Contents lists available at ScienceDirect
Information Sciences
journal homepage: www.elsevier.com/locate/ins

Top-k multi-class SVM using multiple features

Caixia Yan[a], Minnan Luo[a,*], Huan Liu[a], Zhihui Li[b], Qinghua Zheng[a]

[a] SPKLSTN Lab, Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China
[b] Beijing Etrol Technologies Co., Ltd., Beijing, China

ARTICLE INFO

Article history:
Received 14 December 2016
Revised 24 July 2017
Accepted 2 August 2017
Available online xxx

Keywords:
Multiple feature learning
Top-k SVM
Multi-class visual classification

ABSTRACT

Recent studies have demonstrated the advantages of fusing multi-modal features in improving the accuracy of visual object classification. However, regarding a complex classification task with a large number of categories, previous studies on multiple feature fusion are prone to failure resulting from the occurrence of class ambiguity. In this paper, we address this issue by allowing $k (k \geq 2)$ guesses at the top instead of only considering the one with the largest prediction score in the framework of multi-view learning. This strategy relaxes the penalty for making an error in the top-k predictions, which can mitigate the challenge of class ambiguity to some extent. To fuse multiple features effectively, we introduce an adaptive weight for each view and exploit an efficient alternating optimization algorithm to learn the optimal classifiers and their corresponding weights jointly. Extensive experiments on several benchmark datasets illustrate the effectiveness and superiority of the proposed model over the state-of-the-art approaches.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

Multi-modal fusion aims to combine information related to the same object acquired from different types of detectors, multiple sources or various conditions [3,4,18]. This task has attracted more and more research interests due to the widespread use in real-world applications, such as visual object classification [29,30], tactile object recognition [26], robotics [27,42] and medical informatics [46,47]. In particular, Liu et al. [29] originally developed a visual-tactile fusion framework for object recognition to enhance the performance significantly which indeed provided an effective strategy for multi-modal fusion. Inspired by this method, we focus on the application of multi-modal fusion to visual object classification since image/video object contains heterogeneous features acquired from different visual descriptors essentially. Each descriptor can be regarded as an independent acquisition framework, termed as a modality [18]. It is highly expected to take at least two features of image/video into account, each of which contains not only relevant information to the other features but also specific details that are different and irrelevant. This corresponds to the two significant principles ensuring the success of multiple feature learning: *consensus* and *complementary* [44]. The goal of *consensus* principle is to maximize the agreement of multiple views in classification. Under *complementary* principle, classifiers with respect to each view will exchange complementary information with each other and thus learn from each other during the training process.

Visual object classification still remains a challenge, which mainly lies in the following two folds. On one hand, the classification accuracy is likely to undergo a great decline as a result of class ambiguity, which is extremely apparent in

[*] Corresponding author.
E-mail addresses: yancaixia@stu.xjtu.edu.cn (C. Yan), minnluo@mail.xjtu.edu.cn, minnluo@xjtu.edu.cn (M. Luo), hliuxjtu@gmail.com (H. Liu), zhihuilics@gmail.com (Z. Li), qhzheng@mail.xjtu.edu.cn (Q. Zheng).

http://dx.doi.org/10.1016/j.ins.2017.08.004
0020-0255/© 2017 Elsevier Inc. All rights reserved.

Please cite this article as: C. Yan et al., Top-k multi-class SVM using multiple features, Information Sciences (2017), http://dx.doi.org/10.1016/j.ins.2017.08.004

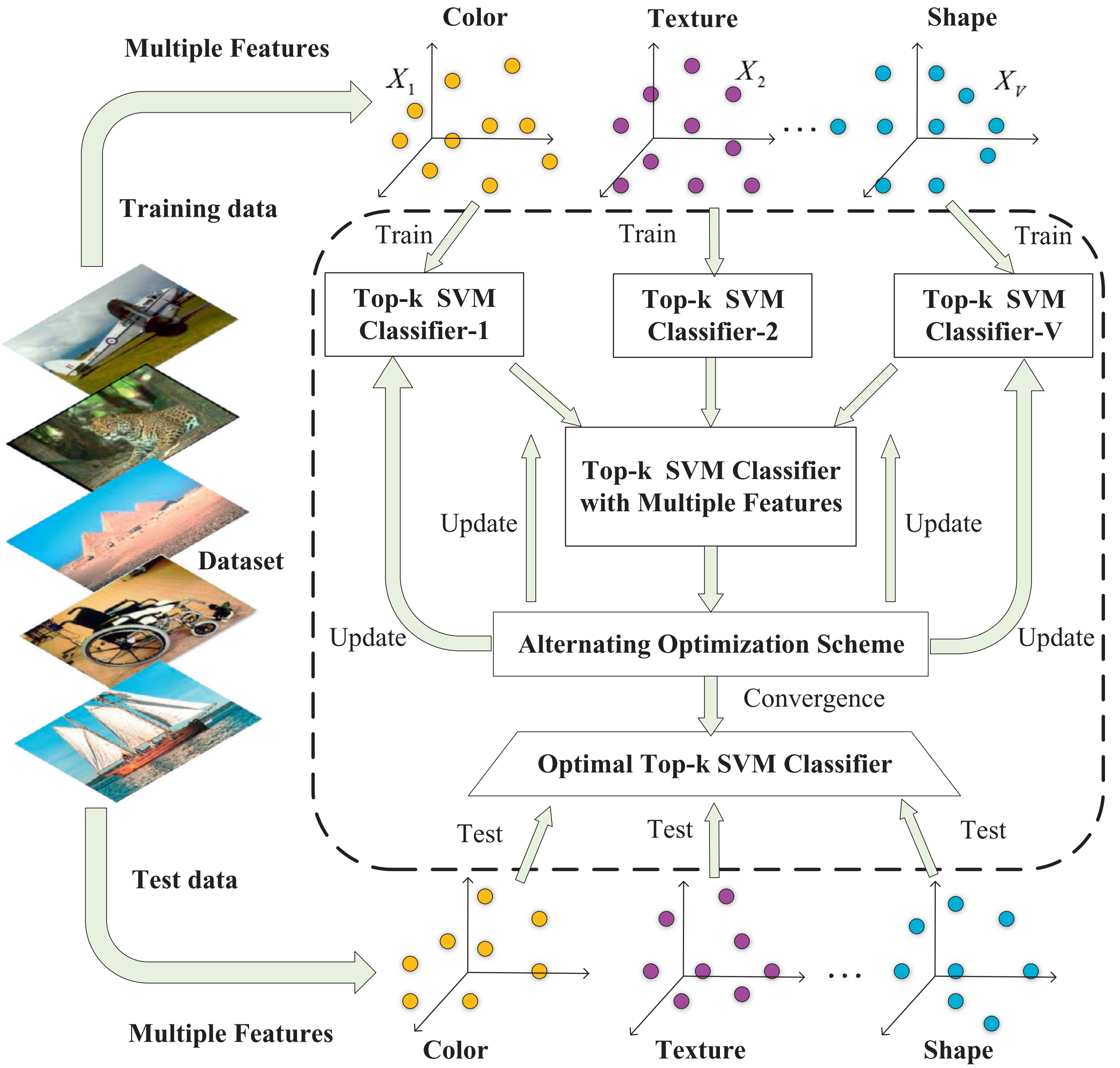


Fig. 1. The framework of the proposed method.

multi-class classification. On the other hand, it is very complicated to learn an effective multi-view classifier which takes both the consistency and complementarity of various features into consideration. Traditional multi-class SVM [2,8] assumes that an object is correctly classified if and only if the category with the largest prediction score is exactly identical with its ground truth label; Otherwise, the classifier wound be penalized for making a mistake. This assumption is usually appropriate for classification with fewer categories. However, for complex classification with massive categories, some classes might be highly confusable [12]. For example, the number of bird categories in ImageNet dataset is up to 861. Two of them, *i.e.*, hawk and eagle, are particularly hard to distinguish since they are extremely similar to each other in many aspects. These confusable classes bring much noise which deteriorate the training process of a multi-class classifier. In fact, the traditional definition of penalty is too strict and might suffer from the problem of over-fitting for a real-world application. Recently, Lapin et al. [20,21] improved this issue by considering the top-*k* predictions rather than only the top one. However, this approach limits to the situation of single feature, which neglects the enormous role of fusing multiple features in improving classification accuracy.

In this paper, we address the issue of combining multiple diverse features in the framework of top-*k* SVM classification. For a better understanding, we demonstrate the overall framework of the proposed method in Fig. 1. We extract multiple features from visual data to provide multi-modal information, and then conduct feature selection on each feature to characterize the intrinsic structure and reduce computational complexity simultaneously [31–33]. Top-k SVM classifier is learned on each single feature independently to alleviate the class ambiguity problem. Each single-view classifier is further associated with a weight to reveal its particular contribution to the final decision in classification. In the training stage, single-view classifiers and their corresponding weights are optimized alternatively. When predicting which category a test sample belongs to, the predictive results of all the single-view classifiers are combined with the learned weights, which falls into a late feature fusion method. Through comparing to the previous studies, we summarize the main contributions of our method as follows:

- Instead of considering the top one prediction only, we propose a novel classification model based on multi-modal feature fusion to mitigate the problem of class ambiguity, which allows k guesses and is not penalized for the first $k-1$ mistakes.
- To combine all top-k single-view classifiers effectively, we allow the classifier learned on each view to participate with an adaptive weight which reflects how "informative" the corresponding view is.
- We exploit an alternating optimization scheme to solve the proposed non-convex problem. Extensive experimental results on several benchmark datasets verify the effectiveness of this algorithm.

