Learning with Limited and Noisy Tagging

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

ABSTRACT
With the rapid development of social networks, tagging has become an important means responsible for such rapid development. A robust tagging method must have the capability to meet the two challenging requirements: limited labeled training samples and noisy labeled training samples. In this paper, we investigate this challenging problem of learning with limited and noisy tagging and propose a discriminative model, called SpSVM-MC, that exploits both labeled and unlabeled data through a semi-parametric regularization and takes advantage of the multi-label constraints into the optimization. While SpSVM-MC is a general method for learning with limited and noisy tagging, in the evaluations we focus on the specific application of noisy image tagging with limited labeled training samples on a benchmark dataset. Theoretical analysis and extensive evaluations in comparison with state-of-the-art literature demonstrate that SpSVM-MC outstands with a superior performance.

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 image tagging, semi-parametric regularization, multi-label constraint, limited labels, discriminative method

1. INTRODUCTION
With the rapid development of social networks, recent years have witnessed fast booming photo sharing websites, such as Flickr and Picasa. These websites allow users to upload personal images and describe the image contents with tags. In particular, appropriate tags allow users to conveniently organize and access to those shared images. Thus, tagging is considered as an important means for the development of multimedia social networks. However, with the typically explosive amount of data to be tagged, it is very time-consuming and labor-intensive for manually labeling data. Consequently, machine learning techniques for image tagging have become an effective alternative attracting much attention in the literature. In order to reduce the manual effort, many semi-supervised learning or active learning methods have been developed. Two requirements are then immediately imposed for a robust image-tagging method.

First, since it is expensive to obtain labeled training samples, it is expected that a robust image tagging method need only a small portion of the labeled training samples. Second, even with the manual labeling of images, there is no guarantee that the provided labels are always correct as human may make mistakes, not even speaking that in many applications the labeled training samples are generated by machine from other sources such as the surrounding text in which case the given labels are very error-prone. On the other hand, even with manual labeling, it is always the case that given labels are not complete to describe everything in an image as human beings typically do not have the patience to give labels for all the details of an image. Both scenarios contribute to the typically noisy labeled training samples where noisy samples mean either samples are incomplete and/or samples are error-prone.

This paper is motivated to address both such requirements for developing a robust tagging method making use of the unlabeled data through a semi-parametric regularization. The geometric structure of the unlabeled data is encoded through a specific family of parametric functions. In particular, we consider the case where the unlabeled data is encoded through a specific family of parametric functions. In particular, we consider the case where the unlabeled data is encoded through a specific family of parametric functions. In particular, we consider the case where the unlabeled data is encoded through a specific family of parametric functions.
Figure 2: Exemplar images with multiple tags

labeled data are expected to determine the marginal distribution of the data if there is a small set of labeled data available along with a relatively large set of unlabeled data. Thus, we must consider the geometric structure of the marginal distribution of the whole data including the labeled data and the unlabeled data. An example is shown in Figure 1 for the binary classification problem. Moreover, in this paper we obtain these parametric functions by applying the Kernel Principal Component Analysis (KPCA) algorithm [15] to the whole data. Since in KPCA the kernel functions in the Reproducing Kernel Hilbert Space (RKHS) are used to extract important principal components, the geometric structure of the marginal distribution of the data can be obtained by the learned parametric functions. We extend the original RKHS to a discriminative method by using these parametric functions learned from the whole data. Specifically, this extension is called semi-parametric regularized discriminative method.

Furthermore, we incorporate the information from the multi-label space into the semi-parametric regularized discriminative method to make it adaptive to the noisy data. In general the 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, the majority 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 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 tags contain noise, 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. In general, it is believed that most tags are correct, although there are many incorrect and missing tags. Thus, these tags can provide additional information to help improve the training accuracy even when instances may be incorrectly labeled in the OVA mode. Figure 2 shows exemplar images with multiple tags. The images in the first row in Figure 2, 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 2, 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 learning with limited and noisy tagging not only exploiting the unlabeled data based on semi-parametric regularization but also denoising the noisy tagging by incorporating the multi-label space information. Even though there is recent literature to address the semi-supervised learning and noisy tagging separately, combining them as a single, overall problem and developing a synergistic solution to address them together in a single framework are still significant and challenging. We first propose a semi-parametric regularized discriminative method to encode the geometric structure of labeled and unlabeled data. Then we propose a distance measure to compute the distance between instances in the multi-label space. Consequently, a new framework is proposed to incorporate the information of the multi-label space into the semi-parametric regularized discriminative classification method, which we call Semi-parametric regularized Support Vector Machine with Multi-label Constraint (SpSVM-MC). While SpSVM-MC is a general method, we demonstrate through extensive evaluations in the application of image tagging using real data that the proposed method performs well in comparison with the peer methods in the literature as an effective and promising solution to the problem of learning with limited and noisy image tagging.

2. RELATED WORK

We begin by reviewing the literature on regularization, and then review the related work on denoising methods for classification.

Many machine learning algorithms including Support Vector Machine (SVM) and Regularized Least Square (RLS) can be interpreted as examples of regularization [14]. Research work [27] on sparse learning mainly exploits one norm as the regularization term. Semi-parametric regularization is also an attractive solution for many practical problems [13, 24, 2, 8]. Sun et al. [16] propose a manifold regularization based semi-supervised semi-parametric regression method. Bouboulis et al. [2] propose an image denoising method by employing a semi-parametric regularization.

Many existing semi-supervised learning methods exploit the regularization terms on the unlabeled data. Transductive SVM [20, 22] may be considered as SVM with an additional regularization term on the unlabeled data. Szummer and Jaakkola [17] propose an information regularization framework to minimize the mutual information on multiple overlapping regions covering the data space. Grandvalet and Bengio [7] use the entropy on the unlabeled data as a regularizer.

The noise among the data is categorized into two types: attribute noise and class noise [28]. 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 [19, 29]. For the latter category, robust algorithms are designed to reduce the impact of the noise in the classification [9, 11, 18].

