doi:10.3772/j.issn.1006-6748.2019.01.013

Multi-label learning algorithm with SVM based association[①]

Feng Pan (冯 攀)*, Qin Danyang[②]*, Ji Ping*, Ma Jingya*, Zhang Yan*, Yang Songxiang**
(* Key Laboratory of Electronic and Communication Engineering, Heilongjiang University, Harbin 150080, P. R. China)
(** School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, P. R. China)

Abstract

Multi-label learning is an active research area which plays an important role in machine learning. Traditional learning algorithms, however, have to depend on samples with complete labels. The existing learning algorithms with missing labels do not consider the relevance of labels, resulting in label estimation errors of new samples. A new multi-label learning algorithm with support vector machine (SVM) based association (SVMA) is proposed to estimate missing labels by constructing the association between different labels. SVMA will establish a mapping function to minimize the number of samples in the margin while ensuring the margin large enough as well as minimizing the misclassification probability. To evaluate the performance of SVMA in the condition of missing labels, four typical data sets are adopted with the integrity of the labels being handled manually. Simulation results show the superiority of SVMA in dealing with the samples with missing labels compared with other models in image classification.

Key words: multi-label learning, missing labels, association, support vector machine (SVM)

0 Introduction

Labels are important characteristics of images and are necessary carriers in image processing. Abundant unlabeled images in existence will cause low efficiency in information extraction. Multi-label learning is an emerging method to annotate images with missing labels. Traditional learning algorithms, however, have to depend on images training with complete labels[1,2], which can be hardly achieved in practical applications. Some new learning algorithms are presented aiming to solve the missing labels problem. Ref. [3] introduced a concept of fuzzy mutual information. Ref. [4] proposed a new multi-label learning formulation by introducing a self-paced function as the regularizer. Ref. [5] established a model of ranking-preserving low-rank factorization with missing labels. Ref. [6] proposed a multi-label classification method that can learn the inductive classifier and explicitly deal with the missing labels. The training method in Ref. [7] had only positive label data and unlabeled data, with which only standard binary classifiers can be learned. Ref. [8] proposed a semi-supervised multi-class learning method. Ref. [9] proposed a weak label learning method, and the multi-label sorting with a group lasso was proposed in Ref. [10], considering multi-label classification as bidirectional sorting. Ignoring the missing labels in research models will cause errors or mistakes in image processing. A missing-label multi-label (MLML) learning method proposed in Ref. [11] set the positive labels and negative labels to different values to solve the binomial problem in some degree, but failed to deal with the complicated distributed data. Ref. [12] showed that multi-label learning based on support vector machine (SVM) was an effective method. Ref. [13] proposed an active learning based on SVM. Ref. [14] minimized the margin and rank loss[15]. However, not full consideration about the relevance of the labels will result in label estimation errors, which may reduce the usability of the image information. Therefore, a new multi-label learning algorithm with SVM based association (SVMA) is proposed in this paper to deal with the possible overfitting problem of SVM. Loss functions and association models among labels will be established by sample smoothness and class smoothness to estimate the missing labels, so as to improve the accuracy of data classification.

[①] Support by the National High Technology Research and Development Program of China (No. 2012AA120802), National Natural Science Foundation of China (No. 61771186), Postdoctoral Research Project of Heilongjiang Province (No. LBH-Q15121), Undergraduate University Project of Young Scientist Creative Talent of Heilongjiang Province (No. UNPYSCT-2017125).

[②] To whom correspondence should be addressed. E-mail: qindanyang@hlju.edu.cn

Received on July 13, 2018

1 SVM classifier

1.1 Support vector and margin

Given a data set $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}$, and $y_i \in \{-1, +1\}$.

Recall that classification learning is used to find a hyperplane to classify data into two categories accurately. In the sample space, the hyperplane is defined by

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0 \tag{1}$$

where, $\boldsymbol{w}$ and b are normal vector and bias of the hyperplane respectively.

The closest samples from the hyperplane are the support vectors which satisfy $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = \pm 1$ and the distance between negative and positive support vectors is margin, e. g. γ in Fig. 1. After obtaining the hyperplane, the data class can be obtained by Eq. (1).

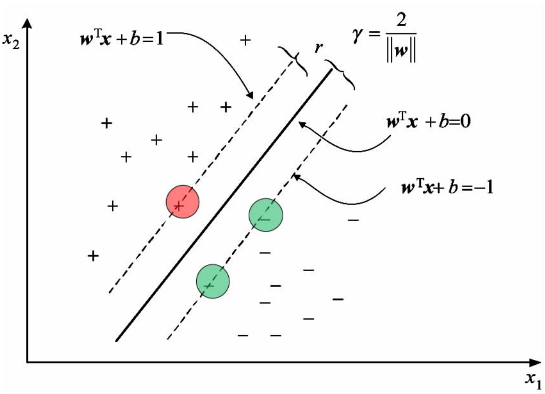

x_2
$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=1$
r
$\gamma = \frac{2}{\|\boldsymbol{w}\|}$
$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=0$
$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=-1$
x_1

Fig. 1 Support vector and margin

1.2 Multi-label SVM

Given data set $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n] \in \boldsymbol{R}^{d\times n}$, and these images can be divided into m classes $\{c_1, \cdots, c_m\}$. The image is annotated by the label matrix $\boldsymbol{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)$, where $\boldsymbol{y}_i \in \{-1, +1\}^{m\times 1}$ is the label of the ith sample, and $y_{ki} = 1 (k = 1, \cdots, m)$ indicates that sample $\boldsymbol{x}_i$ belongs to the kth class. Vector $\boldsymbol{w}_k$ and bias b_k satisfy:

$$g = \min\left(\frac{1}{2}\|\boldsymbol{w}_k\|^2 + c\sum_{i=1}^{n}\xi_{ki}\right) \tag{2}$$

Eq. (2) meets $y_{ki}(\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i + b_k) \geqslant 1 - \xi_{ki}, \forall i, \xi_{ki} \geqslant 0$, where ξ_{ki} is the slack variable, and $c > 0$ is a constant.