The remainder of this paper is organized as follows. In Section 2, a brief review of the related studies is introduced about top-k multi-class SVM and multiple feature learning. We propose in Section 3 a novel multiple feature learning method based on top-k multi-class SVM. The corresponding theoretical analyses and optimization process of the proposed model are demonstrated in Section 4. Extensive experiments are conducted over four benchmark datasets to validate the effectiveness of the proposed algorithm in Section 5. Section 6 concludes our work.

Notations and Definitions: Throughout this paper, we utilize lowercase characters to denote real, bold lowercase characters to denote vectors and uppercase characters to denote matrix. For any c-dimension vector $\mathbf{a}\in\mathbb{R}^c$, $\mathbf{a}_{[k]}(k\le c)$ refers to its kth largest entry. For an arbitrary matrix $A\in\mathbb{R}^{n\times m}$, we denote its ith row and the jth column as $\mathbf{a}^i$ and $\mathbf{a}_j$ respectively. A_{ij} stands for the element corresponding to the ith row and jth column of matrix A. We denote $\mathbf{0}$ and $\mathbf{1}$ as column vectors whose elements are all zero and one, respectively. $\mathbf{e}_i$ is the ith canonical basis vector. The Frobenius norm of matrix A is defined as $\|A\|_F=\sqrt{\sum_i\sum_j a_{ij}^2}=\sqrt{Tr(AA^T)}$. We utilize $\mathbb{1}_S$ to represent the indicator function whose value is 1 if S is ture, and 0 otherwise.

2. Related works

In this section, we briefly review the related works on top-k multi-class SVM and multiple feature learning for visual classification.

2.1. Multiple feature learning

Visual data essentially have a variety of features capturing heterogeneous characteristics. It is widely accepted that combining a set of complementary features can enhance the performance of classification [11]. Among the previous studies, there are two major ways for multiple feature combination, namely early fusion and late fusion [45]. Early fusion combines different features before the training process. A simple way of early fusion is to concatenate distinct feature vectors as the final representation; however, this method may potentially deteriorate each feature's statistical property and limit the exploitation of information from different features [34]. A kernel sparse coding method was therefore developed to address the feature representation problem in classification [25]. Canonical Correlation Analysis (CCA) based early fusion is another classical approach which integrates two types of features through cross-correlating them [16,41]. Recently, Lin et al. [24] conducted feature encoding using a heterogeneous structure fusion (LVFC-HSF) algorithm to capture the intrinsic structure for image representation. Zhong et al. [49] developed a hashing method to fuse multiple features for hyper-spectral imagery classification. This algorithm can transform a high-dimensional float-type feature into extremely low bit binary codes while maintaining the performance. Liu et al. [28] proposed a projective dictionary learning framework for weakly paired multi-modal data fusion which efficiently produced representation vector for data from each modality and fused them by learning the latent pairing relation automatically. These methods are representative of the state-of-the-art in early fusion.

Late fusion combines the predictive values regarding to each feature after the training procedure. For example, Farquhar et al. [9] proposed SVM-2K for feature learning via combining two distinct stages of kernel CCA and SVM into a single optimization. Multiple Kernel Learning (MKL) [19,37,50] techniques were also exploited for late fusion via associating a kernel to each feature and establishing more interpretable decision function based on convex combination of multiple kernels. Recently, a SD-MKL algorithm was proposed to learn score-distributions on the score curves generated from independent data, and then put them into multi-kernel support vector machine (MKSVM) [13]. Li et al. [23] developed a hyperspectral image (HSI) classification method by the probabilistic fusion of pixel and super-pixel level classifiers in a maximum estimation model. Nevertheless, all the methods mentioned above fail to consider the particular contribution of different features for classification.

Please cite this article as: C. Yan et al., Top-k multi-class SVM using multiple features, Information Sciences (2017), http://dx.doi.org/10.1016/j.ins.2017.08.004

2.2. Top-k multi-class SVM

Previous studies on multi-class SVM classification usually adopt top-1 loss, which focuses on the category corresponding to the largest prediction score only [2,8]. For example, Crammer and Singer [8] presented such a method which only compared the true label with the largest one prediction in multi-class SVM classification. Chang and Lin [2] implemented a Library for Support Vector Machines, namely LIBSVM, on the basis of top-1 loss function. Nevertheless, these methods neglect the problem that several classes are likely to have intersections with each other in a multi-class setting [12]. When distinguishing between them, it is inevitably to make mistakes even for a person, not to mention a classifier. Top-*k* multi-class SVM [21] is just proposed to solve this problem by relaxing the penalty for making mistakes. In the framework of top-*k* multi-class SVM, the classifier won't be penalized unless the ground truth label is outside the range of the largest *k* predictions. It was developed on the basis of ranking methods initially [6,35]. Usunier et al. [40] assigned different weights to a ranking of losses corresponding to all the classes and then obtained their weighted sum as the total loss to highlight the top ranked elements. Weston et al. [43] put forward a method by only optimizing the precision at the top of ranked list for image annotation. Swersky et al. [39] developed a probabilistic *n*-Choose-*k* model for multi-class classification which explicitly included a prior distribution together with a count-conditional likelihood. In light of the above methods, Lapin et al. [20] extended top-1 guess to top-*k* guesses, which improved the classification accuracy to some extent by relaxing the penalty. Moreover, Lapin et al. also introduced a top-*k* hinge loss to modify the traditional multi-class SVM and provided efficient optimization schemes in [21]. Chang et al. extend their framework into a robust version in [5]. However, the above-mentioned studies are designed for single modality only and lack the exploitation of multi-modal features, resulting in a limited performance.

3. The proposed methodology

Supposing that the training dataset X consists of n visual samples and each sample is represented by $V(V \geq 2)$ different features. Let $X_v = \{(\mathbf{x}_i^v, y_i) : i = 1, 2, \cdots, n\}$ be the training dataset regarding to the vth feature $(v = 1, 2, \cdots, V)$, where $\mathbf{x}_i^v \in \mathbb{R}^{d^v}$ is the feature representation corresponding to the vth view of the ith training sample; d^v stands for the dimension of the vth feature. $y_i \in \mathcal{Y} := \{1, 2, \cdots, c\}$ means that the ith training sample belongs to the y_ith category; c refers to the number of categories. Let $\mathbf{w}_j^v \in \mathbb{R}^{d^v}$ be the linear classifier learned for the jth category and the vth feature. $W_v = [\mathbf{w}_1^v, \mathbf{w}_2^v, \cdots, \mathbf{w}_c^v] \in \mathbb{R}^{d^v \times c}$ collects all the classifiers corresponding to the vth feature. In the framework of linear SVM, the inner product $\langle \mathbf{w}_j^v, \mathbf{x}_i^v \rangle$ is usually adopted to specify the score of the ith data point belonging to the jth category. For each test sample $\mathbf{x}^v$, we sort its prediction scores in descending order as

$$\langle \mathbf{w}_{[1]}^v, \mathbf{x}^v \rangle \geq \langle \mathbf{w}_{[2]}^v, \mathbf{x}^v \rangle \geq \cdots, \geq \langle \mathbf{w}_{[c]}^v, \mathbf{x}^v \rangle \tag{1}$$

where the bracket $[\cdot]$ denotes a permutation of labels such that $[j]$ is the index of the jth largest score. Note that $(W_v^\top \mathbf{x}_i^v)_{[k]} = \langle \mathbf{w}_{[k]}^v, \mathbf{x}_i^v \rangle$ is valid for $i = 1, 2, \cdots, n$.

Traditional multi-class SVM recognizes the correct category with the top-1 confusing class, *i.e.*, for a test sample $\mathbf{x}_i^v$, the desirable label $\hat{y}$ is obtained according to the max-true rule $\hat{y} = \arg\max_{y \in \mathcal{Y}} \langle \mathbf{w}_y^v, \mathbf{x}_i^v \rangle$. That is to say, the conventional multi-class SVM assumes the predicted label is correct for test sample $\mathbf{x}_i^v$ only when its ground truth label y_i satisfies $\langle \mathbf{w}_{y_i}^v, \mathbf{x}_i^v \rangle > \langle \mathbf{w}_y^v, \mathbf{x}_i^v \rangle$ $(\forall y \neq y_i)$ which turns to

$$\langle \mathbf{w}_{y_i}^v, \mathbf{x}_i^v \rangle > \max_{y \neq y_i} \{\langle \mathbf{w}_y^v, \mathbf{x}_i^v \rangle\} = (W_v^{\backslash y_i \top} \mathbf{x}_i^v)_{[1]}, \tag{2}$$

where the matrix $W_v^{\backslash y_i} \in \mathbb{R}^{d^v \times (c-1)}$ removes the y_ith column of $W_v \in \mathbb{R}^{d^v \times c}$. Based on this criterion, the conventional multi-class SVM is formulated to minimize the following regularized empirical risk $\min_{W_v \in \mathbb{R}^{d^v \times c}} \sum_{i=1}^n l_i^1(W_v) + \lambda \|W_v\|_F^2$, where the top-1 loss function $l_i^1(W_v)$ is defined as

$$\begin{aligned} l_i^1(W_v) &= \max_{y \in \mathcal{Y}} \left\{ \mathbb{1}_{y \neq y_i} + \langle \mathbf{w}_y^v, \mathbf{x}_i^v \rangle - \langle \mathbf{w}_{y_i}^v, \mathbf{x}_i^v \rangle \right\} \\ &= \max \left\{ 0, 1 + (W_v^{\backslash y_i \top} \mathbf{x}_i^v)_{[1]} - \langle \mathbf{w}_{y_i}^v, \mathbf{x}_i^v \rangle \right\}, \end{aligned} \tag{3}$$

for the vth view of the ith training sample, where the second equality holds according to inequality (2). Note that the top-1 assumption in conventional multi-class SVM is too strict in practice due to the multi-label nature of visual example and severe class overlapping problem [12]. Recently, Lapin et al. [20] pioneered top-k $(k \geq 1)$ multi-class SVM via allowing each test sample $\mathbf{x}_i^v$ associating with a set of the k most related labels, denoted by $\hat{Y}$, *i.e.*,

$$\hat{Y} = \arg \max_{Y \subseteq \mathcal{Y}, |Y| = k} \sum_{j \in Y} \langle \mathbf{w}_j^v, \mathbf{x}_i^v \rangle. \tag{4}$$

