} \sum_{i=1}^{n} |y_{i}| \xi_{i}^{(b)} + \sum_{i=1}^{n} |y_{i}| \sum_{j \in \mathcal{N}_{i}} C_{ij} \eta_{ij}^{(ab)} \\
C_{ij} = \begin{cases} C & i = j \\ C^{*}/e^{dis(L_{i}, L_{j})} & i \neq j \quad (C^{*} < C) \\ s.t. \ \forall_{i=1}^{n} : y_{i}(\mathbf{w}^{(a)T}\mathbf{x}_{i}^{(a)} + \hat{b}^{(a)}) \ge |y_{i}|^{2} - |y_{i}| \xi_{i}^{(a)}, \ \xi_{i}^{(a)} \ge 0 \\
y_{i}(\mathbf{w}^{(b)T}\mathbf{x}_{i}^{(b)} + \hat{b}^{(b)}) \ge |y_{i}|^{2} - |y_{i}| \xi_{i}^{(b)}, \ \xi_{i}^{(b)} \ge 0 \\
\forall_{i=1}^{n} and \ \forall i \in \mathcal{N}_{i}:
\end{cases}$$
(7)

$$|\mathbf{w}^{(a)T}\mathbf{x}_{i}^{(a)} + \hat{b}^{(a)} - \mathbf{w}^{(b)T}\mathbf{x}_{j}^{(b)} - \hat{b}^{(b)}| \le \eta_{ij}^{(ab)}, \ \eta_{ij}^{(ab)} \ge 0$$

With the Lagrange multipliers, we have the target function

$$\begin{split} \mathbf{L} &= \frac{1}{2} \|\mathbf{w}^{(a)}\|^2 + \frac{1}{2} \|\mathbf{w}^{(b)}\|^2 + C^{(a)} \sum_{i=1}^n |y_i| \xi_i^{(a)} \\ &+ C^{(b)} \sum_{i=1}^n |y_i| \xi_i^{(b)} + \sum_{i=1}^n |y_i| \sum_{j \in \mathcal{N}_i} C_{ij} \eta_{ij}^{(ab)} \\ &- \sum_{i=1}^n \alpha_i^{(a)} \left[ y_i (\mathbf{w}^{(a)T} \mathbf{x}_i^{(a)} + \hat{b}^{(a)}) - |y_i|^2 + |y_i| \xi_i^{(a)} \right] \\ &- \sum_{i=1}^n \alpha_i^{(b)} \left[ y_i (\mathbf{w}^{(b)T} \mathbf{x}_i^{(b)} + \hat{b}^{(b)}) - |y_i|^2 + |y_i| \xi_i^{(b)} \right] \\ &- \sum_{i=1}^n (\mu_i^{(a)} \xi_i^{(a)} + \mu_i^{(b)} \xi_i^{(b)}) - \sum_{i=1}^n \sum_{j \in \mathcal{N}_i} \left[ \mu_{ij}^{(ab)} \eta_{ij}^{(ab)} \\ &+ \beta_{ij}^{(ab)+} (\eta_{ij}^{(ab)} - \mathbf{w}^{(a)T} \mathbf{x}_i^{(a)} - \hat{b}^{(a)} + \mathbf{w}^{(b)T} \mathbf{x}_j^{(b)} + \hat{b}^{(b)}) \\ &+ \beta_{ij}^{(ab)-} (\eta_{ij}^{(ab)} + \mathbf{w}^{(a)T} \mathbf{x}_i^{(a)} + \hat{b}^{(a)} - \mathbf{w}^{(b)T} \mathbf{x}_j^{(b)} - \hat{b}^{(b)}) \right] \end{split}$$
(8)

where  $\alpha_i^{(a)}$ ,  $\alpha_i^{(b)}$ ,  $\beta_{ij}^{(ab)+}$ ,  $\beta_{ij}^{(ab)-}$ ,  $\mu_i^{(a)}$ ,  $\mu_i^{(b)}$  and  $\mu_{ij}^{(ab)}$  are the Lagrange multipliers. We claim that if the KKT condition of (8) shown in (9) is satisfied, the solution to (7) is the same as that to its dual problem.

$$\begin{array}{l} \nabla \mathbf{L} = 0 \\ \forall_{i=1}^{n}: \quad \alpha_{i}^{(z)} \left[ y_{i}(\mathbf{w}^{(z)T}\mathbf{x}_{i}^{(z)} + \hat{b}^{(z)}) - |y_{i}|^{2} + |y_{i}|\xi_{i}^{(z)} \right] = 0, \\ \mu_{i}^{(z)}\xi_{i}^{(z)} = 0, \quad z = a, b \\ \forall_{i=1}^{n} \quad and \quad \forall j \in \mathcal{N}_{i}: \quad \mu_{ij}^{(ab)}\eta_{ij}^{(ab)} = 0, \\ \beta_{ij}^{(ab)+}(\eta_{ij}^{(ab)} - \mathbf{w}^{(a)T}\mathbf{x}_{i}^{(a)} - \hat{b}^{(a)} + \mathbf{w}^{(b)T}\mathbf{x}_{j}^{(b)} + \hat{b}^{(b)}) = 0, \\ \beta_{ij}^{(ab)-}(\eta_{ij}^{(ab)} + \mathbf{w}^{(a)T}\mathbf{x}_{i}^{(a)} + \hat{b}^{(a)} - \mathbf{w}^{(b)T}\mathbf{x}_{j}^{(b)} - \hat{b}^{(b)}) = 0. \end{array} \right.$$