For SVM is a binary classifier, one-vs-rest strategy is adopted, where samples in the kth class are considered as class A and the other samples are considered as class B , and the kth binary classifier is represented by $f_k(\boldsymbol{x}_i) = \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i + b_k$. $f(\boldsymbol{x}) > 0$ means true if $\boldsymbol{x}$ belongs to the kth class, otherwise it is false.

Although realizing multi-class classification, SVM does not take the associativity between labels into account, so as to fail to solve the multi-label learning problems with missing labels.

Traditional SVM is optimized in this paper with the working process shown in Fig. 2.

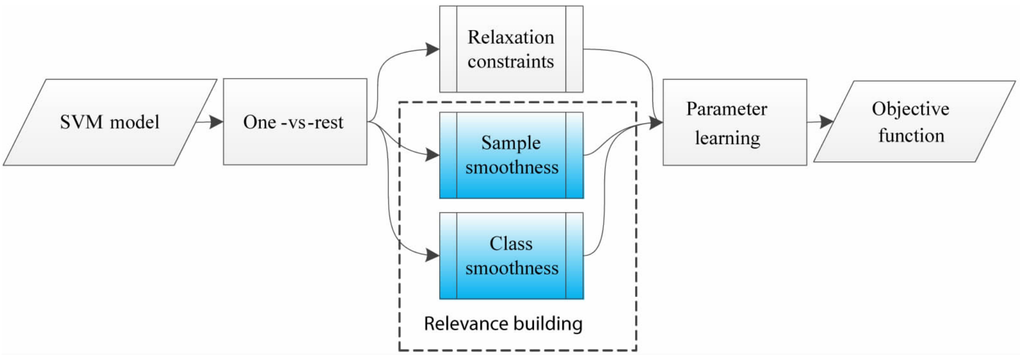


Fig. 2 Implementation process of SVMA

2 Loss function of SVMA

Given a sample $\boldsymbol{x}_i$, and the label $\boldsymbol{y}_i \in \{-1, 0, 1\}^{m\times 1}$, where $\boldsymbol{y}_{ki} = 0$ indicates that the kth label is missing.

As shown in Fig. 2, sample smoothness and class smoothness are used to estimate the missing label. However, the model will become too complicated to solve if SVM is directly combined with them, and it will reduce the flexibility of SVM. Therefore, Eq. (2) can be replaced by

$$g = \min\sum_{i}^{m}\frac{1}{2}\boldsymbol{err}_i^{\mathrm{T}}\boldsymbol{err}_i + \frac{1}{2}tr(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}) \tag{3}$$

where, $\boldsymbol{err}_i$ is a column vector with element $f_{iq} - y_{iq}$, and subscript $q = \{k \mid 1 \leqslant k \leqslant n, f_{ik} \times y_{ik} < (1 - \xi_{ik})\}$ is a sample index. f_{ik} represents $f_{w_i, b_i}(\boldsymbol{x}_k) = \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x}_k + b_i$, where $\boldsymbol{w}_i \in \boldsymbol{R}^d$ is a weighted vector of category c_i, and b_i is the bias. $\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_m] \in \boldsymbol{R}^{d\times m}$ is the weight matrix. Here, the margin will be $2/tr(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})$.

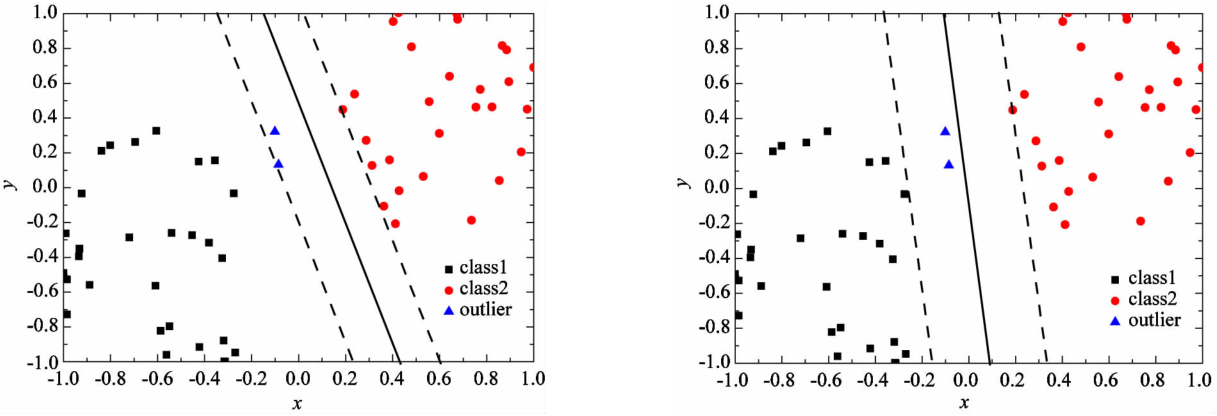
Standard SVM aims to find the largest margin with the whole of samples out of margin, and the model in the paper will minimize the number of samples in the margin and maximize the margin at the same time. The first term in Eq. (3) tries to achieve the first purpose, and the second term can maximize margin through $\boldsymbol{W}$.

There are two types of typical data, which include 28 samples satisfying Gaussian distribution respectively. These data are utilized to compare the performance of SVM and SVMA with consistent slack variables. The relevant distribution parameters are shown in Table 1.

Table 1 Relevant distribution parameters of class 1 and class 2

	Covariance matrix	Mean vector
Class1	[0.07 −0.02; −0.02 0.17]	[−0.63 −0.38]
Class2	[0.06 0.02; 0.02 0.13]	[0.61 0.43]

Further, two outliers, which are [−0.1, 0.32] and [−0.09, 0.13], are provided to affect learning results. Fig. 3 shows the comparing results that both the two models are able to realize classification effectively. However, SVM allows choosing an outlier as the support vector, which is the overfitting problem that increases the probability of a new sample being misclassified, making the margin small and the hyperplane near to the samples in class 2 , while SVMA considers the general features of most samples, and selects a point of class 1 as the support vector, which can obtain such a more reasonable and large margin that improve the robustness of the model. Based on this, the association of labels will be built subsequently to estimate missing labels.