In such a way, top-*k* multi-class SVM considers the predicted label of $\mathbf{x}_i^v$ is correct as long as its prediction score is among the top-*k* largest scores; in other words, its ground truth label y_i satisfies $\langle \mathbf{w}_{y_i}^v, \mathbf{x}_i^v \rangle \geq (W_v^\top \mathbf{x}_i^v)_{[k]}$. Consequently, top-*k* multi-class classification separates the correct class using the top-*k* confusing category and ignores the $k-1$ most confusing categories, which gives some slack for conventional top-1 multi-class classification. This strategy is more reasonable in visual category recognition with massive categories because image/video is usually associated with extremely many labels and these labels inevitably overlap a lot. Correspondingly, the top-k loss function of multi-class SVM is extended as

$$\begin{aligned} l_i^k(W_v) &= \max\left\{0, 1 + (W_v^\top \mathbf{x}_i^v)_{[k]} - \langle \mathbf{w}_{y_i}^v, \mathbf{x}_i^v \rangle\right\} \\ &= \max\left\{0, (\mathbf{1} + (W_v^\top - \mathbf{1}\mathbf{w}_{y_i}^{v\,\top})\mathbf{x}_i^v - \mathbf{e}_{y_i})_{[k]}\right\} \\ &= \max\left\{0, \mathbf{s}^i_{[k]}\right\}, \end{aligned} \tag{5}$$

where the abbreviation $\mathbf{s}^i = \mathbf{1} + (W_v^\top - \mathbf{1}\mathbf{w}_{y_i}^\top)\mathbf{x}_i^v - \mathbf{e}_{y_i} \in \mathbb{R}^c$ is used to specify the score vector of sample $\mathbf{x}_i^v$ for better representation. Note that the conventional loss function in (3) becomes a special case of (5) when the value of k is set to 1.

Due to the non-convexity of $l_i^k(W_v)$, we first attempt to adopt the following convex upper bound to approximate the top-k loss, $\bar{l}_i^k(W_v) = \max\left\{0, \sum_{j=1}^k \mathbf{s}^i_{[j]}\right\}$. Nevertheless, this approximation may still be dominated by the top-1 confusing class; moreover, it ignores the significant influence of outliers which make $\|W_v\|_F$ grow unbounded [48]. In light of this situation, Lapin et al. [20] utilized top-k hinge loss, *i.e.*, $\tilde{l}_i^k(W_v) = \max\left\{0, \frac{1}{k}\sum_{j=1}^k \mathbf{s}^i_{[j]}\right\}$, as another convex substitute for top-k loss. Note that the above mentioned loss functions satisfy

$$l_i^k(W_v) \leq \tilde{l}_i^k(W_v) \leq \bar{l}_i^k(W_v), \tag{6}$$

for any $k \geq 1$. The top-k hinge loss $\tilde{l}_i^k(W_v)$, which is chosen for our model, is a tighter upper bound to make a better approximation of the top-k loss.

For better representation, we define $L_v^k = \frac{1}{n}\sum_{i=1}^n \tilde{l}_i^k(W_v; \mathbf{x}_i^v)$ as the top-k hinge loss with respect to the vth view, and $L^k = [L_1^k, L_2^k, \cdots, L_V^k]$ collects all the loss functions. In order to characterize the importance of multiple views, we associate a weight in $\alpha = [\alpha_1^r, \alpha_2^r, \cdots, \alpha_V^r]$ with the corresponding loss function in L^k. Based on the above-mentioned derivation, we can formulate the following multi-modal optimization problem:

$$\begin{aligned} &\min_{W_v, \alpha_v (\forall v)} \sum_{v=1}^V (\alpha_v^r L_v^k + \frac{\lambda}{2}\|W_v\|_F^2) \\ &\quad s.t. \quad \sum_{v=1}^V \alpha_v = 1, \;\; \alpha_v \geq 0, \;\; v = 1, 2, \cdots, V, \end{aligned} \tag{7}$$

where $r > 1$. Specifically, the total loss function $\sum_{v=1}^V \alpha_v^r L_v^k$ is used to control the classification error. The second term is a Frobenius norm with respect to W_v, *i.e.*, $\|W_v\|_F^2$, to prevent from the trivial solution. λ is introduced to balance the loss function and regularization term. Through minimizing the objective function in (7), we aim to get the optimal solution of $W_v(\forall v)$ and $\alpha_v(\forall v)$ corresponding to each view. As for a testing data $\mathbf{x}$ with multiple features, the category with respect to the largest score is selected in the prediction vector

$$\mathbf{y} = \sum_{v=1}^V \alpha_v^r W_v \mathbf{x}^v. \tag{8}$$

In summary, the optimal classifier obtained in our method takes into consideration not only the class ambiguity issue but also the reasonable combination of multi-modal features. Comparing with the traditional methods, our model is capable of boosting the classification accuracy significantly.

4. Alternating optimization

In this section, we introduce the optimization process of the proposed method in detail. It's noteworthy that the primal objective function in (7) is convex in W_v only or α_v only, but not convex in both variables together. To the best of our knowledge, there is no direct way to find the global minimum due to its non-convexity. Considering the independence of variables, we address the proposed problem with alternating optimization algorithm. The overall process is shown in Algorithm 1.

4.1. *Update weight vector*

With fixed variables $W_v(\forall v)$, the weight parameter $\alpha_v(\forall v)$ corresponding to each view is updated by solving the following optimization problem

$$\min_{\alpha_v(\forall v)} \sum_{v=1}^V \alpha_v^r L_v^k \qquad s.t. \quad \sum_{v=1}^V \alpha_v = 1, \;\; \alpha_v \geq 0, \;\; v = 1, 2, \cdots, V. \tag{9}$$

Please cite this article as: C. Yan et al., Top-k multi-class SVM using multiple features, Information Sciences (2017), http://dx.doi.org/10.1016/j.ins.2017.08.004

Algorithm 1 Alternating optimization.

```
Input: parameters: $k$, $\lambda$ and $r(r > 1)$;training data: $X_v = \{(\mathbf{x}_i^v, y_i) : i = 1, 2, \cdots, n\}$ , $v = 1, 2, \cdots, V$;
Output: primal variables: $W_v(\forall v)$;weights: $\alpha_v(\forall v)$
Initialize : $W_v(\forall v) \leftarrow 0$, $\alpha = [1/V, \cdots, 1/V]$
repeat
    for $v = 1$ to $V$ do
        update primal variables $W_v$ with $\alpha_v$ fixed (see Algorithm 2 for details)
        update weights $\alpha_v$ with $W_v$ fixed according to Eq. (12)
    end for
until convergence
```

To this end, we introduce a Lagrange multiplier β and reformulate the optimization problem above as

$$L(\alpha_v, \beta) = \sum_{v=1}^{V} \alpha_v^r L_v^k - \beta\left(\sum_{v=1}^{V} \alpha_v - 1\right). \tag{10}$$

For the ith view, we calculate the partial derivatives with respect to variables α_v and β respectively by

$$\begin{cases} \frac{\partial L(\alpha_v, \beta)}{\partial \alpha_v} = L_v^k r \alpha_v^{r-1} - \beta \\ \frac{\partial L(\alpha_v, \beta)}{\partial \beta} = \sum_{v=1}^{V} \alpha_v - 1 \end{cases}, \tag{11}$$

where $v = 1, 2, \cdots, V$. According to the KKT conditions, we set both the two partial derivatives in Eq. (11) to zero and arrive at

$$\alpha_v = \frac{[1/(nL_v^k)]^{\frac{1}{r-1}}}{\sum_{v=1}^{V} [1/(nL_v^k)]^{\frac{1}{r-1}}} \quad (v = 1, 2, \cdots, V), \tag{12}$$

where parameter $r > 1$ controls the distribution of weight vector α. It makes α_v negatively correlate with its corresponding loss L_v^k. In other words, the view with a larger loss will have a smaller impact on the final decision, which is reasonable for classification theoretically. For simplicity, we mainly focus on the two extreme cases $r \to 1$ and $r \to \infty$. If $r \to 1$, the optimal solution of the weight vector α is $\alpha_v = 1$ for the view regarding to the minimum loss L_v^k, and $\alpha_v = 0$ for all the other views. For the second extreme case $r \to \infty$, all the weights in α wound be quite close to each other and approximately equal to $1/V$, which can be denoted as $[1/V,\ 1/V,\ \cdots,\ 1/V]$. In general, the value of r should be determined by the degree of complementarity among various views. Larger r is suitable for rich complementary information; Otherwise, smaller r should be selected.

