# Mining Noisy Tagging from Multi-label Space

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

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

## ABSTRACT

In this paper we study the problem of mining noisy tagging. Most of the existing discriminative classification methods to this problem only consider one tag at a time as the classification target, and completely ignore the rest of the given tags at the same time. In this paper we argue that all the given multiple tags can be utilized simultaneously as an additional feature and the information contained in the multi-label space can be taken advantage of to improve the performance of the classification. We first propose a novel distance measure to compute the distance between instances in the multi-label space. Then we propose several novel methods to incorporate the information of the multi-label space into the discriminative classification methods in one view learning or in two views learning to solve a general multi-label classification problem and to mitigate the influence of the noise in the classification. We apply the proposed solutions to the problem with a more specific context — noisy 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 in the existing literature on solving the problem of mining noisy tagging.

## 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

Noisy tagging; multi-label space; image annotation prediction

## 1. INTRODUCTION

Recent research has witnessed a great success in developing effective solutions to many important real-world problems through appropriate tagging. Thus, tagging is considered as an important means to develop solutions to many such real-world problems, especially in the areas of information retrieval, knowledge management, and database. 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. For social tagging, the tagging results can never be expected to be perfect, and it always contains lots of noise, including missing tags and/or incorrect tags. Figure 1 shows exemplar images with noisy tagging. In Figure 1(a), the given annotations of the image are *buildings*, *cityscape*, *sky*, and *street*, while there exist many other things missed to be tagged, such as *person*, *tree*, *water*, *fountain*, and *street lamp*. The image in Figure 1(b) is given tags of *animal*, *plants*, *horses*, *grass*, and *fox*, while it is obvious that the tag *fox* is given incorrectly. In Figure 1(c), the given annotations of the image are *animal* and *fish*, while it is obvious that the animal in the image is *bird* instead of *fish*, and there exists missing tag of *plants*.


**Figure 1: Exemplar images with noisy tagging.**

In general an $N$-class classification problem can always be decomposed into $N$ binary classification problems in the one-vs-all (OVA) mode. 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 to 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 and at the same time completely ignore the rest of the tags. In this paper, we explicitly consider all the given tags simultaneously as an additional feature which further helps improve the classification performance. We claim that when the tags contain noise, taking advantage of all the given tags shall mitigate the influence of the noise compared with only considering one tag at a time as the classification target. Assuming that the correct tags are always the majority in all the given tags of an instance, these tags can provide additional information to help improve the training accuracy even when the instance is incorrectly labeled in the OVA mode. Figure 2 shows exemplar images with multiple tags. The images in the first row, which are all tagged as *fish*, always have the accompanied tags of *water*, *coral*, and *ocean*, while the images in the second row, 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*.

In this paper we study the problem of mining noisy tagging. We


*CIKM'12,* October 29–November 2, 2012, Maui, HI, USA.

**Figure 2: Exemplar images with multiple tags.**

argue that noisy tagging exists in many real-world applications and that the solutions to the problem of mining noisy tagging shall generate great societal and technical impacts. This paper proposes a new approach which makes 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 and in mitigating the influence of the noise in the classification. We first propose a novel distance measure to compute the distance between instances in the multi-label space, which considers the various relationships among the multiple tags. Then we propose three novel methods to incorporate the information of the multi-label space into the discriminative classification methods in one view learning or in two views learning, which we call SVM with Multi-label Soft Membership (SVM-MSM), SVM with Multi-label Constraints (SVM-MC), and Multi-label SVM in two views (MSVM-2K), respectively. We apply the proposed methods to the problem of mining noisy tagging with a more specific context — noisy image annotation, and demonstrate through extensive evaluations using real data that the proposed methods perform well in comparison with the peer methods in the literature as effective and promising solutions to the problem of mining noisy tagging.

## 2. RELATED WORK

The noise among the data is categorized into two types: attribute noise and class noise [10]. 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 [9, 11]. For the latter category, robust algorithms are designed to reduce the impact of the noise in the classification [4, 5, 7]. Lin and Wang [4] propose the fuzzy SVM to classify the noisy data by assigning a fuzzy membership to the cost of a target function. Tang et al. [7] modify the kNN-sparse semi-supervised learning on graph to annotate the tagged images with noise.

Most of the existing discriminative classification methods to the multi-label problem only consider one tag at a time as the classification target, and completely ignore the rest of the given tags at the same time. There has been work of treating sets of labels as single labels [6, 8]. However, all these methods increase the number of the classifiers if label sets between instances differ substantially, and fail to take advantage of all the labels simultaneously in one classifier. Godbole and Sarawagi [3] extend the original datasets with $|S|$ extra features containing the predictions of each binary classifier; then a second round of training $|S|$ new binary classifiers takes place using the extended datasets. However, this proposed method (called SVM-HF) fails to make use of the original given multiple labels of each instance simultaneously.