(a) SVM (b) SVMA

Fig. 3 Comparison between SVM and SVMA

3 Construction of association in SVMA

3.1 Sample smoothness

An adjacency graph $G = (\boldsymbol{X}, \boldsymbol{V})$ can be used to characterize the local geometry of the training samples $\boldsymbol{X}$ as shown in Fig. 4.

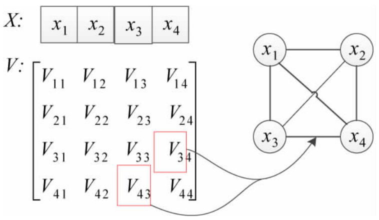

X:
x_1 x_2 x_3 x_4
V:
V_{11} V_{12} V_{13} V_{14}
V_{21} V_{22} V_{23} V_{24}
V_{31} V_{32} V_{33} V_{34}
V_{41} V_{42} V_{43} V_{44}
x_1
x_2
x_3
x_4

Fig. 4 Adjacency graph $G = (\boldsymbol{X}, \boldsymbol{V})$

$\boldsymbol{X}$ in Fig. 4 is the vertex of graph G and $\boldsymbol{V}$ is a weight matrix along with V_{ij} (or V_{ji}) representing the relationship between samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Generally, $\boldsymbol{V}$ is a symmetric matrix with element defined by

$$V_{ij} = \exp \frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\sigma} \tag{4}$$

where, σ is the hyper parameter. It is obvious that V_{ij} represents the closeness of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

Since Eq. (2) does not consider the local geometric structure of the data during learning, Eq. (5) is added as a new constraint.

$$s = \min \frac{1}{2} \sum_{i,j}^{n} V_{ij} \left\| \frac{z_i}{\sqrt{d_i}} - \frac{z_j}{\sqrt{d_j}} \right\|^2 \tag{5}$$

where, $d_i = \sum_{j=1}^{n} V_{ij}$, and $\boldsymbol{z}_i$ is the label of $\boldsymbol{x}_i$ obtained based on the method proposed in the paper. z_{ij}, which is the jth element of $\boldsymbol{z}_i$, is defined as

$$z_{ij} = \mathrm{sign}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x}_j + b_i) \tag{6}$$

here, sign(x) is a sign function with the value being 1 when there is $x > 0$, otherwise is -1. Samples smoothness is considered during learning progress of parameters $\boldsymbol{w}$ and b, as a result, if samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are close, the difference between estimated labels z_i and z_j should be small enough, or it will affect the solution of Eq. (5), which reflects the correlation between the label and the sample.

3.2 Class smoothness

Similar to Section 3.1, $G_c = (\boldsymbol{Y}^{\mathrm{T}}, \boldsymbol{Q})$ is the class graph built on class adjacency matrix $\boldsymbol{Y}^{\mathrm{T}}$ and $\boldsymbol{Q}$, which is the weight matrix with element Q_{ij} defined as

$$Q_{ij} = \exp(-\varphi(1 - \cos(\boldsymbol{u}_i, \boldsymbol{u}_j))) \tag{7}$$

where φ is the hyper parameter and $\boldsymbol{u}_i (i = 1, 2, \cdots, m)$ is the ith column of $\boldsymbol{Y}^{\mathrm{T}}$. The cosine similarity of $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$ is calculated by $\cos(\boldsymbol{u}_i, \boldsymbol{u}_j) = \langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle / \| \boldsymbol{u}_i \| \| \boldsymbol{u}_j \|$.

The smoothness of class-level label matrix $\boldsymbol{Y}$ is represented as

$$c = \min \frac{1}{2} \sum_{i,j}^{m} Q_{ij} \left\| \frac{\tilde{z}_i}{\sqrt{s_i}} - \frac{\tilde{z}_j}{\sqrt{s_j}} \right\|^2 \tag{8}$$

where, $\boldsymbol{Z} = [z_1, \cdots, z_n] \in \boldsymbol{R}^{m \times n}$, $\tilde{z}_i (i = 1, 2, \cdots, m)$ is the ith column of $\boldsymbol{Z}^{\mathrm{T}}$, $s_i = \sum_{j=1}^{m} Q_{ij}$.

It can be seen that Eq. (8) considers the label smoothness when learning $\boldsymbol{w}$ and b that if $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$ are close to each other and estimated vectors $\tilde{z}_i$ and $\tilde{z}_j$ have obvious difference, the value of Eq. (8) will be affected, therefore, two close classes of samples have to generate two similar estimated labels.

4 SVMA weight vector learning

Eqs (9) and (10) can be obtained by simple computations.

$$\frac{1}{2} \sum_{i,j}^{n} V_{ij} \left\| \frac{z_i}{\sqrt{d_i}} - \frac{z_j}{\sqrt{d_j}} \right\|^2 = tr(\boldsymbol{Z}(\boldsymbol{I} - \boldsymbol{D}^{-1/2} \boldsymbol{V} \boldsymbol{D}^{-1/2}) \boldsymbol{Z}^{\mathrm{T}}) = \mathrm{tr}(\boldsymbol{Z} \boldsymbol{L} \boldsymbol{Z}^{\mathrm{T}}) \tag{9}$$

$$\frac{1}{2} \sum_{i,j}^{n} Q_{ij} \left\| \frac{\tilde{z}_i}{\sqrt{d_i}} - \frac{\tilde{z}_j}{\sqrt{d_j}} \right\|^2 = tr(\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{I} - \tilde{\boldsymbol{D}}^{-1/2} \boldsymbol{Q} \tilde{\boldsymbol{D}}^{-1/2}) \boldsymbol{Z}) tr(\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{H} \boldsymbol{Z}) \tag{10}$$

Matrixes $\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-1/2} \boldsymbol{V} \boldsymbol{D}^{-1/2}$ and $\boldsymbol{H} = \boldsymbol{I} - \tilde{\boldsymbol{D}}^{-1/2} \boldsymbol{Q} \tilde{\boldsymbol{D}}^{-1/2}$ respectively are the normalized Laplacian matrix of graph G and G_c which are symmetric matrixes, where $\boldsymbol{D} = diag(d_1, \cdots, d_n)$ and $\tilde{\boldsymbol{D}} = diag(s_1, \cdots, s_m)$ are the diagonal matrixes.