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. Boutell et al. [3] introduce several transformation methods that map the multi-label 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 [25]. Hence, the original annotation problem becomes a multi-instance and multi-label learning problem. Zhou and Zhang [25] 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 SVM, which demonstrates a strong discrimination power [6, 12, 23]. Yang et al. [23] propose an asymmetrical support vector machine for region-based image annotation. There has been work on image tag refinement in the literature [21, 26, 10]. Wang et al.
[21] address the tag optimization problem given the tagging labels of images obtained by a classifier. Zhu et al. [26] propose a classification method to address the noisy tagging problem. Compared with the existing work, our work takes advantage of all the given multiple tags simultaneously not only as the classification target but also as an additional feature.

Among the existing work, the closest work to ours is the semi-parametric regularization method by Guo et al. [8]. We note that the difference from our work is significant. First, the two differ in theory. The essence of semi-parametric regularization lies in the strategy of regularized risk functional. While the regularized risk functional of [8] takes the same form as that in Chapter 4 of [14], we make a fundamental extension of the approach in [14] by adding a regularization constraint \( \|h\|_h\) to parametric functions in Eq. 6. Indeed, our semi-parametric regularization is a variation of the mixed semi-parametric regularization defined in [14]. Second, the two differ in the way of optimization. The new regularized risk functional makes our optimization problem different from the existing work and this plays a key role in the efficiency and accuracy of an algorithm.

3. LEARNING WITH LIMITED TAGGING

Since the labeled data are expensive to acquire, we utilize the information contained in the plentiful unlabeled data which better exploits the geometric structure of the data to solve the problem of learning with limited tagging. We begin with the brief review of the regularized risk functional. Then we propose a method which treats the information contained in the unlabeled data as the additional prior knowledge incorporated into the regularized risk functional to satisfy the first requirement that only a small portion of the labeled training samples are needed to learn a robust classifier.

3.1 Regularized Risk Functional

Suppose that there is a probability distribution \( F \) on \( \mathcal{X} \times \mathcal{Y} \), \( \mathcal{X} \subset \mathbb{R}^n \) according to which data are generated. We assume that the given data consist of \( l \) labeled data points \((x_i, y_i), 1 \leq i \leq l\) which are generated according to \( F \). In this paper, we assume the binary classification problem where the labels \( y_i, 1 \leq i \leq l \), are binary, i.e., \( y_i = \pm 1 \).

If we have no knowledge about the test patterns, we should minimize the expected error over all the possible training patterns. Hence we must minimize the expected loss called risk functional with respect to \( F \) and \( L \)

\[
R[f] = \int \mathcal{L}(x, y, f(x))dF(x, y)
\]

(1)

where \( \mathcal{L} \) is a loss function. A variety of loss functions have been considered in the literature. The simplest loss function is 0/1 loss

\[
\mathcal{L}(x_i, y_i, f(x_i)) = \begin{cases} 
0 & \text{if } y_i = f(x_i) \\
1 & \text{if } y_i \neq f(x_i)
\end{cases}
\]

(2)

For the loss function Eq. 2, Eq. 1 determines the probability of a classification error for any decision function \( f \). Unfortunately, in most applications the probability distribution \( F \) is unknown. All we have at our disposal is the actual training data \((x_i, y_i), 1 \leq i \leq l\). What one usually does is to consider the empirical estimate of the risk functional [20], which is defined as

\[
R_{emp}[f] = \frac{1}{l} \sum_{i=1}^{l} \mathcal{L}(x_i, y_i, f(x_i))
\]

(3)

Minimizing the empirical risk Eq. 3 may lead to numerical instabilities and a bad generalization performance [14]. A possible solution is to add a stabilization (regularization) term \( \Omega[f] \) to the empirical risk functional. This leads to a better conditioning of the problem. Thus, we consider the following class of regularized risk functionals

\[
R_{reg}[f] = R_{emp}[f] + \gamma \Omega[f]
\]

where \( \gamma > 0 \) is the regularization parameter which specifies the tradeoff between the minimization of \( R_{emp}[f] \) and the smoothness or simplicity enforced by a small \( \Omega[f] \). When we equivalently think of the feature space as a reproducing kernel Hilbert space, \( \Omega[f] \) is the norm of the RKHS representation of the feature space

\[
\Omega[f] = \|f\|_K
\]

where \( \| \cdot \|_K \) is the norm in the RKHS \( \mathcal{H}_K \) associated with the kernel \( K \). In this case, we equivalently minimize

\[
f^* = \arg \min_{f \in \mathcal{H}_K} \sum_{i=1}^{l} \mathcal{L}(x_i, y_i, f(x_i)) + \gamma \|f\|_K^2
\]

(4)

The solution to Eq. 4 is determined by the loss function \( L \) and the kernel \( K \). The following classic Representer Theorem [14] states that the solution to the minimization problem in Eq. 4 exists in \( \mathcal{H}_K \) and gives the explicit form of a minimizer.

**Theorem 1.** Denote by \( \Omega : [0, \infty) \to \mathbb{R} \) a strictly monotonically increasing function, by \( \mathcal{X} \) a set, and by \( \Lambda : (\mathcal{X} \times \mathbb{R}^2)^l \to \mathbb{R} \cup \{\infty\} \) an arbitrary loss function. Then each minimizer \( f \in \mathcal{H}_K \) of the regularized risk

\[
\Lambda((x_1, y_1, f(x_1)), \ldots, (x_l, y_l, f(x_l))) + \Omega(\|f\|_K)
\]

admits a representation of the form

\[
f(x) = \sum_{i=1}^{l} \alpha_i K(x_i, x)
\]

(5)

with \( \alpha_i \in \mathbb{R} \).