Compared with the existing work, our work takes advantage of all the given multiple tags simultaneously as well as considers the various relationships among the multiple tags to mitigate the influence of the noise in the classification.

## 3. MODEL FORMULATION

The main idea of the proposed methods is to minimize the differences between the classification results of each instance and its nearest neighbors from the multi-label space in the same view and in different views, respectively. We denote a training dataset as $\mathcal{I}$. Each instance $I_i \in \mathcal{I}$ is tagged with various tags. The whole tag vocabulary for $\mathcal{I}$ forms the $S$-dimensional 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 $\mathcal{L}_r$. Obviously, the dimensionality of $\mathcal{L}_r$ is $S-1$. 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 $I_i$ and each $T_r$, we denote $y_{i,r}$ as the class label of $I_i$, where $y_{i,r} = 2 \cdot d_{i,r} - 1$. In a typical one view learning, $I_i = (\mathbf{x}_i, \mathbf{d}_i)$, where $\mathbf{x}_i$ is the feature descriptor of $I_i$ in view $\mathcal{F}$. In the two views learning, $I_i = (\mathbf{x}_i^{(a)}, \mathbf{x}_i^{(b)}, \mathbf{d}_i)$, where $\mathbf{x}_i^{(a)}$ and $\mathbf{x}_i^{(b)}$ are the feature descriptors of $I_i$ in view $\mathcal{F}^{(a)}$ and in view $\mathcal{F}^{(b)}$, respectively. Ideally, the two views $\mathcal{F}^{(a)}$ and $\mathcal{F}^{(b)}$ for $\mathcal{I}$ are conditionally independent.

### 3.1 A Novel Distance Measure in the Multi-label Space

In the discriminative classification approach we intend to learn a function $f : X \to \mathcal{R}$ which discriminates instances in the two classes. In the SVM-based methods, this function is defined as $f = \mathbf{w}^T\mathbf{x} + \hat{b}$, and the binary classification problem is solved by finding the division plane to separate the instances of the two classes. The optimization problem is presented as follows.

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i, \;\; s.t. \;\; y_i(\mathbf{w}^T\mathbf{x}_i + \hat{b}) \geq 1 - \xi_i, \;\; \xi_i \geq 0 \quad (1)$$

In general an $N$-class classification problem can always be decomposed into $N$ binary classification problems in the OVA mode. From (1) we observe that in the OVA mode, the SVM-based 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 larger the number of the tags is, the more information the tag space contains. When the tags contain noise, taking advantage of all the given tags together simultaneously shall mitigate the influence of the noise compared with only considering one tag at a time as the classification target. 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 $\mathcal{L}_r$. It is a reasonable assumption that the similarity between the classification results of two instances is inversely proportional to the distance between the instances in $\mathcal{L}_r$. The closer the instances in $\mathcal{L}_r$ are, the higher the similarity in the classification is. We denote the feature vector of $I_i$ in $\mathcal{L}_r$ as $\mathbf{t}_{i,r}$, where $\mathbf{t}_{i,r} = (d_{i,1}, \ldots, d_{i,r-1}, d_{i,r+1}, \ldots, d_{i,S})'$. However, measuring the distance for instances $I_i$ and $I_j$ in $\mathcal{L}_r$ by using $||\mathbf{t}_{i,r} - \mathbf{t}_{j,r}||_p$ directly is unreasonable in most cases, for it is based on the tag independence assumption, which is violated due to the potential existence of the relationships among the tags. Actually in the real-world scenarios, there may be various relationships among the tags, e.g., some tags may co-occur frequently, while other tags may never co-occur.

We discuss the relationship between $T_r$ and $T_k$ $(k \in \{1, \ldots, r-1, r+1, \ldots, S\})$ by examining the effect of $|d_{i,k} - d_{j,k}|$ $(\forall I_i, \; I_j \in \mathcal{I})$ on the distance between $I_i$ and $I_j$ in $\mathcal{L}_r$. When $|d_{i,k} - d_{j,k}| = 0$, the effect of $|d_{i,k} - d_{j,k}|$ on the distance between $I_i$ and $I_j$ in $\mathcal{L}_r$ is always zero; when $|d_{i,k} - d_{j,k}| = 1$, the effect of $|d_{i,k} - d_{j,k}|$ on the distance between $I_i$ and $I_j$ in $\mathcal{L}_r$ varies depending on the association degree between $T_r$ and $T_k$. We describe the relationship between $|d_{i,k} - d_{j,k}| = 1$ and the value of $|d_{i,r} - d_{j,r}|$

as follows.