Let 
$$\nu_{ij}^{(ab)} = \beta_{ij}^{(ab)+} - \beta_{ij}^{(ab)-}$$
. From  $\nabla \mathbf{L} = 0$ , we have

$$\begin{aligned} \mathbf{w}^{(a)} &= \sum_{i=1}^{n} \alpha_{i}^{(a)} y_{i} \mathbf{x}_{i}^{(a)} - \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_{i}} \nu_{ij}^{(ab)} \mathbf{x}_{i}^{(a)} \\ \mathbf{w}^{(b)} &= \sum_{i=1}^{n} \alpha_{i}^{(b)} y_{i} \mathbf{x}_{i}^{(b)} + \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_{i}} \nu_{ij}^{(ab)} \mathbf{x}_{j}^{(b)} \\ \sum_{i=1}^{n} (\alpha_{i}^{(a)} y_{i} - \sum_{j \in \mathcal{N}_{i}} \nu_{ij}^{(ab)}) &= \sum_{i=1}^{n} (\alpha_{i}^{(b)} y_{i} + \sum_{j \in \mathcal{N}_{i}} \nu_{ij}^{(ab)}) = 0 \\ \forall_{i=1}^{n} : C^{(a)} |y_{i}| - \alpha_{i}^{(a)} |y_{i}| - \mu_{i}^{(a)} = C^{(b)} |y_{i}| - \alpha_{i}^{(b)} |y_{i}| - \mu_{i}^{(b)} = 0 \\ \forall_{i=1}^{n} and \ \forall j \in \mathcal{N}_{i} : \beta_{ij}^{(ab)+} + \beta_{ij}^{(ab)-} + \mu_{ij}^{(ab)-} - |y_{i}|C_{ij} = 0 \end{aligned}$$

Substituting the results from (9) and (10) to (8), we obtain the dual problem of (7) as follows:

$$\max - \frac{1}{2} \sum_{i,k=1}^{n} \left[ (\alpha_{i}^{(a)} y_{i} \mathbf{x}_{i}^{(a)} - \sum_{j \in \mathcal{N}_{i}} \nu_{ij}^{(ab)} \mathbf{x}_{i}^{(a)}) \\ \cdot (\alpha_{k}^{(a)} y_{k} \mathbf{x}_{k}^{(a)} - \sum_{h \in \mathcal{N}_{k}} \nu_{kh}^{(ab)} \mathbf{x}_{k}^{(a)}) \\ + (\alpha_{i}^{(b)} y_{i} \mathbf{x}_{i}^{(b)} + \sum_{j \in \mathcal{N}_{i}} \nu_{ij}^{(ab)} \mathbf{x}_{j}^{(b)}) \\ \cdot (\alpha_{k}^{(b)} y_{k} \mathbf{x}_{k}^{(b)} + \sum_{h \in \mathcal{N}_{k}} \nu_{kh}^{(ab)} \mathbf{x}_{h}^{(b)}) \right] + \sum_{i=1}^{n} (\alpha_{i}^{(a)} + \alpha_{i}^{(b)}) |y_{i}|^{2}$$
s.t. 
$$\sum_{i=1}^{n} (\alpha_{i}^{(a)} y_{i} - \sum_{j \in \mathcal{N}_{i}} \nu_{ij}^{(ab)}) = 0$$

$$\sum_{i=1}^{n} (\alpha_{i}^{(b)} y_{i} + \sum_{j \in \mathcal{N}_{i}} \nu_{ij}^{(ab)}) = 0$$

$$\forall_{i=1}^{n} : 0 \le \alpha_{i}^{(a)} \le C^{(a)}, \quad 0 \le \alpha_{i}^{(b)} \le C^{(b)}$$
(11)

$$\forall_{i=1}^{n} \ and \ \forall_{j} \in \mathcal{N}_{i}: \ \beta_{ij}^{(ab)+/-} \geq 0, \quad \beta_{ij}^{(ab)+} + \beta_{ij}^{(ab)-} \leq |y_{i}|C_{ij}$$

#### **3.4** Calibrating the Posterior Probability

After the process of training MSS-2K for each  $T_r$   $(1 \le r \le S)$ , the parameters of the classifiers,  $\mathbf{w}_r^{(z)}$  and  $b_r^{(z)}$  (z = a, b), have been optimized and the returned values of the classifiers  $f_{i,r}^{(z)} = \mathbf{w}_r^{(z)T} \mathbf{x}_i^{(z)} + \hat{b}_r^{(z)}$  in view  $\mathcal{F}^{(z)}$  (z = a, b) can be computed. The value of  $f_{i,r}^{(z)}$  should be confined to [0, 1] to cast a posterior probability of the class labels using the Sigmoid model [20] as

$$p_{i,r}^{(z)} = (e^{Af_{i,r}^{(z)} + B}) / (1 + e^{Af_{i,r}^{(z)} + B}), \quad z = a, \ b \tag{12}$$

Hence, the posterior log-likelihood of the class labels in the training dataset of size n is

$$\mathbf{H}_{r}^{(z)} = \log \prod_{i=1}^{n} P[\operatorname{sgn}(y_{i,r}) = 1 | f_{i,r}^{(z)}]^{t_{i,r}} P[\operatorname{sgn}(y_{i,r}) = -1 | f_{i,r}^{(z)}]^{1-t_{i,r}}$$
$$= \sum_{i=1}^{n} t_{i,r} \log p_{i,r}^{(z)} + (1-t_{i,r}) \log(1-p_{i,r}^{(z)}), \quad z = a, b$$
(13)

where  $t_{i,r} = \frac{1}{2}(1 + \operatorname{sgn}(y_{i,r}))$  is an index variable. The regression coefficients A and B for each  $T_r$   $(1 \le r \le S)$  in view  $\mathcal{F}^{(z)}$  (z = a, b) can be obtained numerically by maximizing the posterior log-likelihood.