4.2. *Update primal variables*

Considering the complexity of minimizing primal objective function, we employ a dual optimization method, termed as Fenchel Duality, to update all the single-view classifiers $W_v(\forall v)$ with fixed variable $\alpha_v(\forall v)$. For a better understanding, we first introduce some lemmas and theorems in the framework of multi-view learning.

Definition 1. The Fenchel Conjugate of multi-variable function $G : X_1 \times X_2 \times \cdots \times X_V \to \mathbb{R}$ is defined as

$$G^*(\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_V) = \sup_{\mathbf{x}_i(\forall i)} \left\{ \sum_{i=1}^{V} \langle \mathbf{y}_i, \mathbf{x}_i \rangle - G(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_V) \right\}, \tag{13}$$

where $\mathbf{y}_1$, $\mathbf{y}_2$, $\cdots$, $\mathbf{y}_V$ are conjugate variables.

Lemma 1. *Assuming that function $G(\mathbf{x}_1,\ \mathbf{x}_2,\ \cdots,\ \mathbf{x}_V)$ can be decomposed into the weighted sum of multiple functions,* i.e., $G(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_V) = \sum_{i=1}^{V} \alpha_i g(\mathbf{x}_i)$. *The Fenchel Conjugate function of $G(\mathbf{x}_1,\ \mathbf{x}_2,\ \cdots,\ \mathbf{x}_V)$ can be written as*

$$G^*(\alpha_1\mathbf{y}_1, \alpha_2\mathbf{y}_2, \cdots, \alpha_V\mathbf{y}_V) = \sum_{i=1}^{V} \alpha_i g^*(\mathbf{y}_i), \tag{14}$$

where $g^*(\mathbf{y}_i)(\forall i)$ is the Fenchel Conjugate function of $g(\mathbf{x}_i)(\forall i)$.

Proof. According to Definition 1, we can obtain the Fenchel Conjugate function of $G(\mathbf{x}_1,\ \mathbf{x}_2,\ \cdots,\ \mathbf{x}_V)$ by

$$G^*(\alpha_1\mathbf{y}_1, \alpha_2\mathbf{y}_2, \cdots, \alpha_V\mathbf{y}_V) = \sup_{\mathbf{x}_i(\forall i)} \left\{ \sum_{i=1}^{V} \langle \alpha_i\mathbf{y}_i, \mathbf{x}_i \rangle - G(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_V) \right\}. \tag{15}$$

Then we use $\sum_{i=1}^{V} \alpha_i g(\mathbf{x}_i)$ to replace $G(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_V)$ in Eq. (15) as

$$\sup_{\mathbf{x}_i(\forall i)} \left\{ \sum_{i=1}^{V} \langle \alpha_i \mathbf{y}_i, \mathbf{x}_i \rangle - \sum_{i=1}^{V} \alpha_i g(\mathbf{x}_i) \right\}. \tag{16}$$

By exchanging operation order, formula (16) turns to be $\sum_{i=1}^{V} \alpha_i (\sup_{\mathbf{x}_i} \{\langle \mathbf{y}_i, \mathbf{x}_i \rangle - g(\mathbf{x}_i)\})$, where $\sup_{x_i} \{\langle \mathbf{y}_i, \mathbf{x}_i \rangle - g(\mathbf{x}_i)\}$ is the Fenchel Conjugate function of $g(\mathbf{x}_i)$, *i.e.*, $g^*(\mathbf{y}_i)$. Consequently, the Fenchel Conjugate function of $G(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_V)$ can be derived as $\sum_{i=1}^{V} \alpha_i g^*(\mathbf{y}_i)$. The proof is completed. □

Theorem 1. *Let $A_v = [\mathbf{a}_1^v, \mathbf{a}_2^v, \cdots, \mathbf{a}_n^v] \in \mathbb{R}^{c \times n}$ be the matrix consisting of all the dual variables $\mathbf{a}_i^v \in \mathbb{R}^c$, each of which is regarding to the vth feature of the ith sample. The Fenchel dual objective function of the primal objective function in (7) is formulated as*

$$\begin{aligned} \max_{A_v(\forall v)} & -\sum_{v=1}^{V} \alpha_v^r \frac{1}{n} \sum_{i=1}^{n} \widetilde{l}_i^{k*}(-\lambda n(\mathbf{a}_i^v - \mathbf{a}_{y_i,i}^v \mathbf{e}_{y_i})) - \sum_{v=1}^{V} \frac{\lambda}{2} \|X_v A_v^\top\|_F^2 \\ & s.t. \ \ \langle \mathbf{1}, \mathbf{a}_i^v \rangle = 0, \ \ v = 1, 2, \cdots, V, \end{aligned} \tag{17}$$

where $\widetilde{l}_i^{k*}$ is the Fenchel Conjugate function of $\widetilde{l}_i^k$ and we have $W_v = X_v A_v^\top, \ v = 1, 2, \cdots, V$.

Proof. The primal objective function $P(\{W_v\}_{v=1}^V)$ in Eq. (7) can be separated into two parts

$$\begin{cases} G(\{X_v^\top W_v\}_{v=1}^V) = \sum_{v=1}^{V} \alpha_v^r L_v^k \\ F(\{W_v\}_{v=1}^V) = \sum_{v=1}^{V} \frac{\lambda}{2} \|W_v\|_F^2 \end{cases}. \tag{18}$$

According to the definition of Fenchel Duality [1, Theorem 3.3.5], we can formulate the dual function of $P(\{W_v\}_{v=1}^V)$ as

$$D(\{A_v\}_{i=1}^V) = -G^*(\{-\alpha_v^r A_v^\top\}_{v=1}^V) - F^*(\{X_v A_v^\top\}_{v=1}^V), \tag{19}$$

where $G^*(\{-\alpha_v^r A_v^\top\}_{v=1}^V)$ and $F^*(\{X_v A_v^\top\}_{v=1}^V)$ are the Fenchel Conjugate functions of $G(\{X_v^\top W_v\}_{v=1}^V)$ and $F(\{W_v\}_{v=1}^V)$ respectively. The primal and dual variables satisfy $W_v = X_v A_v^\top, \ v = 1, 2, \cdots, V$.

From Lemma 1, the Fenchel Conjugate function of $G(\{X_v^\top W_v\}_{v=1}^V)$ can be formulated as $\sum_{v=1}^{V} \alpha_v^r L_v^{k*}$, where L_v^{k*} is the Fenchel Conjugate function of L_v^k. According to the derivation of L_v^{k*} (see [20, Theorem 1]), we can obtain

$$G^*(\{-\alpha_v^r A_v^\top\}_{v=1}^V) = \begin{cases} \sum_{v=1}^{V} \alpha_v^r \frac{1}{n} \sum_{i=1}^{n} l_i^{k*}(-n(\mathbf{a}_i^v - \mathbf{a}_{y_i,i}^v \mathbf{e}_{y_i})), & \langle \mathbf{1}, \mathbf{a}_i^v \rangle = 0 \\ +\infty, & otherwise \end{cases}. \tag{20}$$

Based on the fact that the Fenchel Conjugate of $\frac{\lambda}{2}\|W_v\|_F^2$ is $\frac{\lambda}{2}\|\frac{1}{\lambda} X_v A_v^\top\|_F^2$, we can derive the Fenchel Conjugate of $F(\{W_v\}_{v=1}^V)$ as

$$F^*(\{X_v A_v^\top\}_{v=1}^V) = \sum_{v=1}^{V} \frac{\lambda}{2} \|\frac{1}{\lambda} X_v A_v^\top\|_F^2. \tag{21}$$

Setting $A \leftarrow \frac{A}{\lambda}$ for convenience, the Fenchel dual objective function $D(\{A_v\}_{i=1}^V)$ can be obtained by combining the two Fenchel Conjugate functions in Eqs. (20) and (21) as $-\sum_{v=1}^{V} \alpha_v^r \frac{1}{n} \sum_{i=1}^{n} \widetilde{l}_i^{k*}(-\lambda n(\mathbf{a}_i^v - \mathbf{a}_{y_i,i}^v \mathbf{e}_{y_i})) - \sum_{v=1}^{V} \frac{\lambda}{2} \|X_v A_v^\top\|_F^2$ when $\langle \mathbf{1}, \mathbf{a}_i^v \rangle = 0 (\forall v)$; and $-\infty$ otherwise. The proof is completed. □

Based on Fenchel Dual theory, we have transform the minimization of primal objective function into the maximization of dual objective function. Specifically, we exploit proximal stochastic dual coordinate ascent (Prox-SDCA) proposed by Shalev-Shwartz and Zhang [38] as our optimization scheme. At the beginning of each epoch, we randomly sort all the samples in training set to make the convergence faster. During each iteration, only one sample $\mathbf{x}_i$ is utilized to update its corresponding dual variables $\mathbf{a}_i^v (\forall v)$ while fixing other dual variables. Via updating dual variables $\mathbf{a}_i^v (\forall v)$, the corresponding primal variables can be updated according to their inherent relation $W_v = X_v A_v^\top$. Then both the primal and dual objective function values should be recalculated to check the stopping criterion whether the relative dual gap is below ε. All processes are summarized in Algorithm 2. However, it is extremely hard to get the optimal solution of $D(\{A_v\}_{i=1}^V)$ directly. Next, we attempt to transform it into an equivalent problem which is well-defined and easy to solve.