$$\forall I_i,\ I_j \in \mathcal{I}:\ \ |d_{i,k} - d_{j,k}| = 1 \Rightarrow |d_{i,r} - d_{j,r}| = 1,$$
$$\text{when } \forall I_i \in \mathcal{I}:\ d_{i,k} = 0 \Rightarrow d_{i,r} = 0, \text{ and } d_{i,k} = 1 \Rightarrow d_{i,r} = 1 \quad (2)$$
$$\text{or } \forall I_i \in \mathcal{I}:\ d_{i,k} = 0 \Rightarrow d_{i,r} = 1, \text{ and } d_{i,k} = 1 \Rightarrow d_{i,r} = 0 \quad (3)$$

$$\forall I_i,\ I_j \in \mathcal{I}:\ \ |d_{i,k} - d_{j,k}| = 1 \Rightarrow |d_{i,r} - d_{j,r}| = 0,$$
$$\text{when } \forall I_i \in \mathcal{I}:\ d_{i,k} = 0 \Rightarrow d_{i,r} = 0, \text{ and } d_{i,k} = 1 \Rightarrow d_{i,r} = 0 \quad (4)$$
$$\text{or } \forall I_i \in \mathcal{I}:\ d_{i,k} = 0 \Rightarrow d_{i,r} = 1, \text{ and } d_{i,k} = 1 \Rightarrow d_{i,r} = 1 \quad (5)$$

We define $\mathcal{V}_r = \{I_i | d_{i,r} = 1\}$ ($I_i \in \mathcal{I}$ and $r \in \{1, 2, \ldots, S\}$). Formulas (2-5) describe four special relationships between $T_r$ and $T_k$. In reality, when $T_r$ is distributed evenly in $\mathcal{V}_k$ and $\mathcal{I} - \mathcal{V}_k$, $T_k$ is an undistinguished tag for $T_r$; when $T_r$ is distributed unevenly in $\mathcal{V}_k$ and $\mathcal{I} - \mathcal{V}_k$, $T_k$ is a distinguished tag for $T_r$. We show the conditional probabilities of $d_{i,r} = 0$ or 1 given $d_{i,k} = 0$ or 1 as follows, respectively.

$$\begin{aligned} P_{00} &\triangleq P(d_{i,r} = 0 | d_{i,k} = 0) = \frac{|(\mathcal{I} - \mathcal{V}_r) \cap (\mathcal{I} - \mathcal{V}_k)|}{|\mathcal{I} - \mathcal{V}_k|} \\ P_{11} &\triangleq P(d_{i,r} = 1 | d_{i,k} = 1) = \frac{|\mathcal{V}_r \cap \mathcal{V}_k|}{|\mathcal{V}_k|} \\ P_{10} &\triangleq P(d_{i,r} = 1 | d_{i,k} = 0) = 1 - P_{00} \\ P_{01} &\triangleq P(d_{i,r} = 0 | d_{i,k} = 1) = 1 - P_{11} \end{aligned} \quad (6)$$

From (2-6) we observe that, when the value of $P_{00} \cdot P_{11}$ or $P_{10} \cdot P_{01}$ is larger, the possibility that $|d_{i,k} - d_{j,k}| = 1$ leads to $|d_{i,r} - d_{j,r}| = 1$ is larger, and consequently the effect of $|d_{i,k} - d_{j,k}|$ on the distance between $I_i$ and $I_j$ in $\mathcal{L}_r$ is larger. When the value of $P_{00} \cdot P_{01}$ or $P_{10} \cdot P_{11}$ is larger, the possibility that $|d_{i,k} - d_{j,k}| = 1$ leads to $|d_{i,r} - d_{j,r}| = 0$ is larger, and consequently the effect of $|d_{i,k} - d_{j,k}|$ on the distance between $I_i$ and $I_j$ in $\mathcal{L}_r$ is smaller. We also note $P_{00} \cdot P_{11} + P_{10} \cdot P_{01} + P_{00} \cdot P_{01} + P_{10} \cdot P_{11} = 1$.

We denote the association degree vector for each tag $T_r$ as $\mathbf{g}_r$, where $\mathbf{g}_r = (g_{r,1}, \ldots, g_{r,r-1}, g_{r,r+1}, \ldots, g_{r,S})'$, and the elements of $\mathbf{g}_r$ are the association degrees between tag $T_r$ and the other tags. Hence, we define $g_{r,k}$ ($k \in \{1, \ldots, r-1, r+1, \ldots, S\}$) as follows:

$$g_{r,k} = P_{00} \cdot P_{11} + P_{10} \cdot P_{01} \quad (7)$$

Combining the feature vectors of the instances in $\mathcal{L}_r$ with the association degree vector for $T_r$, we obtain the distance between $I_i$ and $I_j$ in $\mathcal{L}_r$ as follows:

$$dis_r(I_i, I_j) = ||(\mathbf{t}_{i,r} - \mathbf{t}_{j,r}) \odot \mathbf{g}_r||_p \quad (8)$$

where $\odot$ indicates the Hadamard product between two vectors. The neighborhood of $I_i$ in $\mathcal{L}_r$ (not including $I_i$ itself) based on the distance measure in formula (8) is denoted as $\mathcal{N}_r(I_i)$. The size $u$ of the neighborhood $\mathcal{N}_r(I_i)$ for each $I_i$ is defined as the count of the nearest neighbors of $I_i$ in $\mathcal{L}_r$. $\mathcal{N}_i^r \triangleq \{j | I_j \in \mathcal{N}_r(I_i)\}$.

## 3.2 SVM with Multi-label Soft Membership

In a typical one view learning, we assume that the data in the target dataset are of the following formulation: $I_i = \{\mathbf{x}_i, y_{i,r}\}_{i=1}^n$, where $y_{i,r} \in \{-1, +1\}$ is the hard class label of $I_i$ for the tag $T_r$. Based on the assumption that the closer the instances in $\mathcal{L}_r$ are, the higher the similarity in the classification is, the information contained in the multi-label space can be added into SVM by introducing the multi-label soft membership to change the class label of $I_i$ from $\{-1, +1\}$ to $[-1, +1]$. We denote $l_{i,r}$ as the multi-label soft membership for $I_i$, where the value of $l_{i,r}$ not only depends on the hard class label of $I_i$, but also depends on the hard class labels of the nearest neighbors of $I_i$ in $\mathcal{L}_r$. With the multi-label soft membership, the typical SVM constraint is changed as follows to take advantage of the information contained in the multi-label space.

$$\forall_{i=1}^n:\ \ \text{sgn}(l_{i,r}) \cdot (\mathbf{w}^T \mathbf{x}_i + \hat{b}) \geq |l_{i,r}| - \xi_i \quad (9)$$

$$l_{i,r} = \frac{y_{i,r} + D \cdot \sum_{j \in \mathcal{N}_i^r} y_{j,r} / e^{dis_r(I_i, I_j)}}{1 + \sum_{j \in \mathcal{N}_i^r} D / e^{dis_r(I_i, I_j)}} \quad (10)$$

where $D$ is a constant, and $0 \leq D < 1$. Hence, we obtain the following optimization:

$$\begin{aligned} &\min_{\mathbf{w}}\ \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^n |l_{i,r}| \xi_i \\ &s.t.\ \forall_{i=1}^n:\ l_{i,r}(\mathbf{w}^T \mathbf{x}_i + \hat{b}) \geq |l_{i,r}|^2 - |l_{i,r}| \xi_i,\ \ \xi_i \geq 0 \end{aligned} \quad (11)$$

## 3.3 SVM with Multi-label Constraints

Another method to take advantage of the information contained in the multi-label space is to introduce the multi-label constraints which minimize the difference between the classification results of each instance and its nearest neighbors in $\mathcal{L}_r$.

$$\forall_{i=1}^n \ and\ \forall j \in \mathcal{N}_i^r:\ \ |\mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \mathbf{x}_j| \leq \eta_{ij},\ \ \eta_{ij} \geq 0 \quad (12)$$

Combining these multi-label constraints with the typical SVM constraints and allowing different regularization constants, we obtain the following optimization:

$$\begin{aligned} &\min_{\mathbf{w}}\ \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i + \sum_{i=1}^n \sum_{j \in \mathcal{N}_i^r} \frac{C^*}{e^{dis_r(I_i, I_j)}} \cdot \eta_{ij} \\ &s.t.\ \forall_{i=1}^n:\ \ y_{i,r}(\mathbf{w}^T \mathbf{x}_i + \hat{b}) \geq 1 - \xi_i,\ \ \xi_i \geq 0 \\ &\quad \forall_{i=1}^n \ and\ \forall j \in \mathcal{N}_i^r:\ \ |\mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \mathbf{x}_j| \leq \eta_{ij},\ \ \eta_{ij} \geq 0 \end{aligned} \quad (13)$$

where $C$ and $C^*$ are constants, and $C^* < C$.

## 3.4 Multi-label SVM in Two Views

In the two views learning, we assume that the data in the target dataset are of the following formulation: $I_i = \{\mathbf{x}_i^{(a)}, \mathbf{x}_i^{(b)}, y_{i,r}\}_{i=1}^n$, where $y_{i,r} \in \{-1, +1\}$ is the hard class label of $I_i$ for the tag $T_r$. Additional constraints are introduced in the two views learning to maximize the similarity between the classification results of the same instances in the two views. These two views constraints are presented as follows:

$$\forall_{i=1}^n:\ \ |\mathbf{w}^{(a)T} \mathbf{x}_i^{(a)} + \hat{b}^{(a)} - \mathbf{w}^{(b)T} \mathbf{x}_i^{(b)} - \hat{b}^{(b)}| \leq \eta_i,\ \ \eta_i \geq 0 \quad (14)$$

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.