SVMA will find a mapping function to satisfy:

$$g = \min \sum_{i}^{m} \frac{1}{2} \boldsymbol{err}_i^{\mathrm{T}} \boldsymbol{err}_i + \frac{1}{2} tr(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{W}) + \beta tr(\boldsymbol{Z} \boldsymbol{L} \boldsymbol{Z}^{\mathrm{T}}) + \gamma \mathrm{tr}(\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{H} \boldsymbol{Z}) \tag{11}$$

where, β and γ are non-negative constants, representing the weight of the sample smoothness and the label smoothness, and can be modified by cross validation method, and $\boldsymbol{Z}$ is related to the sign function. The non-differentiable sign function, however, makes Eq. (11) fail to be solved directly. Thus, an approximating process as Eq. (12) can be adopted to represent $\boldsymbol{Z}$.

$$z_{ij} = \mathrm{sgn}(\boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{x}_j + b_i) \approx 2\sigma(\tau(\boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{x}_j + b_i)) - 1 \tag{12}$$

where $\sigma(a) = 1/(1 + \exp(-a))$ is a sigmoid function and $\tau \geqslant 1$ is a parameter, and $z_{ij} \in [-1, 1]$.

$\boldsymbol{w}_i$ and $\boldsymbol{x}_j$ are defined as Eq. (13) to solve Eq. (11).

$$\boldsymbol{w}_i = \begin{bmatrix} b_i \\ \boldsymbol{w}_i \end{bmatrix}, \boldsymbol{x}_j = \begin{bmatrix} 1 \\ \boldsymbol{x}_j \end{bmatrix} \tag{13}$$

The mapping function is $f_{w_i, b_i}(\boldsymbol{x}_j) = \boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{x}_j$, and $\boldsymbol{w}_i$ can be learnt iteratively by

$$g = \min_{\boldsymbol{w}_i} \left\{ \frac{1}{2} err_i^T err_i + \frac{\| \boldsymbol{w}_i \|^2}{2} + \frac{\beta}{2} \sum_{j,r}^{n} V_{jr} \left(\frac{z_{ij}}{\sqrt{d_j}} - \frac{z_{ir}}{\sqrt{d_r}} \right)^2 + \gamma \frac{1}{2} \sum_{r \neq i}^{m} \sum_{j}^{n} Q_{ir} \left(\frac{z_{ij}}{\sqrt{s_i}} - \frac{z_{rj}}{\sqrt{s_r}} \right)^2 \right\} \tag{14}$$

To simplify the derivation process, four terms in Eq. (14) can be written as A, B, C, and D orderly. The partial derivatives of A and B to $\boldsymbol{w}_i$ are shown in Eq. (15), where $\hat{\boldsymbol{X}} = \{ \boldsymbol{x}_r \mid f_{ir} \times y_{ir} < (1 - \xi_{ir}) \}$ and the partial derivatives of C and D to $\boldsymbol{w}_i$ are shown in Eqs (16) and (17) respectively, where σ_{ij} is the abbreviation of function $\sigma(\tau \boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{x}_j)$.

$$\frac{\partial A}{\partial \boldsymbol{w}_i} = \hat{\boldsymbol{X}} \cdot err_i, \frac{\partial B}{\partial \boldsymbol{w}_i} = \boldsymbol{w}_i \tag{15}$$

According to $\nabla \boldsymbol{w}_i = \frac{\partial A}{\partial \boldsymbol{w}_i} + \frac{\partial B}{\partial \boldsymbol{w}_i} + \frac{\partial C}{\partial \boldsymbol{w}_i} + \frac{\partial D}{\partial \boldsymbol{w}_i}$, Eq. (18) can be obtained:

$$\beta \sum_{j,r}^{n} V_{jr} \left(\frac{2\sigma_{ij} - 1}{\sqrt{d_j}} - \frac{2\sigma_{rj} - 1}{\sqrt{d_r}} \right) \left(\frac{2 \sum_{j}^{n} \boldsymbol{x}_j \sigma_{ij}(1 - \sigma_{ij})}{\sqrt{d_j}} - \frac{2 \sum_{r}^{n} \boldsymbol{x}_r \sigma_{ir}(1 - \sigma_{ir})}{\sqrt{d_r}} \right) = \frac{4\beta \sum_{j}^{n} \boldsymbol{x}_j \sigma_{ij}(1 - \sigma_{ij})}{\sqrt{d_j}} \left(\frac{2\sigma_{ij} - 1}{\sqrt{d_j}} \sum_{r}^{n} V_{jr} - \sum_{r}^{n} \frac{V_{jr}(2\sigma_{rj} - 1)}{\sqrt{d_j}} \right) \tag{16}$$

$$\frac{\partial D}{\partial \boldsymbol{w}_i} = \gamma \sum_{r \neq i}^{m} \sum_{j}^{n} Q_{ir}\left(\frac{2\sigma_{ij} - 1}{\sqrt{s_i}} - \frac{2\sigma_{rj} - 1}{\sqrt{s_r}}\right)\frac{2\sum_{j}^{n} \boldsymbol{x}_j \sigma_{ij}(1 - \sigma_{ij})}{\sqrt{s_r}}$$