As shown in Algorithm 1, we randomly select two-thirds data points from  $\mathscr{D}_r$  to form the set  $\mathscr{K}_r^t$  in each iteration t, which is used to train MSS-2K for each tag  $T_r$   $(1 \le r \le S)$ . The remaining data points in  $\mathscr{D}_r - \mathscr{K}_r^t$  are used to train the sigmoid model, and all the data points in  $\mathscr{D}_r$  are re-weighted as the prior to the next iteration. This iterative process avoids the over-fitting problem [20].

## **3.5** Weighting the Training Data

The methods to weight the training data in Figure 4, i.e., Step 20 in Algorithm 1, are different between dealing with the incorrect tags and handling the missing tags. The details of the methods are shown in Algorithm 2 and Algorithm 3, respectively.

#### 3.5.1 Correcting the incorrect tags

As shown in Figure 2(c), both the asterisks and the circles contain the data from the positive and the negative classes. First, we give different weights to different data points in the training set by integrating the results of the two views and shifting the weights to the range [-1, 1] in Step 2 of Algorithm 2. It is assumed that the number of the correctly tagged data points is always larger than that of the incorrectly tagged data points in each class in the training set. Hence, we reinforce the classification confidence of the instance and set  $|y_{i,r}| = 1$  if the updated values of  $y_{i,r}$  are both larger than e or both smaller than -e in the consecutive two iterations. Here e is a threshold which is set appropriately in  $(\frac{1}{2}, 1)$ to select the instances with the highest confidence to be positive or negative.

Al	Algorithm 2: Weight the Training Data for Incorrect Tags						
1 fc	preach $I_i \in \mathscr{D}_r$ do						
2	$ \begin{vmatrix} \mathbf{Compute} & q_{i,r}^{t+1} = p_{i,r}^{(a)t+1} + p_{i,r}^{(b)t+1} - 1; \\ \mathbf{if} & sgn(q_{i,r}^{t+1}) = sgn(y_{i,r}^{t}) \text{ and }  q_{i,r}^{t+1}  > e \text{ and }  y_{i,r}^{t}  > e \text{ then} \\ &   & sgn(y_{i,r}^{t+1}) = sgn(q_{i,r}^{t+1}),  y_{i,r}^{t+1}  = 1; \end{vmatrix} $						
3	if $sgn(q_{i,r}^{t+1}) = sgn(y_{i,r}^{t})$ and $ q_{i,r}^{t+1}  > e$ and $ y_{i,r}^{t}  > e$ then						
5	else						
6	else						

#### 3.5.2 *Completing the missing tags*

In the binary classification of the OVA mode, the incompletely tagged training set only contains correctly tagged positive instances and untagged instances. As Figure 2(b) shows, the asterisks are all the correctly tagged data points from the positive class, and we do not know which class each of the circles belongs to. In the initialization, we randomly select the untagged data points to be tagged as negative instances with the same number of the positive instances to train the classifiers, i.e., for each  $I_i \in \mathcal{D}_r$ , if  $I_i$  is untagged but selected as an initially negative instance,  $y_{i,r}^0 = -1$ . In the subsequent iterations, different weights are given to all the data points in  $\mathcal{D}_r$ , while the initially positive instances, i.e.  $\operatorname{sgn}(y_{i,r}^0) = 1$ ,

should be kept positive in all the iterations. Hence, in Step 8 of Algorithm 3, we shift the initially positive instances, which are set to negative values in the updating process, from the updating range to the range [0, c] linearly, where *c* is set appropriately in  $(0, \frac{1}{2})$  for these instances are not typical as the representations of the positive category in the training process.

Algorithm 3: Weight the Training Data for Missing Tags
1 foreach $I_i \in \mathscr{D}_r$ do
2 Compute $q_{i,r}^{t+1} = p_{i,r}^{(a)t+1} + p_{i,r}^{(b)t+1} - 1;$
<b>3</b> Find $Max_r = \max_i q_{i,r}^{t+1}$ and $Min_r = \min_i q_{i,r}^{t+1}$ ,
when $sgn(q_{i,r}^{t+1}) = -1$ and $sgn(y_{i,r}^{0}) = 1$ ;
4 foreach $I_i \in \mathscr{D}_r$ do
5   if $sgn(q_{i,r}^{t+1}) = sgn(y_{i,r}^t)$ and $ q_{i,r}^{t+1}  > e$ and $ y_{i,r}^t  > e$ then
5 if $sgn(q_{i,r}^{t+1}) = sgn(y_{i,r}^{t})$ and $ q_{i,r}^{t+1}  > e$ and $ y_{i,r}^{t}  > e$ then 6 $\lfloor sgn(y_{i,r}^{t+1}) = sgn(q_{i,r}^{t+1}),  y_{i,r}^{t+1}  = 1;$
7 else if $sgn(q_{i}, q_{i}) = -1$ and $sgn(y_{i}, q_{i}) = 1$ then
8 $y_{i,r}^{t+1} = c \cdot (q_{i,r}^{t+1} - Min_r) / (Max_r - Min_r),$
8 $\begin{cases} y_{i,r}^{t+1} = c \cdot (q_{i,r}^{t+1} - Min_r) / (Max_r - Min_r), \\ where \ 0 < c < \frac{1}{2}; \end{cases}$
9 else
$10 \qquad \qquad$