Ideally, we can take advantage of the information contained in the multi-label space by introducing the following multi-label constraints which minimize the differences between the classification results of each instance and its nearest neighbors in the same view and in different views, respectively.

$$\forall_{i=1}^n \ and\ \forall j \in \mathcal{N}_i^r:$$
$$|\mathbf{w}^{(a)T} \mathbf{x}_i^{(a)} - \mathbf{w}^{(a)T} \mathbf{x}_j^{(a)}| \leq \eta_{ij}^{(aa)},\ \ \eta_{ij}^{(aa)} \geq 0 \quad (15)$$
$$|\mathbf{w}^{(b)T} \mathbf{x}_i^{(b)} - \mathbf{w}^{(b)T} \mathbf{x}_j^{(b)}| \leq \eta_{ij}^{(bb)},\ \ \eta_{ij}^{(bb)} \geq 0 \quad (16)$$
$$|\mathbf{w}^{(a)T} \mathbf{x}_i^{(a)} + \hat{b}^{(a)} - \mathbf{w}^{(b)T} \mathbf{x}_j^{(b)} - \hat{b}^{(b)}| \leq \eta_{ij}^{(ab)},\ \ \eta_{ij}^{(ab)} \geq 0 \quad (17)$$
$$|\mathbf{w}^{(b)T} \mathbf{x}_i^{(b)} + \hat{b}^{(b)} - \mathbf{w}^{(a)T} \mathbf{x}_j^{(a)} - \hat{b}^{(a)}| \leq \eta_{ij}^{(ba)},\ \ \eta_{ij}^{(ba)} \geq 0 \quad (18)$$

However, adding all these multi-label constraints would increase the dimensionality of the parameters, as well as the computational complexity substantially. It is easy to prove that when given one of these constraints (15-18) with the multi-view constraint (14), the other three multi-label constraints can either be strictly obtained or be approximately obtained with a little larger constraint variable value.

Hence, in order to decrease the dimensionality of the parameters, as well as the computational complexity, without raising a large bias corresponding to the solution to the optimization problem with the ideal constraints (15-18), we utilize the multi-label soft membership introduced in Section 3.2 instead of the constraints (15) and (16) to minimize the differences between the classification results

of each instance and its nearest neighbors in the same view. Further, we only select one of the constraints (17) and (18) to minimize the differences between the classification results of each instance and its nearest neighbors in different views. We denote $\mathcal{N}_r(I_i)$ and $\mathcal{N}_i^r$ as $\mathcal{N}_r(I_i) \cup \{I_i\}$ and $\mathcal{N}_i^r \cup \{i\}$, respectively. Hence, we obtain the optimization of MSVM-2K as follows:

$$\min_{\mathbf{w}^{(a)},\mathbf{w}^{(b)}} \frac{1}{2}\|\mathbf{w}^{(a)}\|^2 + \frac{1}{2}\|\mathbf{w}^{(b)}\|^2 + C^{(a)}\sum_{i=1}^{n}|l_{i,r}|\xi_i^{(a)} + C^{(b)}\sum_{i=1}^{n}|l_{i,r}|\xi_i^{(b)} + \sum_{i=1}^{n}|l_{i,r}|\sum_{j\in\bar{\mathcal{N}}_i^r} C_{ij}\eta_{ij}^{(ab)}$$

$$C_{ij} = \begin{cases} C^{(ab)} & i = j \\ C^{(ab)*}/e^{dis_r(I_i,I_j)} & i \neq j \end{cases} \quad (19)$$

$$s.t.\ \forall_{i=1}^{n}:\ l_{i,r}(\mathbf{w}^{(a)T}\mathbf{x}_i^{(a)} + \hat{b}^{(a)}) \geq |l_{i,r}|^2 - |l_{i,r}|\xi_i^{(a)},\ \xi_i^{(a)} \geq 0$$

$$l_{i,r}(\mathbf{w}^{(b)T}\mathbf{x}_i^{(b)} + \hat{b}^{(b)}) \geq |l_{i,r}|^2 - |l_{i,r}|\xi_i^{(b)},\ \xi_i^{(b)} \geq 0$$

$$l_{i,r} = \frac{y_{i,r} + D\cdot\sum_{j\in\mathcal{N}_i^r} y_{j,r}/e^{dis_r(I_i,I_j)}}{1 + \sum_{j\in\mathcal{N}_i^r} D/e^{dis_r(I_i,I_j)}}$$

$$\forall_{i=1}^{n}\ and\ \forall j \in \bar{\mathcal{N}}_i^r:$$

$$|\mathbf{w}^{(a)T}\mathbf{x}_i^{(a)} + \hat{b}^{(a)} - \mathbf{w}^{(b)T}\mathbf{x}_j^{(b)} - \hat{b}^{(b)}| \leq \eta_{ij}^{(ab)},\ \eta_{ij}^{(ab)} \geq 0$$

where $C^{(a)}$, $C^{(b)}$, $C^{(ab)}$, $C^{(ab)*}$, and $D$ are constants, and $C^{(ab)*} < C^{(ab)}$, $0 \leq D < 1$.