$$= \frac{2\gamma \sum_{j=1}^{n} x_j \sigma_{ij}(1 - \sigma_{ij})}{\sqrt{s_i}}\left(\frac{2\sigma_{ij} - 1}{\sqrt{s_i}} \sum_{r \neq i}^{m} Q_{ir} - \sum_{r \neq i}^{m} \frac{Q_{ir}(2\sigma_{rj} - 1)}{\sqrt{s_r}}\right) \quad (17)$$

$$\nabla \boldsymbol{w}_i = \hat{\boldsymbol{X}}err_i + \boldsymbol{w}_i + 2\sum_{j=1}^{n} x_j \sigma_{ij}(1 - \sigma_{ij})\left(\frac{2\beta}{\sqrt{d_j}} \times \left[\frac{2\sigma_{ij} - 1}{\sqrt{d_j}} \sum_{r}^{n} V_{jr} - \sum_{r}^{n} \frac{V_{jr}(2\sigma_{rj} - 1)}{\sqrt{d_r}}\right]\right.$$

$$\left. + \frac{\gamma}{\sqrt{s_i}}\left[\frac{2\sigma_{ij} - 1}{\sqrt{s_i}} \sum_{r \neq i}^{m} Q_{ir} - \sum_{r \neq i}^{m} \frac{Q_{ir}(2\sigma_{rj} - 1)}{\sqrt{d_r}}\right] + \frac{\gamma}{\sqrt{s_i}}\left[\frac{2\sigma_{ij} - 1}{\sqrt{s_i}} \sum_{r \neq i}^{m} Q_{ir} - \sum_{r \neq i}^{m} \frac{Q_{ir}(2\sigma_{rj} - 1)}{\sqrt{s_r}}\right]\right) \quad (18)$$

Next, in the tth iteration, $\boldsymbol{w}_i$ will update with the gradient falling as shown in

$$\boldsymbol{w}_i^{(t+1)} \leftarrow \boldsymbol{w}_i^t - \alpha_t \nabla \boldsymbol{w}_i^t \quad (19)$$

where α_t can be obtained by Armijo criterion.

5 Performance simulation and evaluation

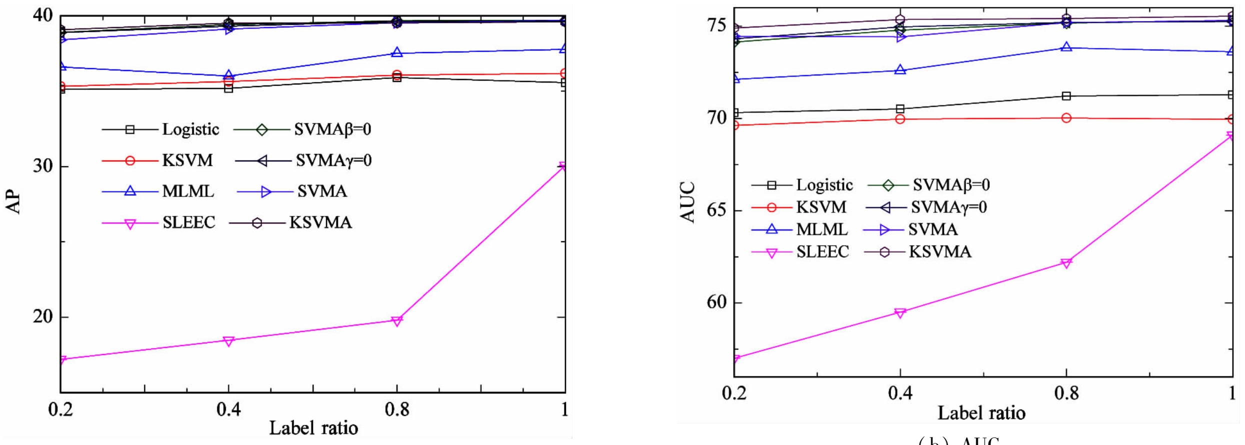
Four typical image sets[16-18] (ESP-GAME, MIR Flickr, NUS-WIDE-Lite, Wiki10) are adopted in the simulation to verify the learning effect of SVMA. Two algorithms proposed recently (MLML and SLEEC) and two widely used algorithms (SVM with RBF kernel and Logistic regression with L1 norm) are added in the simulation to compare and evaluate image process ability with missing labels further. β and γ will be adjusted by cross validation with the range lying in $[10^{-2}, 10^2]$. To handle the data in more complex distribution, SVMA with RBF-kernel (KSVMA) is also adopted. Further, the cases with $\beta = 0$ and $\gamma = 0$ ($\text{SVMA}_{\beta=0}$ and $\text{SVMA}_{\gamma=0}$) are supplemented to evaluate the effects of both sample smoothness and class smooth, and the parameter $\tau = 0$.

Training data are constructed in the simulation in order to verify the effect of all methods with missing labels. For the sake of generality, the label retention rate ranges from 20% to 100% in all data sets with the simulation results shown as Figs 5 – 8.

Two common evaluation parameters, the average precision (AP) shown in Eq. (20) and the area under the ROC curve (AUC) are used in the experiment as the indicators to evaluate the performance of multi-label classification.

$$\text{AP} = \frac{1}{n}\sum_{i}^{n} \frac{1}{|S^i|} \sum_{S_r \in S^i} \frac{|\{s_t \in S^i \mid rank(\boldsymbol{x}_i, s_t) < rank(\boldsymbol{x}_i, s_r)\}|}{rank(\boldsymbol{x}_i, s_r)} \quad (20)$$