## 4. EXPERIMENTS

#### 4.1 Data and Parameter Setting

We apply our methods, including MSS-2K and MITL, to the problem of learning from imperfect tagging with a more specific context — imperfect image annotation. Since the current literature considers the learning from imperfect tagging problem as incomplete tagging learning and noisy data learning two separate problems, two groups of comparative experiments on these two tasks are conducted to evaluate the performances of our methods. On the former task, we compare MSS-2K and MITL with SVM, iter-SVM [17], co-training (with SVMs as classifiers), and SVM-2K [6] using the incompletely tagged images as the training set. On the latter task, we compare MSS-2K and MITL with SVM, fuzzy SVM [16], co-training, and SVM-2K using the noisily tagged images as the training set.

The NUS-WIDE [4] image database is used in the experiments. It includes 269, 648 web images and 81 concepts which we treat as the ground truth tags. Since the database has a very uneven distribution for the number of images corresponding to each concept, we choose the top 75 concepts whose numbers of positive examples are larger than 350 from the database to form the multilabel space  $\mathcal{T}$ . Hence, the dimensionality of the additional feature space of tags ( $\mathcal{L}$ ) for each  $\mathcal{T}_r$  is 74. For each concept, we randomly choose 150 positive examples and 150 negative examples to form the perfectly tagged training set. In the testing set, the numbers of the positive and negative examples are both 100. The left 100 positive examples and 100 randomly selected negative examples form the extra untagged data used only for co-training. For the task of completing missing tags, we formulate the untagged data of the incompletely tagged training set through combining all the negative instances of the perfectly tagged training set with v% randomly selected positive instances of the perfectly tagged training set. The whole incompletely tagged training set, as shown in Figure 2(b), comprises the newly created untagged instances and the residual 1 - v% positive instances. For the task of correcting incorrect tags, s% noise is added into both of the positive and negative

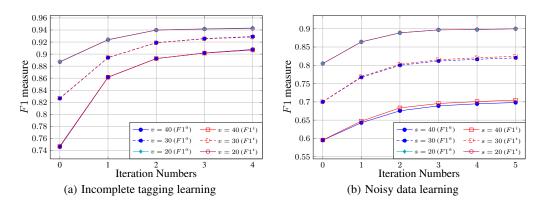


Figure 5: The F1 measure for the imperfectly tagged training set using MITL when u = 4 and R = 1.2.

examples of the perfectly tagged training set for each concept to form the noisily tagged training set. In the experiments, the 500-D bag of words feature based on SIFT descriptions is used as the feature of view  $\mathcal{F}^{(a)}$ , and the 1000-D bag of text words feature which describes the text information correlated to the images provided by the database is used as the feature of view  $\mathcal{F}^{(b)}$ . For the task of completing the missing tags, SVM, co-training, SVM-2K, and MSS-2K use the same initialization as MITL does described in Section 3.5.2 to obtain the initially negative training data from the incompletely tagged training set.

The size u of the neighborhood  $\mathcal{N}(L_i)$  for each  $L_i$  is defined as the count of the nearest neighbors of  $L_i$  in  $\mathcal{L}$ . When u = 1and each  $|y_i| = 1$ , MSS-2K is reduced to SVM-2K. We define R as the ratio between different kinds of regularization constants. For MSS-2K,  $R = \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_i} C_{ij} / (\frac{n}{2} \cdot (C^a + C^b))$ . The returned value of the classifiers in two views combined is denoted as  $f_{i,r}^{(ab)}$ ,  $f_{i,r}^{(ab)} = 0.5 \cdot (f_{i,r}^{(a)} + f_{i,r}^{(b)})$ .

### 4.2 **Results and Discussions**

For each tag  $T_r$ , let  $CT_r$  be the number of correctly predicted examples,  $GT_r$  be the number of the examples which actually have the tag as the ground truth, and  $PT_r$  be the number of all the predicted examples with the tag. Then the precision  $Pre_r$ , recall  $Rec_r$ , and  $F1_r$  measure are defined as

$$Pre_r = \frac{CT_r}{PT_r}; \ Rec_r = \frac{CT_r}{GT_r}; \ F1_r = \frac{2Pre_rRec_r}{Pre_r + Rec_r} = \frac{2CT_r}{PT_r + GT_r}$$

We evaluate the performances of the methods using the standard performance measures of Macro-F1 and Micro-F1. Macro-F1 averages the F1 measures on the predictions of different tags; Micro-F1 computes the F1 measure on the predictions of different labels as a whole.

$$F1^{a} = \frac{1}{S} \sum_{r=1}^{S} F1_{r}; \quad F1^{i} = \frac{2\sum_{r=1}^{S} CT_{r}}{\sum_{r=1}^{S} PT_{r} + \sum_{r=1}^{S} GT_{r}}$$

We describe the F1 measure for the incompletely tagged training set and for the noisily tagged training set using MITL as a function of the iteration in Figure 5(a) and Figure 5(b), respectively. We observe from Figure 5 that the F1 measures increase 6% to 16% for the incompletely tagged training set, and 9% to 12% for the noisily tagged training set after the iteration progresses, which shows that MITL is effective in completing the missing tags and in correcting the incorrect tags for the training set. Due to the elimination of the incompleteness and noise in the training set, the training accuracies increase. The curves in Figure 5(a) level up more than those in Figure 5(b), because the initially positive examples are kept positive during the whole iterations on the task of completing the missing tags. We choose the maximum numbers of iterations for MITL based on the F1 measure of the training set. The curves for the F1 measure of MITL in Figure 5(a) and Figure 5(b) all exhibit their major elevation between the first two iterations, then level off during the subsequent iterations and quickly converge, indicating that MITL converges fast.