Theorem 2. *The problem $\max_{\mathbf{a}_i^v(\forall v)} \left\{ D(\{A_v\}_{i=1}^V) | \langle \mathbf{1}, \mathbf{a}_i^v \rangle = 0, \ v = 1, 2, \cdots, V \right\}$ is equivalent to*

$$\min_{\mathbf{z}^v(\forall v)} \left\{ \sum_{v=1}^{V} (\|\mathbf{b}^v - \mathbf{z}^v\|^2 + \langle \mathbf{1}, \mathbf{z}^v \rangle^2) | \mathbf{z}^v \in \Delta_k(\frac{1}{\lambda n}), \ v = 1, 2, \cdots, V \right\}, \tag{22}$$

under the condition that $\mathbf{a}_{\backslash y_i,i}^v = -\mathbf{z}^v$ and $\mathbf{a}_{y_i,i}^v = \langle \mathbf{1}, \mathbf{z}^v \rangle$, where $\mathbf{a}_{\backslash y_i,i}$ is obtained by removing the y_ith coordinate from $\mathbf{a}_i$. Moreover, $\mathbf{b}^v$ can be represented as $\frac{\mathbf{1}(\alpha_v^r - \mathbf{q}_{y_i}^v) + \mathbf{q}_{\backslash y_i}^v}{K_{ii}^v}$, where $\mathbf{q}^v = W_v^\top \mathbf{x}_i^v - \langle \mathbf{x}_i^v, \mathbf{x}_i^v \rangle \mathbf{a}_i^v$, and $K^v = X_v^\top X_v$. $\Delta_k(\frac{1}{\lambda n}) \triangleq \left\{ \mathbf{u} \mid \langle \mathbf{1}, \mathbf{u} \rangle \le \frac{1}{\lambda n}, 0 \le \mathbf{u}_i \le \frac{1}{k} \langle \mathbf{1}, \mathbf{u} \rangle, i = 1, 2, \cdots, c \right\}$, termed as top-k simplex.

Please cite this article as: C. Yan et al., Top-*k* multi-class SVM using multiple features, Information Sciences (2017), http://dx.doi.org/10.1016/j.ins.2017.08.004

Algorithm 2 Top-k multi-class SVM using single feature.

```
Input: parameters $k$, $\lambda$, $\varepsilon$;training data: $X_v = \{(\mathbf{x}_i^v, y_i) : i = 1, 2, \cdots, n\}, \ \ v = 1, 2, \cdots, V$;
Output: primal variables: $W_v = [\mathbf{w}_1^v, \mathbf{w}_2^v, \cdots, \mathbf{w}_c^v] \in \mathbb{R}^{d^v \times c}$;dual variables:$A_v = [\mathbf{a}_1^v, \mathbf{a}_2^v, \cdots, \mathbf{a}_n^v] \in \mathbb{R}^{c \times n}, \ \ v = 1, 2, \cdots, V$;
Initialize : $W_v \leftarrow 0$, $A_v \leftarrow 0$, $v = 1, 2, \cdots, V$
repeat
    Randomly arrange the order of training data
    for $i = 1$ to $n$ do
        get the prediction scores corresponding to each view: $\mathbf{s}_i^v \leftarrow W_v^\top \mathbf{a}_i^v$
        store previous values of the dual variables regarding to the $i$th sample: $\mathbf{a}_i^{v(old)} \leftarrow \mathbf{a}_i^v$
        update dual variables of each view (see Theorem 2 for details): $\mathbf{a}_i^v \leftarrow update(k, \lambda, \|\mathbf{x}_i^v\|^2, y_i, \mathbf{s}_i^v, \mathbf{a}_i^v)$
        update primal variables of each view: $W_v \leftarrow W_v + \mathbf{x}_i^v(\mathbf{a}_i^v - \mathbf{a}_i^{v(old)})^\top$
    end for
until relative dual gap is below $\varepsilon$
```

Proof. At the ith iteration, our method only updates the dual variables $\mathbf{a}_i^v(\forall v)$ regarding to the ith sample. The dual objective function reduces to

$$\begin{aligned} &\max_{\mathbf{a}_i^v(\forall v)} -\sum_{v=1}^{V}(\alpha_v^r \frac{1}{n}\widetilde{l}_i^{k*}(-\lambda n(\mathbf{a}_i^v - \mathbf{a}_{y_i,i}^v \mathbf{e}_{y_i})) + \frac{\lambda}{2}\|X_v A_v^\top\|_F^2) \\ &\quad s.t. \ \ \langle \mathbf{1}, \mathbf{a}_i^v \rangle = 0, \ \ v = 1, 2, \cdots, V. \end{aligned} \tag{23}$$

The first item $-\sum_{v=1}^{V}\alpha_v^r \frac{1}{n}\widetilde{l}_i^{k*}(-\lambda n(\mathbf{a}_i^v - \mathbf{a}_{y_i,i}^v \mathbf{e}_{y_i}))$ is simplified as $\sum_{v=1}^{V}\alpha_v^r \lambda \mathbf{a}_{y_i,i}^v$ if $-\lambda n(\mathbf{a}_i^v - \mathbf{a}_{y_i,i}^v \mathbf{e}_{y_i}) \in \Delta_k(1)(\forall v)$, and $+\infty$ otherwise (see [20, Proposition 2]). The constraint $-\lambda n(\mathbf{a}_i^v - \mathbf{a}_{y_i,i}^v \mathbf{e}_{y_i}) \in \Delta_k(1)(\forall v)$ is equivalent to $-\mathbf{a}_{\backslash y_i,i}^v \in \Delta_k(\frac{1}{\lambda n})$ and $\mathbf{a}_{y_i,i}^v = \langle \mathbf{1}, -\mathbf{a}_{\backslash y_i,i}^v \rangle (\forall v)$. The top-$k$ simplex $\Delta_k(\frac{1}{\lambda n})$ is a k-dimension polyhedron which is the convex hull of its $k+1$ vertices and is defined as $\Delta_k(\frac{1}{\lambda n}) \triangleq \left\{\mathbf{u} \mid \langle \mathbf{1}, \mathbf{u} \rangle \le \frac{1}{\lambda n}, 0 \le \mathbf{u}_i \le \frac{1}{k}\langle \mathbf{1}, \mathbf{u} \rangle, i = 1, 2, \cdots, c\right\}$, where c is the dimension of $\mathbf{u}$.

Setting $K^v = X_v^\top X_v$ for better representation, the regularization term $\frac{\lambda}{2}\|X_v A_v^\top\|_F^2$ is equivalent to $\frac{\lambda}{2}tr(A_v K^v A_v^\top)$, *i.e.*, $\frac{\lambda}{2}K_{ii}^v \langle \mathbf{a}_i^v, \mathbf{a}_i^v \rangle + \lambda \sum_{j \ne i} K_{ij}^v \langle \mathbf{a}_i^v, \mathbf{a}_j^v \rangle + \frac{\lambda}{2}const$. Assuming that $\mathbf{z}^v = -\mathbf{a}_{\backslash y_i,i}^v$, we can get $\mathbf{a}_{y_i,i}^v = \langle \mathbf{1}, \mathbf{z}^v \rangle$ due to the fact that $\langle \mathbf{1}, \mathbf{a}_i^v \rangle = 0$. It is noteworthy that $\langle \mathbf{a}_i^v, \mathbf{a}_i^v \rangle = \langle \mathbf{1}, \mathbf{z}^v \rangle^2 + \langle \mathbf{z}^v, \mathbf{z}^v \rangle$ and $\langle \mathbf{q}^v, \mathbf{a}_i^v \rangle = \mathbf{q}_{y_i}^v \langle \mathbf{1}, \mathbf{z}^v \rangle - \langle \mathbf{q}_{\backslash y_i}^v, \mathbf{z}^v \rangle$ where $\mathbf{q}^v = \sum_{j \ne i} K_{ij}^v \mathbf{a}_j^v$. $\mathbf{q}^v$ can be calculated using the old $\mathbf{a}_i^v$, *i.e.*, $\sum_{j \ne i} K_{ij}^v \mathbf{a}_j^v = A_v K_i^v - K_{ii}^v \mathbf{a}_i^v = W_v^\top \mathbf{x}_i^v - \langle \mathbf{x}_i^v, \mathbf{x}_i^v \rangle \mathbf{a}_i^v$.