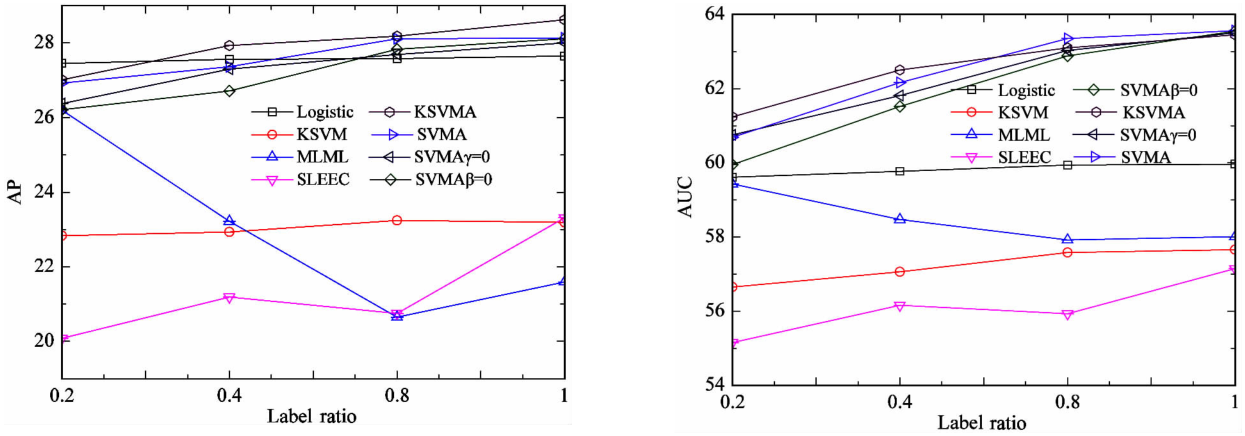
where $rank(\boldsymbol{x}_i, s_t)$ is the order of class s_t lying in the sorted list of sample $\boldsymbol{x}_i$, and S^i is the real positive sample label of $\boldsymbol{x}_i$. The larger the value of AP is, the more accurately the algorithm will perform, and AUC is the region below ROC curve[19], the value of which in the simulation is the average of multiple ROC curves. The larger the value of AUC is, the better the classification capability will be.



(a) AP

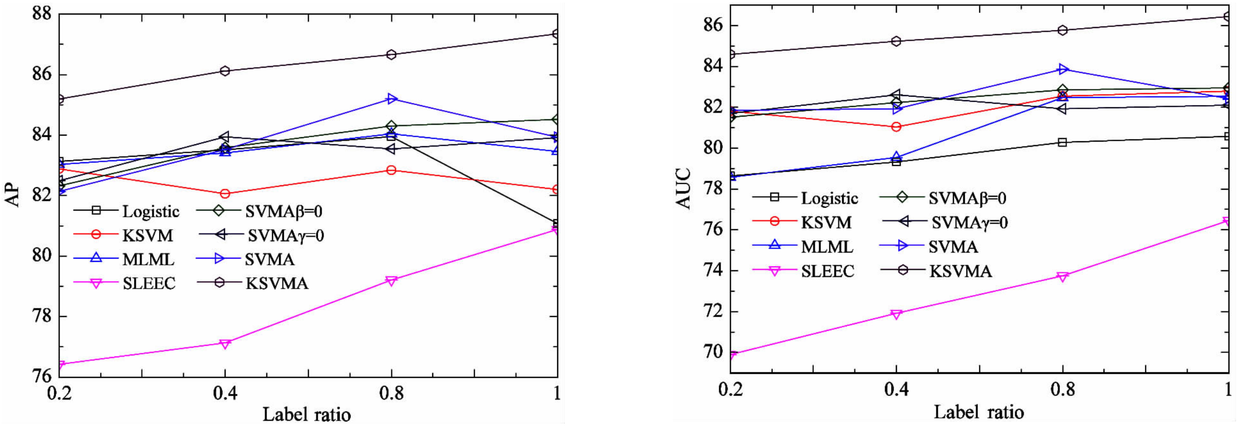
(b) AUC

Fig. 5 Simulation results in ESP game



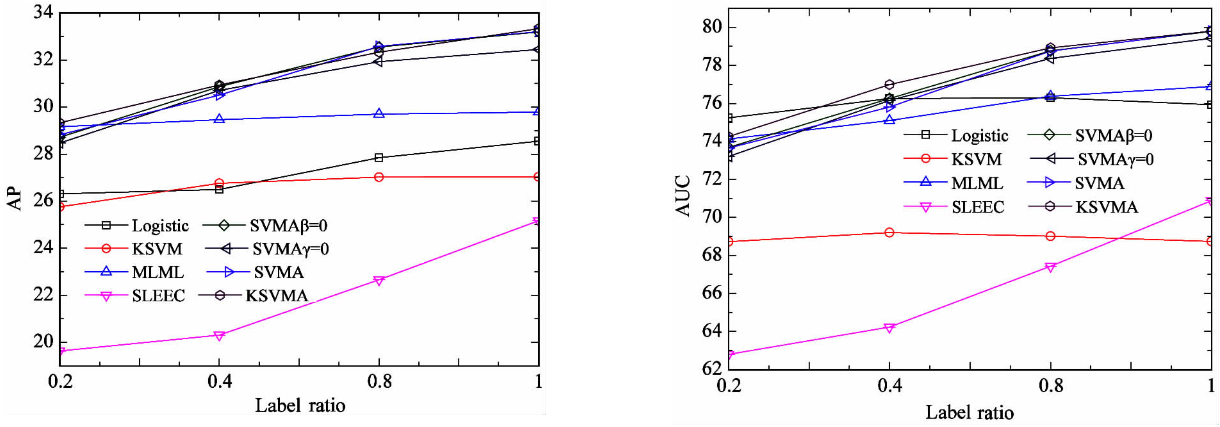
(a) AP (b) AUC

Fig. 6 Simulation results in MIR Flickr



(a) AP (b) AUC

Fig. 7 Simulation results in NUS-WIDE-Lite



(a) AP (b) AUC

Fig. 8 Simulation results in Wiki10

Different algorithms in different data sets show different classification ability, the general comparison results are nearly the same. Compared with the other algorithms, simulation results show that SVMA proposed in this paper can bring better classification performance in the condition of missing labels because of the multi-label learning with association.