In Table 1, we summarize the F1 measure for the testing set with the imperfectly tagged training set in View  $\mathcal{F}^{(a)}$ , in View  $\mathcal{F}^{(b)}$ , and in two views combined, respectively. On incomplete tagging learning, we compare our methods with SVM, iter-SVM, co-training, and SVM-2K when v is selected as 40, 30, 20, and 10, respectively. On noisy data learning, we compare our methods with SVM, fuzzy SVM, co-training, and SVM-2K when s is selected as 40, 30, 20, and 10, respectively.

Compared with our proposed method MITL, SVM with the imperfectly tagged training set considers the training set as the perfectly tagged set and mistakenly takes the missing or incorrect tags as perfect tags. MSS-2K also considers the training set as the perfectly tagged set, while it utilizes the information contained in the multi-label space which helps eliminate the incompleteness and noise in the imperfectly tagged training set. Although iter-SVM and fuzzy SVM consider the training set as incompletely tagged or noisily tagged dataset, respectively, they cannot use the information contained in multiple views synergistically, and cannot consider all the given multiple tags simultaneously as an additional feature as MSS-2K and MITL do. SVM-2K utilizes the information contained in multiple views synergistically, but it mistakenly takes the missing or incorrect tags as perfect tags in the imperfectly tagged training set. Consequently, from Table 1, we see that MITL performs better than SVM, iter-SVM, SVM-2K, and MSS-2K with the incompletely tagged training set, and performs better than SVM, fuzzy SVM, SVM-2K, and MSS-2K with the noisily tagged training set all in view  $\mathcal{F}^{(a)}$ , in view  $\mathcal{F}^{(b)}$ , and in two views combined when v and s are selected as 40, 30, 20, and 10, respectively. The performance of MSS-2K is inferior to the performance of MITL but superior to the performances of the other methods. The performance of MITL is much better than those of co-training both with incompletely tagged and noisily tagged training sets when v and s are chosen as 40, 30, and 20, respectively. When the ratios of the incomplete tags and the noisy tags are small, i.e., v = 10 and s = 10, MITL performs slightly better than co-training because of the usage of the extra untagged training set in co-training, which is not available to MITL. Co-training considers the imperfectly tagged training data as perfectly tagged data and holds both the

## Table 1: The F1 measure for the testing set.

(a) The  $F1^a \setminus F1^i$  for the testing set on incomplete tagging learning in View  $\mathcal{F}^{(a)}$  and View  $\mathcal{F}^{(b)}$ , respectively.

		$F1^a \backslash F1^i$ in	View $\mathcal{F}^{(a)}$		$F1^a \backslash F1^i$ in View $\mathcal{F}^{(b)}$				
	v = 40	v = 30	v = 20	v = 10	v = 40	v = 30	v = 20	v = 10	
SVM	$0.6441 \langle 0.6492$	$0.6816 \\ 0.6844$	$0.6971 \langle 0.6988$	$0.7142 \ 7164$	0.7836\0.7855	$0.8080 \setminus 0.8089$	$0.8255 \setminus 0.8263$	$0.8420 \ 0.8438$	
Iter-SVM	$0.6617 \setminus 0.6675$	0.6903 (0.6931)	0.7033 (0.7051	0.7156 (0.7179)	0.8100\0.8077	0.8149 (0.8160	0.8301 (0.8292)	0.8440 (0.8440)	
Co-training	$0.6577 \setminus 0.6625$	$0.6877 \\ 0.6901$	$0.7060 \\ 0.7076$	$0.7212 \langle 0.7233$	0.7887\0.7905	$0.8140 \\ 0.8147$	$0.8322 \ 0.8328$	0.8505 (0.8521)	
SVM-2K	$0.6539 \\ 0.6605$	$0.6904 \\ 0.6930$	$0.6978 \\ 0.6993$	0.7139 (0.7159	0.7889\0.7895	0.8117 (0.8124)	$0.8292 \ 0.8304$	0.8483 (0.8494)	
MSS-2K	$0.6811 \setminus 0.6859$	0.7092 (0.7118)	$0.7194 \\ 0.7208$	0.7259 (0.7274)	0.8101\0.8119	0.8271 (0.8276)	0.8370 (0.8383)	0.8505 (0.8519)	
MITL	$0.7043 \langle 0.7088$	0.7227 (0.7243)	$0.7264 \langle 0.7266$	$0.7280 \langle 0.7290$	$0.8346 \langle 0.8360$	0.8443 (0.8445)	$0.8476 \langle 0.8467$	0.8554 (0.8563)	

(b) The  $F1^a \setminus F1^i$  for the testing set on noisy data learning in View  $\mathcal{F}^{(a)}$  and View  $\mathcal{F}^{(b)}$ , respectively.