We can obtain the equivalent problem of (23) by collecting all the corresponding items together and multiplying with $-\frac{2Kii^v}{\lambda}$ as

$$\min_{\mathbf{z}^v(\forall v)} \sum_{v=1}^{V}(\mathbf{z}^v)^2 - 2\frac{\mathbf{1}(\alpha_v^r - \mathbf{q}_{y_i}^v) + \mathbf{q}_{\backslash y_i}^v}{K_{ii}^v}\mathbf{z}^v + \langle \mathbf{1}, \mathbf{z}^v \rangle^2. \tag{24}$$

Thanks to the independence of multiple views, we address the problem in (24) via decomposing it into V subproblems. Take the vth view as an example, the combination of the first two terms is a quadratic function with respect to $\mathbf{z}^v$, which reaches its minimum at $\mathbf{z}^v = \frac{\mathbf{1}(\alpha_v^r - \mathbf{q}_{y_i}^v) + \mathbf{q}_{\backslash y_i}^v}{K_{ii}^v}$. The optimal solution of $\mathbf{z}^v$ can be obtained by

$$\min_{\mathbf{z}^v}\left\{\|\frac{\mathbf{1}(\alpha_v^r - \mathbf{q}_{y_i}^v) + \mathbf{q}_{\backslash y_i}^v}{K_{ii}^v} - \mathbf{z}^v\|^2 + \langle \mathbf{1}, \mathbf{z}^v \rangle^2 \mid \mathbf{z}^v \in \Delta_k(\frac{1}{\lambda n})\right\}, \tag{25}$$

which is a regularized projection problems onto the top-k simplex $\Delta_k(\frac{1}{\lambda n})$. Therefore, the above mentioned optimization problem in (24) can be solved via minimizing all its subproblems respectively. The proof is completed. □

Next, we mainly focus on solving the projection problem corresponding to the vth view in (22) and discuss how to divide it into different situations to improve efficiency.

Definition 2. The projection problem in (22) can be divided into the following two situations depending on the relationship between $\langle \mathbf{1}, \mathbf{z}^v \rangle$ and upper bound $\frac{1}{\lambda n}$ according to Lapin et al. [20], that is

$$\begin{cases} \min_{\mathbf{z}^v}\{\|\mathbf{b}^v - \mathbf{z}^v\|^2 + \langle \mathbf{1}, \mathbf{z}^v \rangle^2 \mid \langle \mathbf{1}, \mathbf{z}^v \rangle = \frac{1}{\lambda n}, 0 \le \mathbf{z}_i^v \le \frac{1}{k\lambda n}\} & \langle \mathbf{1}, \mathbf{z}^v \rangle = \frac{1}{\lambda n} \\ \min_{\mathbf{z}^v}\{\|\mathbf{b}^v - \mathbf{z}^v\|^2 + \langle \mathbf{1}, \mathbf{z}^v \rangle^2 \mid 0 \le z_i^v \le \frac{\langle \mathbf{1}, \mathbf{z}^v \rangle}{k}\} & \langle \mathbf{1}, \mathbf{z}^v \rangle < \frac{1}{\lambda n} \end{cases}. \tag{26}$$

When $\langle \mathbf{1}, \mathbf{z}^v \rangle = \frac{1}{\lambda n}$, it is known as the *continuous quadratic knapsack problem* because the upper bound on $\mathbf{z}_i^v$ is a constant. This is a well-defined problem and can be solved by using several highly efficient algorithms. We eventually utilize the method in [17] since it is easy to implement with linear time complexity in practice. If $\langle \mathbf{1}, \mathbf{z}^v \rangle < \frac{1}{\lambda n}$, assuming that $\mathbf{z}^{v*} \in \mathbb{R}^d$

Please cite this article as: C. Yan et al., Top-k multi-class SVM using multiple features, Information Sciences (2017), http://dx.doi.org/10.1016/j.ins.2017.08.004

Table 1
Summary of the benchmark datasets used in our experiments.

Datasets	Caltech101	NUSWIDE	Corel 5k	CCV
View 1	Gabor(48)	CH(65)	CS(32)	SIFT(4000)
View 2	WM(40)	CM(226)	CL(12)	STIP(5000)
View 3	CENTRIST(254)	CORR(145)	EH(80)	MFCC(5000)
View 4	HOG(1984)	SIFT(2000)	RS(35)	–
View 5	GIST(512)	WT(129)	Haar(195)	–
View 6	LBP(928)	–	DC(16)	–
View 7	–	–	SC(64)	–
View 8	–	–	HT(62)	–
Number of Data	9144	30,000	5000	9317
Number of Classes	102	31	50	20
Type of Dataset	Image	Image	Image	Video
Training Data	4000	17,928	2000	4659
Test and Validate Data	5144	12,072	3000	4658

is the optimal solution to the above optimization problem and $B \triangleq \left\{i|\mathbf{z}_i^{v*} = \frac{\langle \mathbf{1},\mathbf{z}^v\rangle}{k}\right\}$, $C \triangleq \left\{i|0 < \mathbf{z}_i^{v*} < \frac{\langle \mathbf{1},\mathbf{z}^v\rangle}{k}\right\}$, $D \triangleq \left\{i|\mathbf{z}_i^{v*} = 0\right\}$ is a partition of $\mathbf{z}^{v*}$. We can obtain three distinct cases which is $B = \varnothing\ C = \varnothing$, $B \neq \varnothing\ C = \varnothing$, and $C \neq \varnothing$ respectively.

To reduce the computation complexity, we first handle the easiest case of zero projection($B = \varnothing\ C = \varnothing$) and constant projection($B \neq \varnothing\ C = \varnothing$). If both of them don't work, we would deal with the continuous quadratic knapsack problem which is more difficult to solve than the former two cases. If that also fails, we have to proceed with the most complex situation($C \neq \varnothing$) which requires sorting and looping over all the feasible partitions of B, C, D until finding an optimal solution.

5. Experiment

In this section, extensive experiments over four real-world datasets are conducted to evaluate the performance of the proposed method. Each dataset has a certain number of features and classes, whose details are summarized in Table 1. Several examples of the four benchmark datasets are illustrated in Fig. 2.

5.1. Dataset description

- *Caltech101* [10] is an image dataset with 102 categories, including 101 object categories and 1 background category. This dataset contains 9144 samples for scene recognition with 6 features extracted from each image: *i.e.*, 48-D Gabor, 40-D wavelet moments (WM), 254-D CENTRIST, 1984-D HOG, 512-D GIST and 928-D LBP. A total of 4000 samples are selected for the training stage, and the rest 5144 samples are used for testing.
- *NUSWIDE* [7] consists of 30,000 images in 31 classes for object recognition, gathered from Flickr. All the 5 published features provided by the website are utilized, *i.e.*, 65-D color Histogram (CH), 226-D color moments (CM), 145-D color correlation (CORR), 74-D edge distribution and 129-D wavelet texture. We select 17928 samples for training and the remaining ones for testing.
- *Corel* [36] is a photo gallery dataset with 599 categories, each of which contains 100 samples. The Haar feature and 7 MPEG-7 visual features of each sample are extracted. In our experiment, we use a subset of the whole dataset including 50 classes, each of which contains 100 samples, namely Corel 5k. In addition, 40 samples of each category are used for training, and the remaining are used for testing.
- *CCV* [15] is the Columbia Consumer Video dataset including 9317 YouTube videos, which belong to 20 semantic categories with the average length of 80 s. Three popular audio/visual features are used, which are 4000-D SIFT, 5000-D STIP, and 5000-D MFCC respectively. We select 4659 samples for training and the rest for testing.

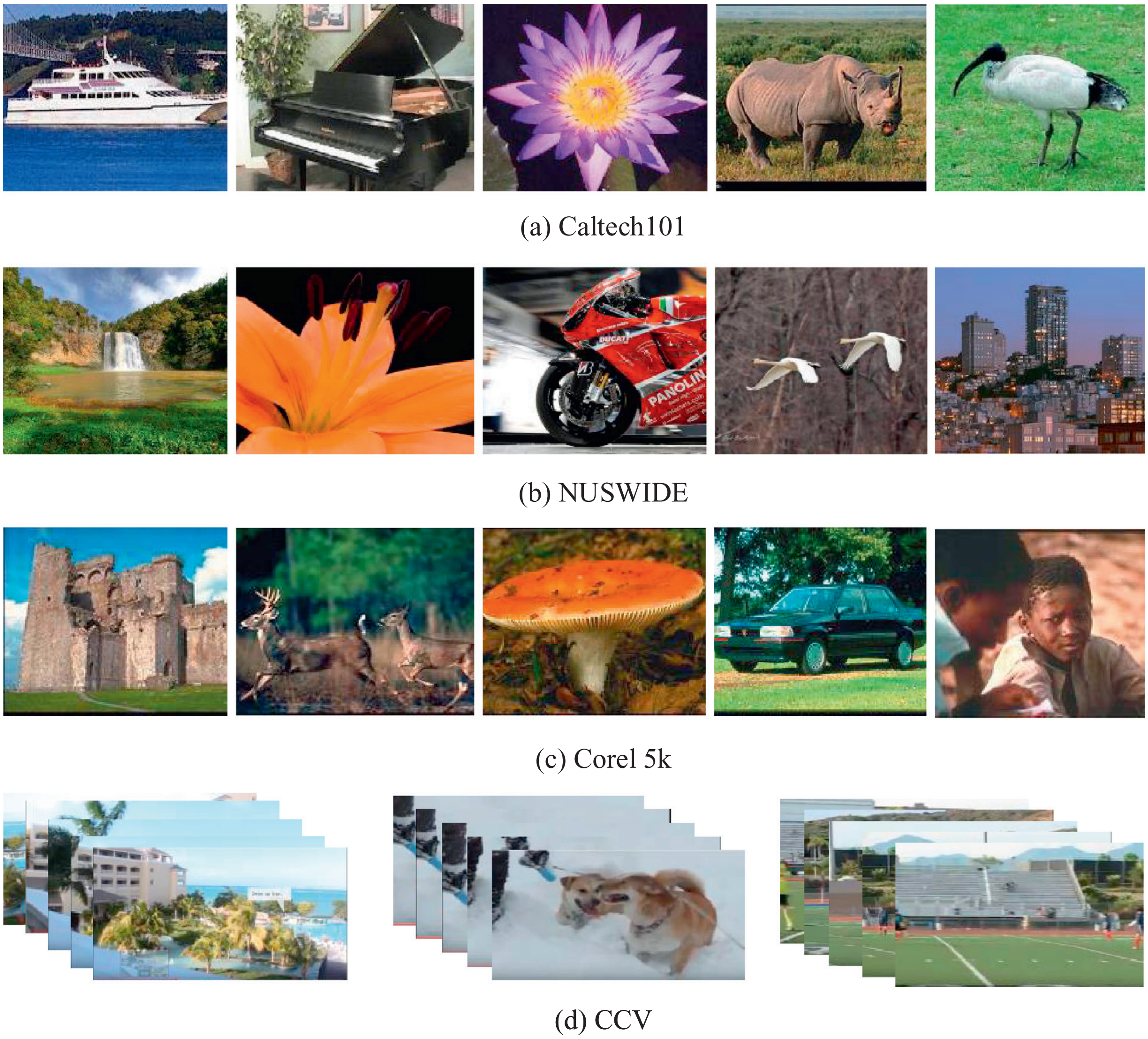
5.2. Experiment competitors and settings