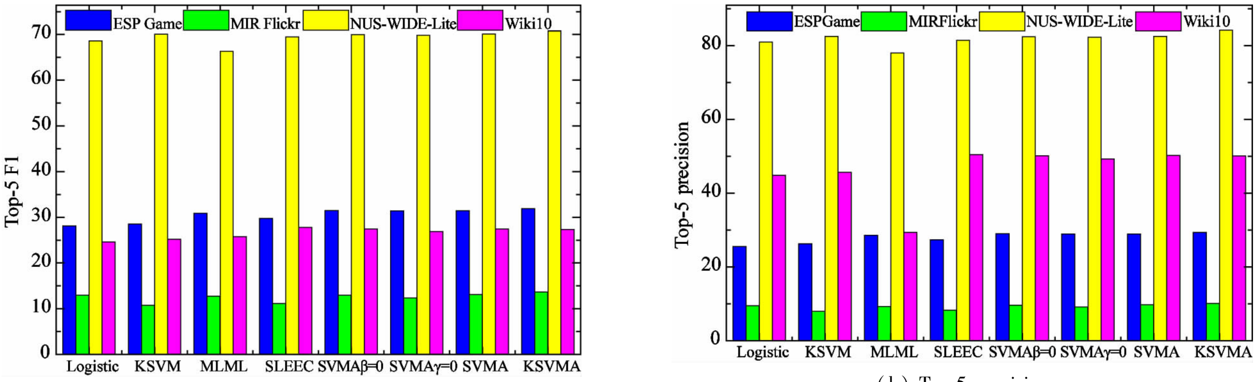
Next, the classification performance of each classifier with complete labels will be analyzed by three typical methods Top-5 F1, Top-5 precision and *P*@ *k*.

Specifically, the first five categories in each ordered list of the test images are considered positive, while the others are negative. A discrete label matrix will be obtained by Top-5 accuracy and Top-5 F1.

P@ *k* is the abbreviation of precision at *k* which focuses on the prediction results of the first *k* positive values. They are often used as the evaluation criteria during label ranking. Given the real label vector and the predicted value, accuracy *k* can be calculated as Eq. (21).

$$P@k(\hat{y},\ y) = \frac{1}{k}\sum_{i\in rank_k(\hat{y})} y_i \qquad (21)$$

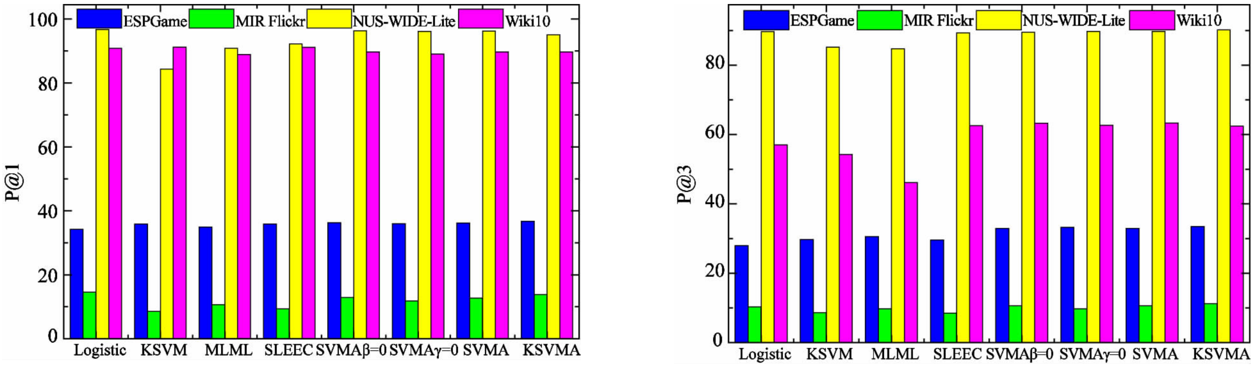
According to Eq. (21), *P*@ *k* with $k = 5$ is equivalent to Top-5 precision. Performance comparisons of the algorithms evaluated by the methods above are shown in Fig. 9 and Fig. 10.



(a) Top-5 F1

(b) Top-5 precision

Fig. 9 Top-5 evaluation (%) comparison of different algorithms



(a) P@1

(b) P@3

Fig. 10 P@1 and P@3 evaluation (%) comparisons of different algorithms

Fig. 11 shows the convergence curves for the algorithm on four typical databases. It can be seen that the algorithm proposed in this paper has a better convergence performance than the other three, which indicates the proposed algorithm SVMA has higher computational efficiency than the common algorithms.

In order to evaluate the effect of τ in Eq. (12), the percentage of the training samples coming from the NUS-WIDE-Lite data set with missing labels will range from 20% to 100%. The simulations are performed for 10 times to obtain the mean values of AUC and AP for different missing label rates as shown in Fig. 12. It can be seen from Fig. 12 that SVMA has better performance when τ is 1.