		$F1^a \backslash F1^i$ in	View $\mathcal{F}^{(a)}$		$F1^a \backslash F1^i$ in View $\mathcal{F}^{(b)}$				
	s = 40	s = 30	s = 20	s = 10	s = 40	s = 30	s = 20	s = 10	
SVM	$0.5698 \\ 0.5761$	$0.6279 \\ 0.6330$	$0.6830 \\ 0.6852$	$0.7084 \\ 0.7093$	0.6065\0.6133	0.7062 (0.7099)	0.7796 (0.7815)	$0.8314 \\ 0.8307$	
Fuzzy SVM	0.5891 (0.5954)	$0.6430 \\ 0.6481$	$0.6922 \\ 0.6944$	$0.7137 \\ 0.7146$	0.6288\0.6314	$0.7233 \\ 0.7261$	$0.7901 \langle 0.7903$	0.8352 (0.8342)	
Co-training	0.5768 (0.5819)	0.6353 (0.6401)	0.6895 (0.6915)	0.7148 (0.7157	0.6127\0.6176	0.7129 (0.7163)	0.7857 (0.7874)	0.8361 (0.8352)	
SVM-2K	$0.5708 \\ 0.5776$	$0.6315 \ 0.6375$	$0.6990 \\ 0.7018$	$0.7087 \langle 0.7098$	0.6080\0.6116	0.7020 (0.7056)	$0.7702 \\ 0.7721$	0.8362 (0.8362)	
MSS-2K	$0.6036 \langle 0.6130$	$0.6648 \ 0.6737$	$0.7071 \langle 0.7097$	$0.7210 \langle 0.7213$	0.6365\0.6433	0.7451 (0.7489)	0.8076 (0.8090)	$0.8414 \\ 0.8410$	
MITL	$0.6265 \backslash 0.6349$	$0.6819 \langle 0.6896$	$0.7181 \langle 0.7208$	0.7268 ackslash 0.7272	0.6808\0.6863	0.7865 (0.7888)	$0.8328 \backslash 0.8338$	0.8529 (0.8521)	

(c) The  $F1^a \setminus F1^i$  for the testing set in two views combined on incomplete tagging learning and on noisy data learning, respectively.

	$F1^{a}\setminus F1^{i}$ in two views combined on incomplete tagging learning					$F1^{a}\backslash F1^{i}$ in two views combined on noisy data learning			
	v = 40	v = 30	v = 20	v = 10		s = 40	s = 30	s = 20	s = 10
SVM	$0.7850 \\ 0.7887$	0.8207\0.8221	$0.8381 \\ 0.8397$	0.8532 (0.8549)	SVM	$0.6140 \langle 0.6210$	0.7137 (0.7184)	0.7956 (0.7984)	$0.8452 \\ 0.8443$
Iter-SVM	$0.8177 \setminus 0.8161$	0.8285 (0.8304)	$0.8433 \\ 0.8430$	0.8543 (0.8552)	Fuzzy SVM	$0.6393 \\ 0.6423$	0.7337 (0.7384)	$0.8101 \\ 0.8101$	$0.8498 \\ 0.8489$
Co-training	0.7927 (0.7959)	$0.8236 \ 0.8248$	$0.8470 \ 0.8484$	$0.8636 \\ 0.8651$	Co-training	$0.6190 \ 0.6258$	$0.7229 \\ 0.7271$	$0.8030 \ 0.8046$	0.8563 (0.8554)
SVM-2K	$0.7962 \ 0.7996$	$0.8271 \\ 0.8281$	$0.8443 \\ 0.8455$	0.8597 (0.8610)	SVM-2K	$0.6138 \\ 0.6188$	$0.7183 \\ 0.7230$	0.7965 (0.7983)	0.8516 (0.8510)
MSS-2K	0.8175\0.8209	$0.8444 \\ 0.8454$	0.8527 (0.8543)	0.8623 (0.8639)	MSS-2K	$0.6387 \langle 0.6466$	$0.7614 \\ 0.7665$	$0.8249 \\ 0.8265$	$0.8584 \\ 0.8580$
MITL	$0.8419 \\ 0.8447$	0.8589 (0.8595)	$0.8606 \langle 0.8619$	$0.8670 \backslash 0.8681$	MITL	0.6841 ackslash 0.6907	$0.7986 \langle 0.8030$	$0.8475 \backslash 0.8492$	$0.8672 \backslash 0.8662$

initially positive and initially negative instances unchanged. In the training process, co-training extracts the instances with the highest confidence to be positive or negative from the extra untagged set added into the training set iteratively and progressively from multiple views. The higher the ratios of the imperfect tags in the training set, the more noisy the instances extracted from the extra untagged set to be added into the training set by co-training. Consequently, co-training performs much worse than MITL when more imperfect tags are contained in the training set. In general, MITL and MSS-2K perform well with the imperfectly tagged training set not only on the task of completing the missing tags, but also on the task of correcting the incorrect tags. Combining the two views delivers a better performance than using any of the two views independently, while we also observe from Table 1 that after the training process, the classifiers in each of the two views  $\mathcal{F}^{(a)}$  and  $\mathcal{F}^{(b)}$  can work independently without the information of the other view, and perform more precisely when predicting tags of any new, untagged image. Further, from Table 1 we also observe that our proposed methods perform much better than the other methods when the incompleteness ratio v% increases and the noise ratio s% increases, respectively.