According to Theorem 1, we may use regularizers other than \( \Omega(\|f\|_K) = \gamma \|f\|_K^2 \) which is a strictly monotonic increasing function of \( \|f\|_K \). This allows us in principle to design algorithms that are more closely aligned with the recommendations given by the bounds derived from the statistical learning theory. Moreover, given the loss function \( L \) and the kernel \( K \), if we use the regularizer \( \Omega(\|f\|_K) = \gamma \|f\|_K^2 \), we could substitute Eq. 5 into Eq. 4 to obtain a minimization problem of the variables \( \alpha_i, 1 \leq i \leq l \).

3.2 Semi-parametric Regularization Learning

In this section, we first propose a semi-parametric regularization approach by including a family of parametric functions. Then we demonstrate how to make use of the whole data including labeled and unlabeled data to obtain the parametric functions. Based on the semi-parametric regularization, we propose a novel semi-parametric regularized SVM to solve the problem of learning with limited tagging.

3.2.1 Semi-parametric Regularization

In the case that the training set only contains a small portion of the labeled samples, the plentiful unlabeled samples may contain some additional knowledge beyond the knowledge which the labeled ones contain. This additional knowledge can be taken advantage of to better exploit the geometric structure of the data. One method to utilize this additional knowledge is to introduce some parametric functions added into the regularized risk functional to refine the decision function.
Suppose that this additional prior knowledge is described as a family of parametric functions \( \{ \psi_p \}_{p=1}^{M} : \mathcal{X} \rightarrow \mathbb{R} \). These parametric functions may be incorporated into the regularized risk functional in Eq. 4 in different ways. In this paper, we assume that \( \text{span}\{ \psi_p \} \triangleq \Psi \subset \mathcal{H}_K \) has the norm \( \| \cdot \|_\psi \), in which \( \mathcal{H}_K \) is the complementary space of \( \mathcal{H}_K \), and minimize the following regularized risk functional

\[
\hat{f}^* = \arg \min_{f} \sum_{i=1}^{l} L(x_i, y_i, \tilde{f}(x_i)) + \gamma_1 \| f \|_K^2 + \gamma_2 \| h \|_\psi^2
\]

where \( \tilde{f} \triangleq f + h \) with \( f \in \mathcal{H}_K \) and \( h \in \Psi \). Consequently, we extend the original RKHS \( \mathcal{H}_K \) by including a family of parametric functions \( \psi_p \). The following semiparametric representer theorem is an immediate extension of Theorem 1, that can be easily proven.

**Theorem 2.** Denote by \( \Omega : [0, \infty) \times [0, \infty) \rightarrow \mathbb{R} \) a strictly monotonically increasing function, by \( X \) a set, and by \( \Lambda : \{X \times \mathbb{R}^2 \} \rightarrow \mathbb{R} \cup \{\infty\} \) an arbitrary loss function. Let \( \{ \psi_p \}_{p=1}^{M} : \mathcal{X} \rightarrow \mathbb{R} \) be a set of \( M \) real valued functions with the property that the \( l \times l \) matrix \( \psi_p(x) \) has rank \( M \) and span \( \{ \psi_p \} \triangleq \Psi \subset \mathcal{H}_K \) has the norm \( \| \cdot \|_\psi \). Then for any \( f \neq \tilde{f} + h \) with \( f \in \mathcal{H}_K \) and \( h \in \Psi \), minimizing the regularized risk

\[
\Lambda((x_1, y_1, \tilde{f}(x_1), \ldots, (x_l, y_l, \tilde{f}(x_l))) + \Omega(\| f \|_K, \| h \|_\psi)
\]

admits a representation of the form

\[
\tilde{f}(x) = \sum_{i=1}^{l} \alpha_i K(x_i, x) + \sum_{p=1}^{M} \beta_p \psi_p(x)
\]

with \( \alpha_i, \beta_p \in \mathbb{R} \).

In Theorem 2, the parametric functions \( \{ \psi_p \}_{p=1}^{M} \) can be any functions. In this paper, the parametric function is learned from the whole data including the labeled and unlabeled data to incorporate the information contained in the unlabeled data. The details of the method to learn the parametric functions are discussed in Section 3.2.2. We denote by \( \psi_p(x) \) the parametric function and by \( \beta_p \) the corresponding coefficient. Thus, the minimizer of Eq. 6 is

\[
\hat{f}^*(x) = \sum_{i=1}^{l} \alpha_i^* K(x_i, x) + \sum_{p=1}^{M} \beta_p^* \psi_p(x)
\]

where \( K \) is the kernel in the original RKHS \( \mathcal{H}_K \).

### 3.2.2 Learning Parametric Functions

\( \{ \psi_p(x) \}_{p=1}^{M} \) is obtained by applying the KPCA algorithm [15] to the whole data set including \( l \) labeled data points \( (x_i, y_i), 1 \leq i \leq l \) and \( v \) unlabeled data points \( x_i, l + 1 \leq i \leq l + v \). In other words, we consider the unlabeled data points as the additional prior knowledge to obtain the parametric functions, which make our approach different from a general semi-supervised learning algorithm. KPCA finds the principal axes in the feature space which carry more variance than any other directions by diagonalizing the covariance matrix

\[
C = \frac{1}{l+v} \sum_{j=1}^{l+v} \phi(x_j) \phi(x_j)^T
\]

where \( \phi \) is a mapping function in the RKHS. To find the principal axes, we solve the eigenvalue problem, \( C U = U \Xi \). Let \( \Xi \) denote the \( M \) largest eigenvalues and \( U \) the corresponding eigenvector space. Substituting Eq. 9 to this equation, we note that all solutions

**Figure 3:** (a) Input points before kernel PCA; (b) Output after kernel PCA, with Gaussian kernel.

U lie in the span of \( \phi(x_1), \ldots, \phi(x_{l+v}) \). There exist coefficients \( \theta_1, \ldots, \theta_{l+v} \), such that

\[
U = \sum_{i=1}^{l+v} \theta_i \phi(x_i)
\]

Given the data point \( x \), the projection onto the principal axes is given by \( \phi(x)^T U \). Consequently, we let \( \psi(x) = \phi(x)^T U \). Fig. 3 shows an illustrative example for a classification problem in which we wish to use KPCA to identify three concentric clouds of points. As shown in this example, the first principal component is sufficient to distinguish the different groups. From this example, it is clear that the data points projected onto the most important principal axis still keep the original neighborhood relationship. In general, after projection on the principal axes, similar data points stay close and dissimilar data points are kept far away from each other. In the proposed model, we choose the top \( M \) largest eigenvalues and the corresponding eigenvectors from the results of KPCA to learn the parametric function for the training data.