5.2.1. Comparison methods

In the experiment, we compare our model, termed as top-k multi-class SVM using multiple features, with both single-view and multi-view baseline methods.

- $LIBSVM_{SF}$ [2] is a classical implementation of Support Vector Machine which supports multi-class classification. It exploits one-against-one approach to train classifiers on the samples from any two different classes, and then uses a voting scheme to decide the predicted label for each sample.
- $Top - k\ SVM_{SF}$ [20] is an improved method for $LIBSVM_{SF}$ which relaxes the penalty from top-1 guess to top-k guesses in prediction. With this algorithm, the classifier won't be penalized as long as the true label of a sample is within the largest k guesses. It demonstrates state-of-the-art performance in multi-class classification.
- *Feature Concatenation(FC)* is one of the simplest feature fusion methods by concatenating all the features into a single one with equal weight. The fused feature is regarded as the input of $Top - k\ SVM_{SF}$.

Please cite this article as: C. Yan et al., Top-k multi-class SVM using multiple features, Information Sciences (2017), http://dx.doi.org/10.1016/j.ins.2017.08.004



(a) Caltech101

(b) NUSWIDE

(c) Corel 5k

(d) CCV

Fig. 2. Example images from the four datasets. Images in the same row are from the same dataset.

- *MFL* [22] is a kernel-based feature fusion method to integrate multiple types of features obtained from both linear and nonlinear transformations. It does not require any regularization parameter to control the weights of considered features, which makes the fusion process more flexible.
- *SimpleMKL* [37] aims to find a linear combination of multiple kernels by introducing a weighted 2-norm regularization with an additional constraint on the weights, encouraging sparse kernel combinations. The linear combination is learned by solving a standard SVM optimization problem.
- $P-Fusion$ [14] first trains single-view SVM classifier on each feature, and then adopts a probabilistic fusion method to ensemble the results of all the classifiers. It attempts to combine the certainty degree of each single-feature SVM classifier as the weight of the probabilistic output.

5.2.2. Experiment setting

As for the two baseline methods designed for single-view data, classifiers are trained and tested on each view separately. Due to space limit, we only report the highest accuracy achieved by the best single-view classifier. In terms of LIBSVM$_{SF}$, we first adjust the kernel type $-t$ and set it to 2 eventually, representing RBF kernel, which has been verified to show better performance. Regarding to RBF kernel function, only one parameter $-g$ is required to be tuned in the range of $[10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}]$ to achieve the highest accuracy. The other single-view method, *i.e.*, top-k SVM$_{SF}$, mainly has two parameters, *i.e.*, the number of guesses k and regularization parameter λ. We search k in the range from 1 to 15

Please cite this article as: C. Yan et al., Top-k multi-class SVM using multiple features, Information Sciences (2017), http://dx.doi.org/10.1016/j.ins.2017.08.004

Table 2
Top-$\widetilde{k}$ accuracy on Caltech101 dataset.

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-10	Top-15
LIBSVM$_{SF}$	65.07%	71.95%	76.27%	79.10%	80.93%	87.45%	90.70%
top-2 SVM$_{SF}$	67.73%	74.84%	78.62%	81.18%	82.93%	87.99%	90.98%
FC	60.54%	67.86%	72.68%	76.56%	78.51%	86.18%	88.90%
MFL	68.62%	77.74%	81.01%	83.38%	85.93%	90.55%	92.29%
SimpleMKL	65.37%	73.44%	77.41%	81.71%	84.94%	87.29%	90.63%
P–Fusion	66.26%	74.61%	79.54%	82.70%	85.58%	88.83%	91.81%
top-1 SVM$_{MF}$	71.27%	78.54%	82.93%	85.46%	**87.33%**	91.76%	93.51%
top-2 SVM$_{MF}$	**71.92%**	**79.12%**	83.09%	85.50%	**87.33%**	91.76%	93.97%
top-3 SVM$_{MF}$	70.96%	78.93%	**83.32%**	**85.57%**	87.25%	91.99%	94.13%
top-4 SVM$_{MF}$	69.25%	78.58%	83.01%	85.54%	87.17%	91.95%	94.05%
top-5 SVM$_{MF}$	68.93%	78.58%	82.62%	85.34%	87.05%	91.99%	94.13%
top-10 SVM$_{MF}$	65.28%	76.48%	81.18%	83.98%	86.31%	**92.22%**	**94.17%**
top-15 SVM$_{MF}$	62.13%	74.88%	79.82%	83.09%	85.30%	91.41%	93.86%

with incremental step 1 and λ in the range of $[10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}]$ to get results corresponding to different parameter pairs.

In terms of the multi-view baselines, FC method adopts the same setting as in top-k SVM$_{SF}$, since it just acts as a pre-processing step. For each feature described previously in the benchmark datasets, MFL combines the original linear feature h_{linear} and its nonlinear kernel counterpart $K_{nonlinear}$ into a new feature representation, *i.e.*, $[h_{linear}, K_{nonlinear}]$, on the basis of Gaussian RBF kernel. For SimpleMKL, we consider both Gaussian kernels with 10 different bandwidths σ and polynomial kernels of degree 1 to 3, to compute multiple kernel matrices based on multiple features. For P-Fusion, we search the optimal spatial bandwidth in the range of [1, 2, ⋯, 10] and range bandwidth of mean-shift segmentation in [10, 11, ⋯, 20] respectively.

When it comes to our method, three parameters require to be tuned, *i.e.*, k, λ and r. Specifically, k and λ are utilized for controlling the performance of a single-view classifier, which are almost identical with the two parameters in top-k SVM$_{SF}$. The parameter r controls the weight vector distribution among all the views. In practice, k and λ are tuned just as in top-k SVM$_{SF}$ and r is searched in the range from 1.5 to 10.5 with incremental step 1. Besides, the stopping criteria of alternating optimization is determined by checking both the number of iterations and the convergence of primal objective function. In our experiment, we set the maximum number of iterations to 100. Furthermore, both the top-k SVM$_{SF}$ and our method report the accuracies with respect to multiple choices of k. The main emphasis is placed on how the performance changes along with not only the fusion of multiple features but also the number of guesses.

5.2.3. Evaluation metric

In the experiment, top-$\widetilde{k}$ accuracy is selected as our evaluation metric. In the previous studies, top-$\widetilde{k}$ accuracy is usually employed in ranking algorithms for information retrieval, which focuses on the precision of the top-$\widetilde{k}$ retrieval results [35]. With top-$\widetilde{k}$ accuracy, a test sample is correctly classified as long as its ground truth label is within the largest $\widetilde{k}$ predictions. Let n be the number of test samples; $\mathbf{s}_i \in \mathbb{R}^c$ is the prediction score vector and y_i is the ground truth label of the ith sample. $S = [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n] \in \mathbb{R}^{c\times n}$ collects prediction score vectors of all the samples ranked in descending order and $F = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_n] \in \mathbb{R}^{c\times n}$ collects all the categories regarding to each prediction score in S. That is to say, the corresponding category of predicted score S_{ij} is F_{ij}. The top-$\widetilde{k}$ accuracy $t_{\widetilde{k}}$ can be calculated as

$$t_{\widetilde{k}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{\widetilde{k}} \mathbb{1}_{y_i=F_{ji}}. \tag{27}$$

In order to study the classification performance with respect to $\widetilde{k}$, we report the accuracies regarding to various values of $\widetilde{k}$ in {1, 2, 3, 4, 5, 10, 15}.

5.3. Experiment results

In this section, we compare our method with both single-view and multi-view competitors. Experiment results on all the four datasets are given in Tables 2–5 respectively, where the best result in each column is highlighted in bold. In terms of the single-view baseline method, we conduct experiments on each view of a certain dataset and report the highest accuracy only, which is illustrated in the top section of each table. The top-$\widetilde{k}$ accuracies of multi-view baselines are shown in the middle section. The classification results of our method, denoted as top-k SVM$_{MF}$, are demonstrated in the bottom section with various choices of k. From the results, we can give the following observations and analyses:

- Our model achieves the best accuracy among the other multi-view competitors over all the benchmark datasets, demonstrating the superiority and good performance of the proposed method. Comparing with the multi-view baseline with

Table 3
Top-$\tilde{k}$ accuracy on NUSWIDE dataset.