# 4. EXPERIMENTS

## 4.1 Data and Parameter Setting

We apply our methods, including SVM-MSM, SVM-MC, and MSVM-2K, to the problem of mining noisy tagging with a more specific context — noisy image annotation. Two groups of comparative experiments on one view learning task and on two views learning task are conducted to evaluate the performances of our methods, respectively.

The NUS-WIDE [1] image database is used in the experiments. It includes 269, 648 web images and 81 concepts which we treat as the ground truth tags. We choose the top 75 concepts whose numbers of positive examples are larger than 350 from the database to form the multi-label space $\mathcal{T}$. Hence, the dimensionality of the additional feature space of tags ($\mathcal{L}_r$) for each $T_r$ 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. In the experiments, $s\%$ noise is added into both of the positive and negative examples of the perfectly tagged training set for each concept to form the noisily tagged training set. On one view learning, the 500-D bag of words feature based on SIFT descriptions is used as the feature of view $\mathcal{F}$. On two views learning, 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 $\mathcal{F}^{(b)}$. The returned value of the classifier in one view is denoted as $f_i$, $f_i = \mathbf{w}^T\mathbf{x}_i + \hat{b}$. The returned value of the classifiers in two views combined is denoted as $f_i^{(ab)}$, $f_i^{(ab)} = 0.5\cdot(\mathbf{w}^{(a)T}\mathbf{x}_i^{(a)} + \hat{b}^{(a)} + \mathbf{w}^{(b)T}\mathbf{x}_i^{(b)} + \hat{b}^{(b)})$.

As we described before, the size $u$ of the neighborhood $\mathcal{N}_r(I_i)$ for each $I_i$ is the count of the nearest neighbors of $I_i$ in $\mathcal{L}_r$. When $u = 0$, SVM-MSM and SVM-MC are both reduced to SVM, and MSVM-2K is reduced to SVM-2K [2]. We define $R$ as the ratio between the regularization constants for instances and the regularization constants for their nearest neighbors. For SVM-MSM, $R = D$; for SVM-MC, $R = C^*/C$; for MSVM-2K, $R_1 = D$, $R_2 = 2\cdot C^{(ab)}/(C^{(a)} + C^{(b)})$, and $R_3 = C^{(ab)*}/C^{(ab)}$. In the experiments, we set $R_1 = R_2 = R_3 = R$ for MSVM-2K.

**Table 1: The top 8 tags ($T_k$) with the highest association degree $g_{r,k}$ for $T_r$.**

| Tag $T_r$ | The top 8 tags ($T_k$) with the highest association degree $g_{r,k}$ for $T_r$ |
|---|---|
| boats | harbor, reflection, lake, vehicle, ocean, water, town, sunset |
| buildings | cityscape, town, castle, house, nighttime, tower, street, window |
| clouds | sunset, sun, rainbow, sky, valley, lake, mountain, beach |
| coral | fish, whales, swimmers, ocean, animal, water, beach, leaf |
| flowers | plants, leaf, garden, grass, frost, coral, wedding, animal |
| horses | running, animal, grass, police, zebra, sand, cow, sports |
| military | plane, airport, police, vehicle, fire, road, clouds, harbor |
| mountain | valley, glacier, rocks, snow, rainbow, lake, waterfall, reflection |
| reflection | harbor, lake, boats, water, bridge, sunset, cityscape, valley |
| running | sports, horses, dog, zebra, elk, animal, beach, sand |
| sun | sunset, moon, ocean, lake, beach, reflection, harbor, tree |
| window | town, cars, vehicle, house, street, buildings, train, castle |

**Table 2: The F1 measure for the testing set with the noisily tagged training set.**

(a) The $F1^a\backslash F1^i$ for the testing set on one view learning.

| | $F1^a\backslash F1^i$ *in View* $\mathcal{F}$ | | | |
|---|---|---|---|---|
| | $s = 40$ | $s = 30$ | $s = 20$ | $s = 10$ |
| SVM | 0.5665\0.5709 | 0.6356\0.6387 | 0.6693\0.6719 | 0.7074\0.7087 |
| Fuzzy SVM | 0.5743\0.5801 | 0.6462\0.6480 | 0.6778\0.6802 | 0.7172\0.7193 |
| SVM-HF | 0.5770\0.5858 | 0.6388\0.6441 | 0.6742\0.6778 | 0.7065\0.7084 |
| SVM-MSM | 0.5969\0.6133 | **0.6785\0.6827** | 0.7018\0.7054 | **0.7327\0.7321** |
| SVM-MC | **0.6135\0.6291** | 0.6712\0.6790 | **0.7039\0.7097** | 0.7243\0.7248 |

(b) The $F1^a\backslash F1^i$ for the testing set on two views learning.