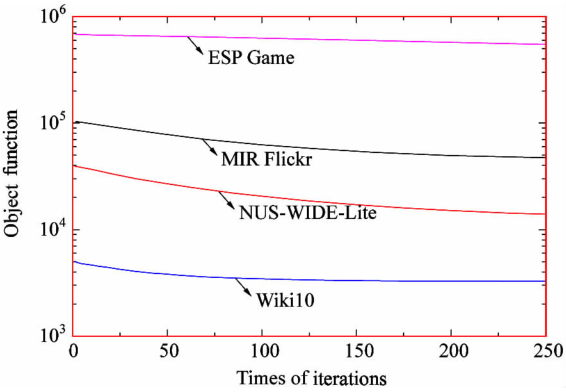
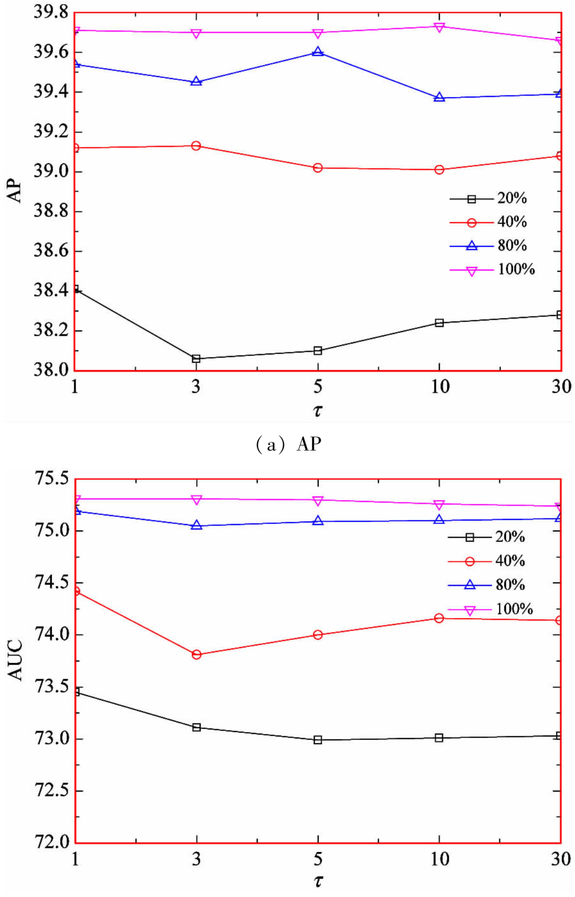


Fig. 11 The convergence curves with different times of iterations



(a) AP

(b) AUC

Fig. 12 Simulation results of SVMA with different values of τ

6 Conclusion

To solve the problems of multi-label learning under condition of missing labels, a new algorithm SVMA is proposed in the paper by establishing an effective mapping function, which can not only provide a margin large enough, but can minimize the number of samples in the margin so as to increase the robustness to noise. Class smoothness and sample smoothness are adopted and the association among labels is constructed to estimate the missing labels. The simulation results show that the proposed algorithm SVMA will achieve better average classification accuracy than the other typical algorithms in the absence or completeness of the labels. Moreover, the rapid convergence of SVMA makes it possess significant practical application value. There are still some issues to be further considered. The complexity cost of SVMA will be evaluated and the theoretic model of convergence will be established in the future research.

References

[1] Li X, Zhao X, Zhang Z, et al. Joint multilabel classification with community-aware label graph learning [J]. *IEEE Transactions on Image Processing*, 2015, 25(1): 484-493

[2] Zhang M L, Wu L. Lift: multi-label learning with label-specific features[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2015, 37(1):107-120

[3] Lin Y, Hu Q, Liu J, et al. Streaming feature selection for multilabel learning based on fuzzy mutual information [J]. *IEEE Transactions on Fuzzy Systems*, 2017, 25 (6):1491-1507

[4] Li C, Wei F, Yan J, et al. A self-paced regularization framework for multilabel learning. [J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2016, 29 (6):2660-2666

[5] Li X, Shen B, Liu B D, et al. Ranking-preserving low-rank factorization for image annotation with missing labels [J]. *IEEE Transactions on Multimedia*, 2018, 20(5): 1169-1178

[6] Ma J, Fan J, Wang W. Multi-label classification for images with missing labels[C]. In: Proceedings of the 15th International Conference on Industrial Informatics, Emden, Germany, 2017. 1050-1055

[7] Fung G P C, Yu J X, Lu H, et al. Text classification without negative examples revisit[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2006, 18(1):6-20

[8] Yi L, Rong J, Liu Y. Semi-supervised multi-label learning by constrained non-negative matrix factorization. [C]. In: Proceedings of National Conference on Artificial Intelligence and the 8th Innovative Applications of Artificial Intelligence Conference, Boston, USA, 2006. 421-426

[9] Sun Y Y, Zhang Y, Zhou Z H. Multi-label learning with weak label[C]. In: Proceedings of the 24th AAAI Conference on Artificial Intelligence, Atlanta, USA, 2010. 593-598

[10] Bucak S S, Jin R, Jain A K. Multi-label learning with incomplete class assignments[C]. In: Proceedings of Computer Vision and Pattern Recognition, Colorado Springs, USA, 2011. 2801-2808

[11] Wu B, Lyu S, Hu B G, et al. Multi-label learning with missing labels for image annotation and facial action unit recognition [J]. *Pattern Recognition*, 2015, 48 (7): 2279-2289

[12] Vapnik V N. The Nature of Statistical Learning Theory [M]. Springer, 2000

[13] Li X, Wang L, Sung E. Multilabel SVM active learning for image classification[C]. In: Proceedings of International Conference on Image Processing. Singapore, 2004. 2207-2210

[14] Elisseeff A, Weston J. A kernel method for multi-labelled classification[C]. In: Proceedings of International Conference on Neural Information Processing Systems: Natural and Synthetic, Vancouver, Canada, 2001. 681-687

[15] Singer R E S Y. BoosTexter: A Boosting-based System for text categorization[J]. *Machine Learning*, 2000, 39(2-3):135-168

[16] Von Ahn A L. Method for labeling images through a computer game [P]. US patent: 7980953. 2011

[17] Huiskes M J, Lew M S. The MIR flickr retrieval evaluation[C] In: Proceedings of ACM International Conference on Multimedia Information Retrieval. Vancouver, Canada, 2008. 39-43

[18] Chen X, Mu Y, Yan S, et al. Efficient large-scale image annotation by probabilistic collaborative multi-label propagation[C]. In: Proceedings of ACM International Conference on Multimedia, Firenze, Italy, 2010. 35-44

[19] Fawcett T. An introduction to ROC analysis[J]. *Pattern Recognition Letter*, 2006, 27(8):861-874

Feng Pan, born in 1993. He received his B. Sc. degree from Hangzhou Dianzi University in 2017. He is studying at Heilongjiang University for postgraduate diploma and majors in the information and communication engineering. His research interests include indoor vision positioning and crowd source sensing.