### 3.2.3 Semi-parametric Regularized Support Vector Machine

Based on the semi-parametric regularization, a semi-parametric regularized SVM is proposed to solve the problem of learning with limited tagging. We first review the SVM approach to the binary classification problem which is the focus of this paper. In the binary classification problem, the classic SVM attempts to solve the following optimization problem on the labeled data.

\[
\min \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} \xi_i
\]

\[
s.t. \ y_i (\langle w, \phi(x_i) \rangle + b) \geq 1 - \xi_i
\]

\[
\xi_i \geq 0 \ i = 1, \ldots, l
\]

where \( \phi \) is a nonlinear mapping function determined by the kernel and \( b \) is a regularized term.

Again, the solution is given by

\[
\hat{f}^*(x) = \langle w^*, \phi(x) \rangle + b^* = \sum_{i=1}^{l} \alpha_i^* K(x_i, x) + b^*
\]

To solve Eq. 11 we introduce one Lagrange multiplier for each constraint in Eq. 11 using the Lagrange multipliers technique and
obtain a quadratic dual problem of the Lagrange multipliers.

\[
\min \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \mu_i \mu_j K(x_i, x_j) - \sum_{i=1}^{l} \mu_i \\
\text{s.t. } \sum_{i=1}^{l} \mu_i y_i = 0 \\
0 \leq \mu_i \leq C, \ i = 1, \ldots, l
\]  

(12)

where \(\mu_i\) is the Lagrange multiplier.

We have \(w^* = \sum_{i=1}^{l} \mu_i y_i \phi(x_i)\) from the solution to Eq. 12.

Note that the following conditions must be satisfied according to the Karush-Kuhn-Tucker theorem [20]:

\[
\mu_i (y_i (\langle w, \phi(x_i) \rangle + b) + \xi_i - 1) = 0, \ i = 1, \ldots, l
\]  

(13)

The optimal solution of \(b\) is determined by the above conditions. Therefore, the solution is given by

\[
f^*(x) = \sum_{i=1}^{l} \alpha_i^* K(x_i, x) + b^*
\]

where \(\alpha_i^* = \mu_i y_i\).

The semi-parametric regularized SVM algorithm solves the optimization problem in Eq. 6:

\[
\min \frac{1}{2} \|w\|^2 + \frac{1}{2} \sum_{p=1}^{M} \beta_p^2 + C \sum_{i=1}^{l} \xi_i \\
\text{s.t. } y_i (\langle w, \phi(x_i) \rangle + b + \sum_{p=1}^{M} \beta_p \psi_p(x_i)) \geq 1 - \xi_i \\
\xi_i \geq 0, \ i = 1, \ldots, l
\]  

(14)

4. LEARNING WITH LIMITED AND NOISY TAGGING

Since the noisy taggins exists in the training data in many real-world applications, we take advantage of the information contained in the multi-label space to satisfy the second requirement of utilizing the noisy tagged training data and at the same time mitigating the influence of the noise in the classification. In this section, we first propose a distance measure to compute the nearest neighbors for each instance in the multi-label space. Then we extend the semi-parametric risk functional by including the multi-label constraint. In particular, we propose a novel discriminative model called semi-parametric regularized SVM with multi-label constraint to solve the problem of learning with limited and noisy tagging.

4.1 A Distance Measure in the Multi-label Space

We denote a training dataset as \(\mathcal{I}\). Each instance \(I_i \in \mathcal{I}\) is labeled with various tags. The whole tag vocabulary for \(\mathcal{I}\) forms the \(S\)-dimensional multi-label space \(\mathcal{L}\). 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 \(d_r = (d_{r,1}, d_{r,2}, \ldots, d_{r,S})\) be the tag representation for \(I_i\), where \(d_{r,j} \in \{0, 1\}, \ 1 \leq r \leq S\) represents the occurrence of the \(r\)-th tag \(T_r\) for \(I_i\). For each \(I_i\) and each \(T_r\), we denote \(y_{r,i}\) as the class label of \(I_i\), where \(y_{r,i} = 2 \cdot d_{r,i} - 1\), \(I_i = (x_i, d_r)\), where \(x_i\) is the feature descriptor of \(I_i\).

In general an \(N\)-class classification problem can always be decomposed into \(N\) binary classification problems in the OVA mode. From Eq. 14 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.

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 \(x_{t,r}\), where \(t_{i,r} = (d_{i,1}, \ldots, d_{r-1}, d_{r+1}, \ldots, d_{S})\).

However, measuring the distance for instances \(I_i\) and \(I_j\) in \(\mathcal{L}_r\) by using \(||x_{t_i,r} - x_{t_j,r}||\) 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 \(|d_{i,k} - d_{i,k}||(\forall i, j \in \mathcal{I})\) 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.

\[
\forall I_i, I_j \in \mathcal{I}: \ |d_{i,k} - d_{j,k}| = 1 \Rightarrow |d_{i,r} - d_{j,r}| = 1,
\]

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

\[
\forall I_i, I_j \in \mathcal{I}: |d_{i,k} - d_{j,k}| = 1 \Rightarrow |d_{i,r} - d_{j,r}| = 0,
\]

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

(15)

We define \(V_r = \{I_i | d_{i,r} = 1\} (I_i \in \mathcal{I} \text{ and } r \in \{1, 2, \ldots, S\})\). Eq. 15 describes 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.

\[
P_{00} \triangleq P(d_{i,r} = 0 | d_{i,k} = 0) = \frac{|(\mathcal{I} - \mathcal{V}_k) \cap (\mathcal{I} - \mathcal{V}_k)|}{|\mathcal{I} - \mathcal{V}_k|}
\]

(16)

From Eq. 15 to Eq. 16 we observe that, when the value of \(P_{00} \cdot P_{11} + P_{01} \cdot P_{01}\) 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 \(P_{00} \cdot P_{01} + P_{11} \cdot P_{01}\) 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 \(P_{00} \cdot P_{11} + P_{01} \cdot P_{01} + P_{01} \cdot P_{01} + P_{01} \cdot P_{11} = 1\).