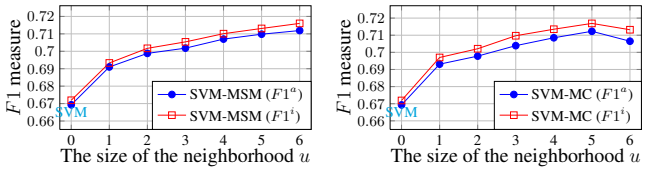
| | $F1^a\backslash F1^i$ *in two views* ($\mathcal{F}^{(a)}$ *and* $\mathcal{F}^{(b)}$) *combined* | | | |
|---|---|---|---|---|
| | $s = 40$ | $s = 30$ | $s = 20$ | $s = 10$ |
| SVM | 0.6468\0.6533 | 0.7557\0.7582 | 0.8062\0.8088 | 0.8505\0.8505 |
| Fuzzy SVM | 0.6539\0.6581 | 0.7619\0.7629 | 0.8187\0.8195 | 0.8556\0.8559 |
| SVM-HF | 0.6478\0.6538 | 0.7590\0.7616 | 0.8098\0.8027 | 0.8536\0.8541 |
| Co-training | 0.6570\0.6624 | 0.7642\0.7667 | 0.8137\0.8158 | 0.8602\0.8603 |
| SVM-2K | 0.6480\0.6534 | 0.7609\0.7636 | 0.8080\0.8103 | 0.8511\0.8509 |
| MSVM-2K | **0.6798\0.6896** | **0.7914\0.7932** | **0.8410\0.8403** | **0.8647\0.8627** |

## 4.2 Results and Discussions

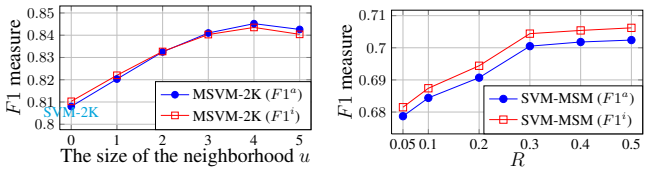
We evaluate the performances of the methods using the standard performance measures of Macro-$F1$ ($F1^a$) and Micro-$F1$ ($F1^i$). Macro-$F1$ averages the $F1$ measures on the predictions of different tags; Micro-$F1$ computes the $F1$ measure on the predictions of different labels as a whole.

We randomly select 12 exemplar tags from the 75 concepts as the target tag $T_r$, and describe the top 8 tags ($T_k$) with the highest association degree $g_{r,k}$ for each $T_r$ in Table 1. From Table 1 we observe that some tags, e.g., *buildings* and *cityscape*, may co-occur frequently, while other tags, e.g., *sun* and *moon*, may never co-occur. All these relationships among the tags are utilized to determine the most distinguished tags for the given target tag. By using the novel distance measure we have proposed, the most distinguished tags are selected for each target tag $T_r$ as the most important elements of the feature to measure the distances between instances in $\mathcal{L}_r$. This proposed distance measure is more reasonable and effective than directly using all the tags as the indiscriminate elements of the feature in finding the neighborhood of each instance in $\mathcal{L}_r$.

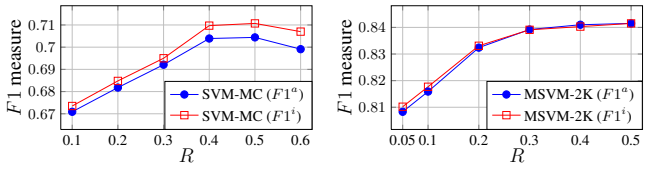
In Table 2(a), we summarize the $F1$ measure for the testing set when $u = 3$, $R = 0.4$, and $s$ is selected as 40, 30, 20, and 10 using SVM, fuzzy SVM [4], SVM-HF [3], SVM-MSM, and SVM-MC on one view learning, respectively. Table 2(b) shows the $F1$ measure for the testing set when $u = 3$, $R = 0.4$, and $s$ is selected as 40, 30, 20, and 10 using SVM, fuzzy SVM, SVM-HF, co-training, SVM-2K, and MSVM-2K on two views learning, respectively. From Table 2 we observe that SVM-MSM and SVM-MC




(a) One view learning when $s = 20$. (b) One view learning when $s = 20$.