We describe the F1 measure for the testing set as a function of u when using MSS-2K and MITL on incomplete tagging learning and on noisy data learning in Figure 6(a) and Figure 6(b), respectively. We observe from Figure 6 that the F1 measures for the testing set all increase when the size of the neighborhood for each  $\mathcal{N}(L_i)$  increases, which shows that it is helpful to use the nearest neighbors of each  $L_i$  in  $\mathcal{L}$  to further improve the classification performance, and that it is able to deliver a more effective classification performance when considering all the given tags simultaneous as an additional feature in the proposed methods. The curves for the F1 measures of MSS-2K and MITL in Figure 6 all exhibit their major elevation from u = 1 to u = 4, then level off or even decline a little when u continues to increase, indicating that there is no need to choose a much larger u which would increase the computational complexity. As we describe before, when u = 1 and each  $|y_i| = 1$ , MSS-2K is reduced to SVM-2K. Figure 6 shows that the performances of MSS-2K and MITL are much better than that of SVM-2K.

Figure 7(a) and Figure 7(b) show the F1 measure for the testing set as a function of R when using MSS-2K and MITL on incomplete tagging learning and on noisy data learning, respectively. As we defined before, R represents the ratio between different kinds of regularization constants. When R increases, the proportion of the two views constraints in the optimization also increases, while the proportion of the typical SVM constraints in the optimization decreases. We observe from Figure 7 that when R increases, the curves for the F1 measures of MSS-2K and MITL ascend, which shows that the two views constraints are more valuable than the typical SVM constraints in MSS-2K and MITL, and that it is helpful to take advantage of the information contained in the multi-label space to further improve the classification performance. Further, in Figure 7, the performance of MITL is better than that of MSS-2K regardless of what value R is, indicating the effectiveness of the general framework of MITL.

Figure 8 shows the F1 measure of each tag for the testing set using MITL on incomplete tagging learning when v = 30, and Figure 9 shows the F1 measure of each tag for the testing set using MITL on noisy data learning when s = 20.

As a further case study to demonstrate the superiority of the MITL performance, Figure 1(a) is one of the imperfectly tagged training images in the training data of NUS-WIDE database with the given ground truth tags to this image as *buildings*, *cityscape*, *sky*, and *street*, and Figure 1(b) is another imperfectly tagged training images in the training data of NUS-WIDE database with the given ground truth tags to this image as *animal* and *fish*. After the learning with MITL, Figure 1(a) is also tagged with the missing tags *person*, *tree*, and *water*, which actually are the objects in the

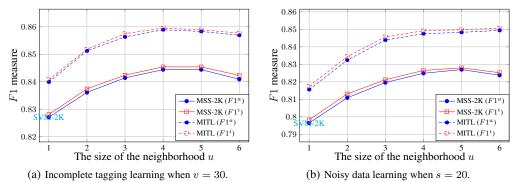


Figure 6: The F1 measure for the testing set in two views combined using MSS-2K and MITL when R = 1.2.

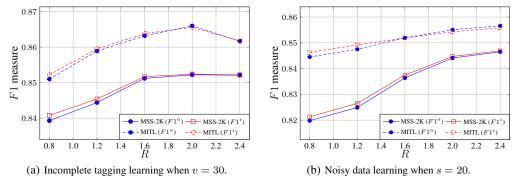


Figure 7: The F1 measure for the testing set in two views combined using MSS-2K and MITL when u = 4.

image but are not originally given in the ground truth. With MITL, the animal in Figure 1(b) is tagged as *bird* instead of *fish*, which is identified as a noisy tag originally given in the ground truth incorrectly. Further, Figure 1(b) is also tagged with the missing tag *plants*, which actually is the object in the image but is not originally given in the ground truth. This shows that MITL is an effective discriminative method not only to complete the missing tags for the incompletely tagged training set, but also to correct the incorrect tags originally given in the noisily tagged training set.

#### 5. CONCLUSION

This paper studies the arguably new and important problem of learning from imperfect tagging. We argue that this problem exists in many real-world applications and that the solutions to this problem shall generate great societal and technical impacts. We then develop a general and effective solution which we call Multi-view Imperfect Tagging Learning (MITL) to this problem. The novelty of the proposed approach is that it can not only use the incompletely tagged training data to complete the missing tags, but also use the noisily tagged training data to correct the incorrect tags through an inductive learning. Multiple views are used in the training process while each view can be utilized independently when predicting tags of any new, untagged data, and the performances of the classifiers in the multiple views are significantly improved after the training. Further, a novel discriminative classification method called MSS-2K is proposed under the framework of MITL, which explicitly makes use of the given multiple labels simultaneously as an additional feature to deliver a more effective classification performance than the existing literature where one label is considered at a time as the classification target while the rest of the given labels are completely ignored at the same time. We apply our methods MITL and MSS-2K to the problem with a more specific context — imperfect

image annotation, and evaluate them on a standard dataset from the related literature in comparison with peer methods from the recent literature. Experimental results demonstrate the effectiveness and promise of our methods as solutions to the problem of learning from imperfect tagging.

# 6. ACKNOWLEDGMENTS

This work is supported in part by National Basic Research Program of China (2012CB316400), and ZJU–Alibaba Joint Lab. Zhongfei Zhang is also supported in part by US NSF (IIS-0812114, CCF-1017828).

# 7. REFERENCES

- K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *Journal* of Machine Learning Research, 3:1107–1135, 2003.
- [2] D. M. Blei and M. I. Jordan. Modeling annotated data. In Proceedings of the 26th International ACM SIGIR Conference, 2003.
- [3] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.
- [4] T. S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 1–9, 2009.
- [5] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys, 40:1–60, 2008.
- [6] J. Farquhar, D. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak. Two view learning: SVM-2K, theory and

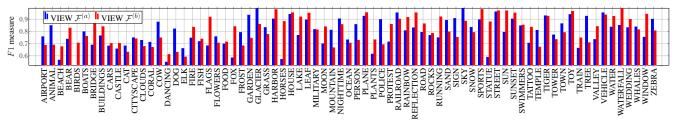


Figure 8: The F1 measure of the 75 tags for the testing set using MITL on incomplete tagging learning when v = 30, u = 4, and R = 1.2.