We denote the association degree vector for each tag \(T_r\) as \(g_r\), where \(g_r = (g_{r,1}, \ldots, g_{r,r-1}, g_{r,r+1}, \ldots, g_{r,S})\), and the elements of \(g_r\) 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.

\[
g_{r,k} = P_{00} \cdot P_{11} + P_{01} \cdot P_{01}
\]

(17)
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:

\[
dis_r(I_i, I_j) = ||(t_{i,r} - t_{j,r}) \odot g_r||_p
\] (18)

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 Eq. 18 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 \( \mathcal{L}_r \). \( \mathcal{N}_r^i \triangleq \{ j | I_j \in \mathcal{N}_r(I_i) \} \).

### 4.2 Semi-parametric Regularization with Multi-label Constraints

The proposed 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 \( \mathcal{L}_r \).

\[
\forall i \in [1, M] \text{ and } \forall j \in \mathcal{N}_r^i : \quad |\bar{f}(x_i) - \bar{f}(x_j)| \leq \eta_{ij}, \quad \eta_{ij} \geq 0
\] (19)

Furthermore, we consider this constraint as a loss function defined as

\[
L(f(x_i), \bar{f}(x_j)) = |\bar{f}(x_i) - \bar{f}(x_j)|
\] (20)

Combining these multi-label constraint with the semi-parametric regularized risk functional in Eq. 6, we have the following regularized risk functional

\[
\begin{align*}
\hat{f}^* = \arg\min_f & \sum_{i=1}^M L(x_i, y_i, \hat{f}(x_i)) + \sum_{i=1}^M \sum_{j \in \mathcal{N}_r^i} \bar{L}(\bar{f}(x_i), \bar{f}(x_j)) \\
+ \gamma_1 ||f||_K^2 + \gamma_2 ||h||_\Psi^2
\end{align*}
\] (21)

### 4.3 Semi-parametric Regularized SVM with Multi-label Constraints

In particular, in this paper we combine these multi-label constraints with the semi-parametric regularized SVM to solve the problem of learning with limited and noisy tagging. The optimization is as follows: where \( C \) and \( C^* \) are constants, and \( C^* < C \).

\[
\begin{align*}
&\min_{w,\beta,\phi,\eta} \frac{1}{2} ||w||^2 + \frac{1}{2} \sum_{p=1}^M \beta_p^2 + C \sum_{i=1}^l \xi_i + \sum_{i=1}^l \sum_{j \in \mathcal{N}_r^i} \frac{C^*}{\dis_r(l, i, j)} \cdot \eta_{ij} \\
s.t. \quad \forall i \in [1, M] : \quad &y_i - \bigg( \langle w, \phi(x_i) \rangle + \sum_{p=1}^M \beta_p \psi_p(x_i) + b \bigg) \geq 1 - \xi_i, \quad \xi_i \geq 0 \\
\forall i \in [1, M] & \text{ and } \forall j \in \mathcal{N}_r^i : \quad \eta_{ij} \geq 0,
\end{align*}
\] (22)

With the Lagrange multipliers, we have the target function

\[
\begin{align*}
L &= \frac{1}{2} ||w||^2 + \frac{1}{2} \sum_{p=1}^M \beta_p^2 + C \sum_{i=1}^l \xi_i + \sum_{i=1}^l \sum_{j \in \mathcal{N}_r^i} \frac{C^*}{\dis_r(l, i, j)} \cdot \eta_{ij} \\
- \sum_{i=1}^M \mu_i \bigg[ y_i - \bigg( \langle w, \phi(x_i) \rangle + \sum_{p=1}^M \beta_p \psi_p(x_i) + b \bigg) - 1 - \xi_i \bigg] \\
- \sum_{i=1}^l \rho_i \xi_i - \sum_{i=1}^l \sum_{j \in \mathcal{N}_r^i} \bigg[ \tau_{ij} \eta_{ij} + \lambda_{ij}^+ \bigg( \eta_{ij} - \langle w, \phi(x_i) \rangle \bigg) \\
- \sum_{p=1}^M \beta_p \psi_p(x_i) + \langle w, \phi(x_i) \rangle + \sum_{p=1}^M \beta_p \psi_p(x_j) \bigg] + \lambda_{ij}^- \eta_{ij} \\
+ \langle w, \phi(x_i) \rangle + \sum_{p=1}^M \beta_p \psi_p(x_i) - \langle w, \phi(x_j) \rangle - \sum_{p=1}^M \beta_p \psi_p(x_j) \bigg] \bigg]
\] (23)

where \( \mu_i, \lambda_{ij}^+, \lambda_{ij}^-, \tau_i, \) and \( \tau_{ij} \) are the Lagrange multipliers. We claim that if the KKT condition of Eq. 23 shown in Eq. 24 is satisfied, the solution to Eq. 22 is the same as that to its dual problem.

\[
\begin{align*}
\nabla L &= 0 \\
\nabla L^*_\omega &= 0 \\
\nabla L^*_\beta &= 0
\end{align*}
\] (24)