(c) Two views learning when $s = 20$. (d) One view learning when $s = 20$.




(e) One view learning when $s = 20$. (f) Two views learning when $s = 20$.

**Figure 3: (a)(b)(c) The F1 measure for the testing set as a function of u when R = 0.4 using SVM-MSM, SVM-MC, and MSVM-2K, respectively; (d)(e)(f) the F1 measure for the testing set as a function of R when u = 3 using SVM-MSM, SVM-MC, and MSVM-2K, respectively.**

perform better than SVM, fuzzy SVM, and SVM-HF with the noisily tagged training set on one view learning, and that MSVM-2K performs better than SVM, fuzzy SVM, SVM-HF, co-training, and SVM-2K with the noisily tagged training set on two views learning when $s$ is selected as 40, 30, 20, and 10, respectively. It shows that taking advantage of all the given tags mitigates the influence of the noise compared with only considering one tag at a time as the classification target both on one view learning and on two views learning. Further, our proposed methods perform much better than the comparing methods when the noise ratio $s\%$ increases.

We describe the $F1$ measure for the testing set as a function of $u$ when $s = 20$ and $R = 0.4$ using SVM-MSM, SVM-MC, and MSVM-2K with the noisily tagged training set in Figure 3(a), (b), and (c), respectively. We observe that the $F1$ measures for the testing set all increase when the size of the neighborhood for each $\mathcal{N}_r(I_i)$ increases, which shows that it is helpful to use the nearest neighbors of each $I_i$ in $\mathcal{L}_r$ to further improve the performance of the classification. The curves for the $F1$ measures of SVM-MSM, SVM-MC, and MSVM-2K all exhibit their major elevation from $u = 0$ 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 $u$ since some instances far away may be regarded as the neighbors which may cause the decline of the $F1$ measure. As we describe before, when $u = 0$, SVM-MSM and SVM-MC are both reduced to SVM, and MSVM-2K is reduced to SVM-2K. Figure 3(a), (b), and (c) also show that the performances of SVM-MSM and SVM-MC are much better than that of SVM, and the performance of MSVM-2K is much better than that of SVM-2K.

Figure 3(d), (e), and (f) show the $F1$ measure for the testing set as a function of $R$ when $s = 20$ and $u = 3$ using SVM-MSM, SVM-MC, and MSVM-2K, respectively. As we defined before, when $R$ increases, the effect of nearest neighbors from the multi-label space in the optimization also increases. We observe that when $R$ increases, the curves for the $F1$ measures of SVM-MSM, SVM-MC, and MSVM-2K ascend, which also shows that it is helpful to use the nearest neighbors of each $I_i$ in $\mathcal{L}_r$ to mitigate the influence of the noise in the classification.

## 5. CONCLUSION

This paper studies the important problem of mining noisy tagging. We propose several effective solutions to this problem either in one view learning or in two views learning. The novelty of the proposed solutions is that they incorporate the information contained in the multi-label space into the discriminative classification methods to mitigate the influence of the noise in the classification. A novel distance measure is also proposed in this paper to compute the distance between instances in the multi-label space, which considers the various relationships among the multiple tags to obtain the most reasonable and promising neighbors for each instance. We apply the proposed solutions to the problem with a more specific context — noisy 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 in the existing literature in solving the problem of mining noisy 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

[1] 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 ACM CIVR*, pages 1–9, 2009.

[2] J. Farquhar, D. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak. Two view learning: SVM-2K, theory and practice. In *Advances in Neural Information Processing Systems*. MIT Press, 2006.

[3] S. Godbole and S. Sarawagi. Discriminative methods for multi-labeled classification. In *Proceedings of the 8th PAKDD*, 2004.

[4] C. F. Lin and S. de Wang. Training algorithms for fuzzy support vector machines with noisy data. *Pattern Recognition Letters*, 25:1647–1656, 2004.

[5] 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.

[6] J. Read, B. Pfahringer, and G. Holmes. Multi-label classification using ensembles of pruned sets. In *ICDM'08*, pages 995–1000, 2008.

[7] 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.

[8] G. Tsoumakas and I. Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. In *Proceedings of the 18th ECML*, 2007.

[9] J. Van Hulse and T. M. Khoshgoftaar. Class noise detection using frequent itemsets. *Intelligent Data Analysis*, 10:487–507, 2006.

[10] X. Zhu and X. Wu. Class noise vs. attribute noise: a quantitative study of their impacts. *Artificial Intelligence*, 22:177–210, 2004.

[11] 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.