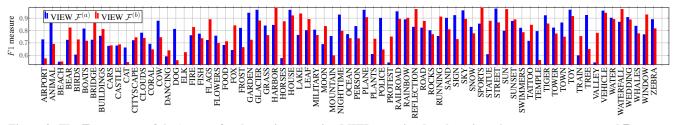


Figure 9: The F1 measure of the 75 tags for the testing set using MITL on noisy data learning when s = 20, u = 4, and R = 1.2.

practice. In Advances in Neural Information Processing Systems. MIT Press, 2006.

- [7] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Proceedings of CVPR*, pages 1002–1009, 2004.
- [8] G. P. C. Fung, J. X. Yu, H. Lu, and P. S. Yu. Text classification without negative examples revisit. *IEEE Transactions on Knowledge and Data Engineering*, 18:6–20, 2006.
- [9] K.-S. Goh, E. Chang, and K.-T. Cheng. SVM binary classifier ensembles for image classification. In *Proceedings* of the tenth international conference on Information and knowledge management, pages 395–402, 2001.
- [10] J. Grappone and G. Couzin. Search Engine Optimization: An Hour a Day. Wiley Publishing, Indianapolis, 2006.
- [11] P. Heymann, D. Ramage, and H. G. Molina. Social tag prediction. In *Proceedings of the 31st International ACM SIGIR Conference*, pages 531–538, 2008.
- [12] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Proceedings of Neural Information Processing Systems*, 2003.
- [13] G. Li, S. C. H. Hoi, and K. Chang. Two-view transductive support vector machines. In SDM, pages 235–244, 2010.
- [14] X. Li and B. Liu. Learning to classify text using positive and unlabeled data. In *Proceedings of the 18th international joint* conference on Artificial intelligence, pages 587–592, 2003.
- [15] Y. Li, D. J. Crandall, and D. P. Huttenlocher. Landmark classification in large-scale image collections. In *Proceedings of International Conference on Computer Vision*, pages 1957 – 1964, 2009.
- [16] C. F. Lin and S. de Wang. Training algorithms for fuzzy support vector machines with noisy data. *Pattern Recognition Letters*, 25:1647–1656, 2004.
- [17] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu. Building text classifiers using positive and unlabeled examples. In *Proceedings of ICDM*, pages 179–186, 2003.
- [18] D. Liu, X.-S. Hua, M. Wang, and H.-J. Zhang. Image retagging. In *Proceedings of ACM MM*, pages 491–500, 2010.
- [19] Y. Liu and Y. F. Zheng. Soft SVM and its application in video-object extraction. *IEEE Transactions on Signal Processing*, 55(7-1):3272–3282, 2007.
- [20] J. C. Platt. Probabilistic outputs for support vector machines

and comparisons to regularized likelihood methods. In Advances in Large Margin Classifiers. MIT Press, 1999.

- [21] X. Qi and Y. Han. Incorporating multiple SVMs for automatic image annotation. *Pattern Recognition*, 40:728–741, 2007.
- [22] Z. Qi, M. Yang, Z. Zhang, and Z. Zhang. Mining partially annotated images. In *Proceedings of the 17th International* ACM SIGKDD Conference, pages 1199–1207, 2011.
- [23] D. S. Rosenberg, V. Sindhwani, P. L. Bartlett, and P. Niyogi. A kernel for semi-supervised learning with multi-view point cloud regularization. In *IEEE Signal Processing Magazine*, 2009.
- [24] Y.-Y. Sun, Y. Zhang, and Z.-H. Zhou. Multi-label learning with weak label. In *Proceedings of AAAI Conference*, pages 593–598, 2010.
- [25] M. Szomszor, C. Cattuto, H. Alani, K. O'Hara, A. Baldassarri, V. Loreto, and V. D. Servedio. Folksonomies, the semantic web, and movie recommendation. In 4th European Semantic Web Conference: Bridging the Gap between Semantic Web and Web 2.0, 2007.
- [26] J. Tang, R. Hong, S. Yan, T.-S. Chua, G.-J. Qi, and R. Jain. Image annotation by knn-sparse graph-based label propagation over noisily tagged web images. ACM *Transactions on Intelligent Systems and Technology*, 2:14:1–14:15, 2011.
- [27] J. Van Hulse and T. M. Khoshgoftaar. Class noise detection using frequent itemsets. *Intelligent Data Analysis*, 10:487–507, 2006.
- [28] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Content-based image annotation refinement. In *Proceedings of CVPR*, pages 1–8, 2007.
- [29] C. Yang, M. Dong, and J. Hua. Region-based image annotation using asymmetrical support vector machine-based multi-instance learning. In *Proceedings of CVPR*, 2006.
- [30] Z.-H. Zhou and M.-L. Zhang. Multi-instance multilabel learning with application to scene classification. In *Proceedings of NIPS*, pages 1609–1616, 2007.
- [31] G. Zhu, S. Yan, and Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *Proceedings* of ACM MM, pages 461–470, 2010.
- [32] X. Zhu, X. Wu, and Q. Chen. Eliminating class noise in large datasets. In *Proceeding of International Conference on Machine Learning*, pages 920–927, 2003.