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

\[
\begin{align*}
\frac{\partial L}{\partial w} &= 0 \Rightarrow w = \sum_{i=1}^M \mu_i y_i \phi(x_i) - \sum_{i=1}^l \sum_{j \in \mathcal{N}_r^i} \nu_{ij} \phi(x_i) - \phi(x_j), \\
\frac{\partial L}{\partial \beta_p} &= 0 \Rightarrow \sum_{i=1}^M \mu_i y_i \phi(x_i) = \sum_{i=1}^l \sum_{j \in \mathcal{N}_r^i} \nu_{ij} \psi_p(x_i) - \psi_p(x_j), \\
\frac{\partial L}{\partial \nu_{ij}} &= 0 \Rightarrow \sum_{i=1}^M \mu_i y_i = 0, \\
\frac{\partial L}{\partial \tau_i} &= 0 \Rightarrow C = \mu_i - \tau_i = 0, \\
\frac{\partial L}{\partial \tau_{ij}} &= 0 \Rightarrow \sum_{i=1}^l \sum_{j \in \mathcal{N}_r^i} \nu_{ij} = 0
\end{align*}
\] (25)

Substituting the results from Eq. 24 and Eq. 25 to Eq. 23, we obtain the dual problem of Eq. 22 as follows:

\[
\begin{align*}
\max_{\mu, \nu} & -\frac{1}{2} \sum_{i=1}^M \sum_{j \in \mathcal{N}_r^i} \left( \mu_i y_i - \phi(x_i) - \sum_{j \in \mathcal{N}_r^i} \nu_{ij} \phi(x_j) - \phi(x_j) \right) + \sum_{i=1}^M \mu_i \\
- \frac{1}{2} \sum_{p=1}^M \sum_{i=1}^l \sum_{j \in \mathcal{N}_r^i} \left( \mu_i y_i \phi(x_i) - \sum_{j \in \mathcal{N}_r^i} \nu_{ij} \phi(x_j) - \phi(x_i) \right) \\
- \frac{1}{2} \sum_{p=1}^M \sum_{i=1}^l \sum_{j \in \mathcal{N}_r^i} \left( \mu_i y_i \phi(x_i) - \sum_{j \in \mathcal{N}_r^i} \nu_{ij} \phi(x_j) - \phi(x_i) \right)
\end{align*}
\]
tagged training set, and the left 90% of the positive and negative examples are both 50 concepts, we randomly choose 150 instances and the weight of the loss function for their nearest neighbors is defined as the ratio between the weight of the loss function for in- form the multi-label space bers in the optimization. For SpSVM-MC, terrors in the optimization. For SpSVM-MC, it includes S^VM [1] and Guo (Semi-parametric regularization SVM) [8] using the noisily tagged training images combined with the plentiful untagged images to evaluate the performances of these methods.

The NUS-WIDE [4] image database is used in the experiments. It includes 269,648 web images and 81 concepts we treat as the ground truth tags. We choose the top 75 concepts whose numbers of positive examples are larger than 200 from the database to form the multi-label space T. Hence, the dimensionality of the additional feature space of tags (L_r) for each T_r is 74. For each concept, we randomly choose 150 positive examples and 150 negative examples as the training data. In the testing set, the numbers of the positive and negative examples are both 50. 10% examples are randomly selected from the training data to form the perfectly tagged training set, and the left 90% examples from the training data are used as the untagged training set. In the experiments, s% noise is added into both the positive and negative examples of the perfectly tagged training set to form the noisily tagged training set. We denote the 500-D bag of words feature based on SIFT descriptions as V1, and denote the 1000-D bag of text words feature which describes the text information correlated to the images provided by the database as V2. Two groups of comparative experiments are conducted to evaluate the performances of these methods, which use V1 and V2 as the feature of the data, respectively.

As we described before, the size u of the neighborhood N_{i_u}(I_u) for each I_u is the count of the nearest neighbors of I_u in L_r. We define R as the ratio between the weight of the loss function for instances and the weight of the loss function for their nearest neighbors in the optimization. For SpSVM-MC, R = C^*/C. In the experiments, we choose the top M largest eigenvalues and the corresponding eigenvectors from the results of KPCA to learn the parametric function for the training data. M = 0 means that we learn no parametric function for the training data.

In addition, we design an experiment for SpSVM-MC with a different distribution on untagged training set. For this experiment, we reduce the numbers of negative samples on untagged training set to make sure that the proportion of positive and negative samples is 2:1. And we denote this experiment setup as SpSVM-MC(2:1).

5.2 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 P_{rec}, recall R_{rec}, and F1 measure are defined as

\[
\begin{align*}
\text{P}_{rec} &= \frac{CT_r}{PT_r}; \\
\text{R}_{rec} &= \frac{CT_r}{GT_r}; \\
F_1 &= \frac{2\text{P}_{rec}\text{R}_{rec}}{\text{P}_{rec} + \text{R}_{rec}} = \frac{2\sum_{r=1}^{s} CT_r}{\sum_{r=1}^{s} PT_r + \sum_{r=1}^{s} GT_r}
\end{align*}
\]

We evaluate the performances of the methods using the standard performance measures of Macro-F1 (F1^*) and Micro-F1 (F1^\dagger). 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^* = \frac{1}{S} \sum_{r=1}^{s} F1_r; \quad F1^\dagger = \frac{2}{\sum_{r=1}^{s} CT_r} \frac{\sum_{r=1}^{s} PT_r + \sum_{r=1}^{s} GT_r}{\sum_{r=1}^{s} PT_r + \sum_{r=1}^{s} GT_r}

We randomly select 12 exemplar tags from the 75 concepts as the target tag T_r, and describe the top 10 tags (T_k) with the highest association degree g_r,k for each T_r in Table 1. From Table 1 we observe that the most association tags (T_k) are different between different target tags (T_r). 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 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 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 L_r.

In Table 2, we summarize the F1 measure of the testing set when M = 4, u = 3, R = 0.2, and s is selected as 30, 25, 20, 15, 10, and 5 using feature V1 and V2 for SVM, fuzzy SVM, SVM-HF, S^VM, Guo, SpSVM-MC, and SpSVM-MC(2:1), respectively. Note the impressive and robust performance of SpSVM-MC in comparison with those of the other methods with the rather limited labeled training samples (about 10%).

On the task of learning with limited and noisy image tagging, compared with our proposed method SpSVM-MC, SVM with the noisily tagged training set considers the training set as the perfectly tagged set and mistakenly takes the incorrect tags as perfect tags. Although fuzzy SVM considers the training set as noisily tagged dataset, it cannot take advantage of all the given multiple tags simultaneously as an additional feature as SpSVM-MC does. SVM-HF extends the original dataset with |S| extra features containing the predictions of each binary classifier; then |S| new binary classifiers are trained using the extended datasets. However, SVM-HF also considers the training set as the perfectly tagged set, and fails to make use of the original given multiple labels of each instance. Further, all the comparing methods, including SVM, fuzzy SVM, and SVM-HF, fail to take advantage of the information contained in the untagged data to improve the learning performance. S^VM tries to learn low-density separators by maximizing the margin over tagged and untagged examples, which utilizes the information contained in all the training data, including the tagged ones and the untagged ones. Guo makes use of the information contained in the untagged data through a semi-parametric regularization. However, S^VM and Guo with the noisily tagged training set consider the training set as the perfectly tagged set and fail to take advantage of all the given multiple tags simultaneously as an additional feature as SpSVM-MC does. Consequently, from Table 2, we see that SpSVM-MC performs better than SVM, fuzzy SVM, SVM-HF, S^VM and Guo on the task of learning with limited and noisy...
Tables 1 and 2 present the top 10 tags and the F1 measure for testing set, respectively.

Table 1: The top 10 tags \( T_r \) with the highest association degree \( g_{r,k} \) for \( T_r \).

<table>
<thead>
<tr>
<th>Tag ( T_r )</th>
<th>The top 10 tags ( T_r ) with the highest association degree ( g_{r,k} ) for ( T_r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>boats</td>
<td>harbor, reflection, lake, vehicle, ocean, water, town, sunset, beach, cityscape</td>
</tr>
<tr>
<td>buildings</td>
<td>cityscape, town, castle, house, nighttime, tower, street, window, harbor, temple</td>
</tr>
<tr>
<td>clouds</td>
<td>sunset, sun, rainbow, sky, valley, lake, mountain, beach, ocean, harbor</td>
</tr>
<tr>
<td>coral</td>
<td>fish, whales, swimmers, ocean, animal, water, beach, leaf, lake, flowers</td>
</tr>
<tr>
<td>flowers</td>
<td>plants, leaf, garden, grass, frost, coral, wedding, animal, sky, tree</td>
</tr>
<tr>
<td>horses</td>
<td>running, animal, grass, police, zebra, sand, cow, sports, frost, snow</td>
</tr>
<tr>
<td>military</td>
<td>plane, airport, police, vehicle, fire, road, clouds, harbor, sky, person</td>
</tr>
<tr>
<td>mountain</td>
<td>valley, glacier, rocks, snow, rainbow, lake, waterfall, reflection, clouds, house</td>
</tr>
<tr>
<td>reflection</td>
<td>harbor, lake, boats, water, bridge, sunset, cityscape, valley, nighttime, mountain</td>
</tr>
<tr>
<td>running</td>
<td>sports, horses, dog, zebra, elk, animal, beach, sand, grass, swimmers</td>
</tr>
<tr>
<td>sun</td>
<td>sunset, moon, ocean, lake, beach, reflection, harbor, tree, clouds, water</td>
</tr>
<tr>
<td>window</td>
<td>town, cars, vehicle, house, street, buildings, train, castle, police, airport</td>
</tr>
</tbody>
</table>

Table 2: The F1 measure for the testing set.

(a) The \( F_1 \) for testing set using \( V_1 \) as the feature of the data.

<table>
<thead>
<tr>
<th></th>
<th>( F_1 ) ( s = 30 )</th>
<th>( F_1 ) ( s = 25 )</th>
<th>( F_1 ) ( s = 20 )</th>
<th>( F_1 ) ( s = 15 )</th>
<th>( F_1 ) ( s = 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.0529(0.5661)</td>
<td>0.0527(0.5660)</td>
<td>0.0527(0.5659)</td>
<td>0.0527(0.5658)</td>
<td>0.0521(0.5661)</td>
</tr>
<tr>
<td>Fuzzy SVM</td>
<td>0.0530(0.5661)</td>
<td>0.0527(0.5660)</td>
<td>0.0527(0.5659)</td>
<td>0.0527(0.5658)</td>
<td>0.0521(0.5661)</td>
</tr>
<tr>
<td>Guo</td>
<td>0.0530(0.5661)</td>
<td>0.0527(0.5660)</td>
<td>0.0527(0.5659)</td>
<td>0.0527(0.5658)</td>
<td>0.0521(0.5661)</td>
</tr>
<tr>
<td>S^VM</td>
<td>0.0540(0.5662)</td>
<td>0.0527(0.5660)</td>
<td>0.0527(0.5659)</td>
<td>0.0527(0.5658)</td>
<td>0.0521(0.5661)</td>
</tr>
<tr>
<td>S^VM-MC</td>
<td>0.0541(0.5662)</td>
<td>0.0527(0.5660)</td>
<td>0.0527(0.5659)</td>
<td>0.0527(0.5658)</td>
<td>0.0521(0.5661)</td>
</tr>
</tbody>
</table>

(b) The \( F_1 \) for testing set using \( V_2 \) as the feature of the data.

<table>
<thead>
<tr>
<th></th>
<th>( F_1 ) ( s = 30 )</th>
<th>( F_1 ) ( s = 25 )</th>
<th>( F_1 ) ( s = 20 )</th>
<th>( F_1 ) ( s = 15 )</th>
<th>( F_1 ) ( s = 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.0648(0.6683)</td>
<td>0.0670(0.6691)</td>
<td>0.0672(0.6702)</td>
<td>0.0673(0.6710)</td>
<td>0.0673(0.6712)</td>
</tr>
<tr>
<td>Fuzzy SVM</td>
<td>0.0652(0.6694)</td>
<td>0.0670(0.6701)</td>
<td>0.0672(0.6710)</td>
<td>0.0673(0.6712)</td>
<td>0.0673(0.6712)</td>
</tr>
<tr>
<td>Guo</td>
<td>0.0654(0.6695)</td>
<td>0.0670(0.6710)</td>
<td>0.0672(0.6710)</td>
<td>0.0673(0.6712)</td>
<td>0.0673(0.6712)</td>
</tr>
<tr>
<td>S^VM</td>
<td>0.0655(0.6695)</td>
<td>0.0670(0.6710)</td>
<td>0.0672(0.6710)</td>
<td>0.0673(0.6712)</td>
<td>0.0673(0.6712)</td>
</tr>
<tr>
<td>S^VM-MC</td>
<td>0.0674(0.6931)</td>
<td>0.0713(0.7295)</td>
<td>0.0753(0.7492)</td>
<td>0.0760(0.7784)</td>
<td>0.0793(0.7945)</td>
</tr>
</tbody>
</table>

image annotation using either feature \( V_1 \) or feature \( V_2 \) when \( s \) is selected as 30, 25, 20, 15, 10, and 5, respectively. Further, based on the fact that SpSVM-MC(2:1) obtains the similar result as that of SpSVM-MC, it is shown that our method is capable of tackling with different kinds of distribution on untagged training set.

In general, Table 2 shows that the semi-parametric regularization which extends to the untagged data helps better exploit the geometric structure of the marginal distribution of the data, and 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. Further, SpSVM-MC performs much better than the comparing methods when the noise ratio \( s \) increases, which further verifies the robustness of SpSVM-MC.

We describe the \( F_1 \) measure for the testing set as a function of \( M \) when \( u = 3, R = 0.2, \) and \( s = 20 \) using feature \( V_1 \) and feature \( V_2 \) for SpSVM-MC in Figure 4(a) and Figure 5(a), respectively. From Figure 4(a) and Figure 5(a) we observe that the \( F_1 \) measures for the testing set all increase when \( M \) increases, which shows that the semi-parametric regularization is helpful to further improve the performance of the classification by better exploiting the geometric structure of the marginal distribution of the data. The curves for the \( F_1 \) measures of SpSVM-MC in Figure 4(a) and Figure 5(a) all exhibit their major elevation from \( M = 0 \) to \( M = 4 \), then level off when \( M \) continues to increase, indicating that there is no need to choose a much larger \( M \) since the top few largest eigenvalues and the corresponding eigenvectors can sufficiently capture the geometric structure of the data.

Figure 4(b) and Figure 5(b) show the \( F_1 \) measure for the testing set as a function of \( u \) when \( M = 3, R = 0.2, \) and \( s = 20 \) using feature \( V_1 \) and feature \( V_2 \) for SpSVM-MC, respectively. We observe from Figure 4(b) and Figure 5(b) that the \( F_1 \) measures for the testing set all increase when the size of the neighborhood for each \( N_u (I_i) \) increases, which shows that it is helpful to use the nearest neighbors of each \( I_i \) in \( L_r \) to further improve the performance of the classification.

We describe the \( F_1 \) measure for the testing set as a function of \( R \) when \( M = 3, u = 3, \) and \( s = 20 \) using feature \( V_1 \) and feature \( V_2 \) for SpSVM-MC in Figure 4(c) and Figure 5(c), respectively. As we defined before, \( r = C^r / C \) represents the ratio between the weight of the loss function for instances and the weight of the loss function for their nearest neighbors. When \( R \) increases, the effect of nearest neighbors from the multi-label space in the optimization.
also increases. We observe from Figure 4(c) and Figure 5(c) that when R increases, the curves for the F1 measures of SpSVM-MC 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.

After the training process, we use the trained model to reclassify the noisily tagged training set. Figure 6 describes the F1 measure of the 75 tags for the noisily tagged training set using feature V2 for SVM and SpSVM-MC when \( s = 20, M = 4, u = 3, \) and \( R = 0.2 \). Figure 7 shows the F1 measure of the 75 tags for the testing set using feature V2 for SVM and SpSVM-MC when \( s = 20, M = 4, u = 3, \) and \( R = 0.2 \). From Figure 6 and Figure 7 we observe that the performances of SpSVM-MC are better than those of SVM in most of the concepts not only for the testing set, but also for the noisily tagged training set, indicating that our proposed method is effective and promising to correct the incorrect tags originally given in the noisily tagged training set during the training process, and consequently to improve the performance of the classification.

6. CONCLUSION

Tagging has become an important means responsible for the rapid development of social networks. A robust tagging method must have the capability to meet the two challenging requirements: limited labeled training samples and noisy labeled training samples. We have studied the challenging problem of learning with limited and noisy tagging and have developed a powerful discriminative model, called SpSVM-MC, that exploits both labeled and unlabeled data through a semi-parametric regularization and takes advantage of the multi-label constraints into the optimization. While SpSVM-MC is a general method for learning with limited and noisy tagging, we have reported the extensive evaluations in the specific application of noisy image tagging with limited labeled training samples on a benchmark dataset. Theoretic analysis and extensive evaluations in comparison with state-of-the-art literature demonstrate that SpSVM-MC outstands with a superior performance.

Acknowledgments

We thank the area chair and reviewers for their constructive comments. This work is supported in part by the National Basic Research Program of China (2012CB316400), Zhejiang University — Alibaba Financial Joint lab, and Zhejiang Provincial Engineering Center on Media Data Cloud Processing and Analysis. ZZ is also supported in part by US NSF (IIS-0812114, CCF-1017828).

7. REFERENCES

Figure 6: The F1 measure of the 75 tags for the noisy tagged training set using feature V2 for SVM and SpSVM-MC when s = 20, M = 4, u = 3, and R = 0.2.

Figure 7: The F1 measure of the 75 tags for the testing set using feature V2 for SVM and SpSVM-MC when s = 20, M = 4, u = 3, and R = 0.2.