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-10	Top-15
LIBSVM$_{SF}$	27.65%	37.90%	46.59%	53.55%	58.80%	78.23%	88.68%
top-1 SVM$_{SF}$	30.52%	44.40%	53.31%	60.15%	65.71%	80.85%	87.62%
FC	33.74%	50.27%	60.49%	67.25%	73.09%	86.85%	92.96%
MFL	39.16%	55.33%	64.56%	70.74%	75.42%	87.97%	92.83%
SimpleMKL	35.75%	52.53%	62.33%	68.90%	74.09%	87.56%	93.11%
P–Fusion	38.47%	54.37%	63.91%	70.34%	74.93%	88.06%	92.99%
top-1 SVM$_{MF}$	40.11%	55.53%	64.93%	71.02%	**76.11%**	89.15%	94.32%
top-2 SVM$_{MF}$	**40.14%**	55.65%	64.99%	71.04%	76.08%	89.21%	94.30%
top-3 SVM$_{MF}$	40.09%	55.60%	**65.01%**	71.04%	76.01%	**89.30%**	94.22%
top-4 SVM$_{MF}$	39.93%	**55.83%**	64.99%	71.12%	76.02%	89.26%	94.27%
top-5 SVM$_{MF}$	39.96%	**55.83%**	64.94%	71.07%	76.06%	89.28%	94.28%
top-10 SVM$_{MF}$	39.45%	55.30%	64.86%	71.01%	75.93%	89.21%	94.32%
top-15 SVM$_{MF}$	38.98%	55.04%	64.73%	**71.17%**	76.03%	89.08%	**94.52%**

Table 4
Top-$\tilde{k}$ accuracy on Corel 5k dataset.

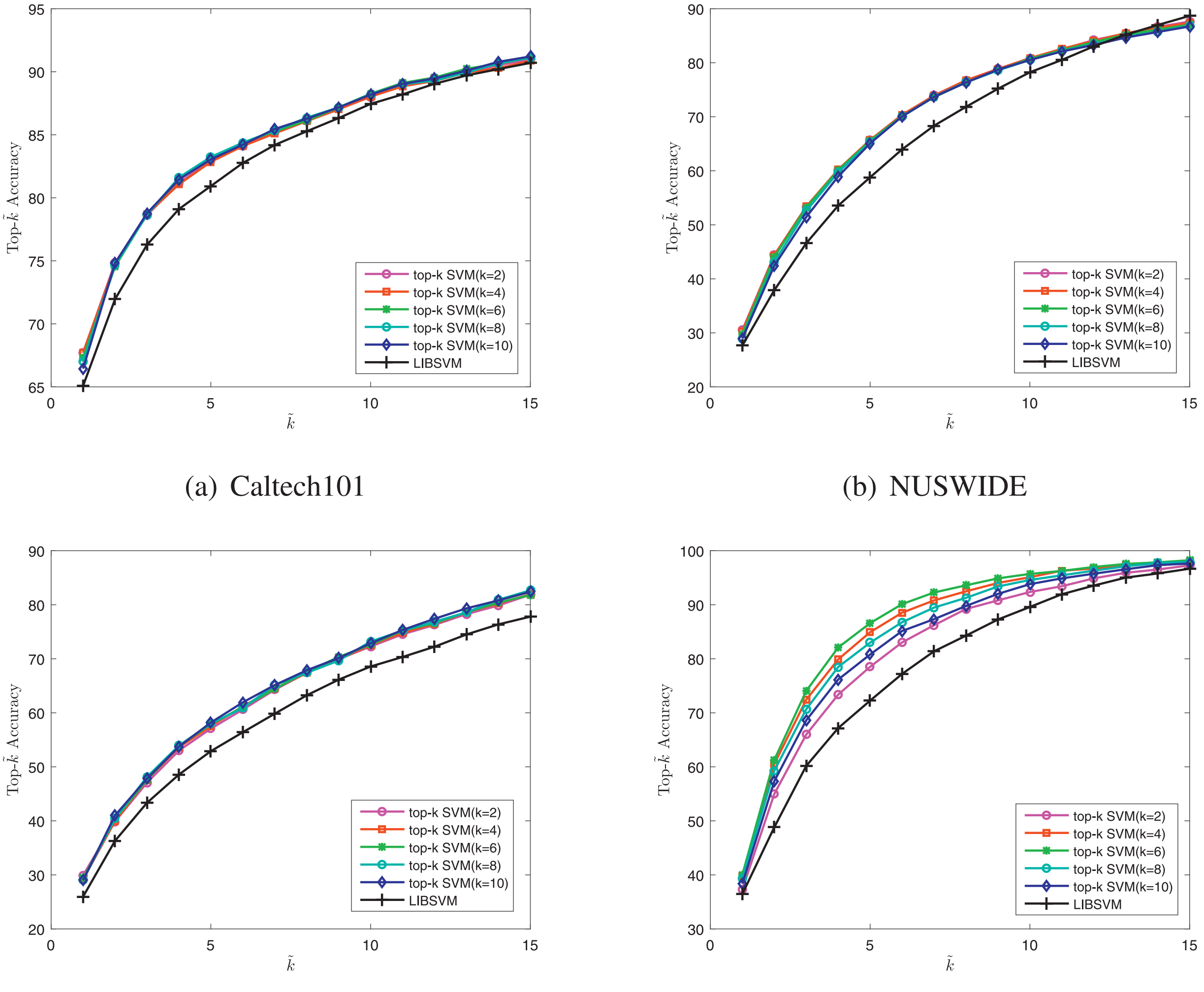
Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-10	Top-15
LIBSVM$_{SF}$	25.97%	36.30%	43.33%	48.57%	52.90%	68.57%	77.80%
top-2 SVM$_{SF}$	29.93%	39.93%	47.40%	53.53%	57.53%	72.53%	80.93%
FC	26.40%	38.00%	47.13%	53.93%	59.60%	73.40%	81.07%
MFL	31.53%	42.46%	51.00%	56.47%	60.87%	73.83%	83.33%
SimpleMKL	27.20%	40.93%	49.20%	54.80%	58.33%	73.73%	82.53%
P–Fusion	29.60%	41.46%	50.00%	55.73%	59.00%	74.53%	83.67%
top-1 SVM$_{MF}$	33.80%	45.26%	51.80%	56.33%	61.00%	74.27%	**84.47%**
top-2 SVM$_{MF}$	**34.13%**	**45.47%**	52.13%	56.93%	61.53%	74.20%	83.13%
top-3 SVM$_{MF}$	33.40%	44.80%	52.00%	57.07%	61.60%	74.47%	83.13%
top-4 SVM$_{MF}$	33.33%	44.87%	52.00%	56.53%	61.40%	74.87%	83.26%
top-5 SVM$_{MF}$	33.27%	44.93%	**52.60%**	**57.20%**	**61.67%**	**75.33%**	83.40%
top-10 SVM$_{MF}$	32.47%	43.20%	51.60%	56.67%	61.13%	73.13%	82.80%
top-15 SVM$_{MF}$	29.13%	41.27%	49.73%	55.33%	59.20%	73.80%	83.40%

Table 5
Top-$\tilde{k}$ accuracy on CCV dataset.

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-10	Top-15
LIBSVM$_{SF}$	36.46%	48.82%	60.17%	67.17%	72.26%	89.59%	96.68%
top-5 SVM$_{SF}$	40.03%	61.30%	74.06%	82.09%	86.60%	95.70%	98.24%
FC	40.38%	60.14%	71.56%	78.57%	83.25%	95.58%	98.11%
MFL	48.75%	68.98%	78.12%	85.51%	89.22%	96.09%	98.63%
SimpleMKL	45.79%	65.84%	75.68%	83.97%	86.91%	95.58%	98.67%
P–Fusion	47.42%	66.13%	76.01%	84.22%	87.59%	95.88%	98.58%
top-1 SVM$_{MF}$	50.30%	68.47%	79.68%	85.57%	89.48%	97.47%	99.31%
top-2 SVM$_{MF}$	50.30%	70.06%	80.24%	86.30%	90.25%	97.55%	99.18%
top-3 SVM$_{MF}$	**51.46%**	**71.65%**	81.27%	86.68%	90.55%	97.42%	**99.48%**
top-4 SVM$_{MF}$	50.86%	71.48%	81.44%	87.46%	91.07%	97.51%	**99.48%**
top-5 SVM$_{MF}$	50.30%	71.35%	**81.62%**	**87.84%**	**91.28%**	97.42%	**99.48%**
top-10 SVM$_{MF}$	40.59%	62.54%	76.68%	84.88%	89.99%	**97.94%**	99.31%
top-15 SVM$_{MF}$	33.89%	54.04%	69.29%	79.38%	85.61%	97.42%	99.44%

the best performance, the top-1 accuracy improves 3.30% on Caltech101, 0.98% on NUSWIDE, 2.60% on Corel 5k and 2.71% on CCV respectively. In particular, the kernel based feature fusion method MFL achieves better performance than the other multi-view baselines over all the datasets, which is better suited to our problem.

- From the results of multi-view methods, we can observe that most of them outperform the two single-view methods, *i.e.*, LIBSVM$_{SF}$ and top-k SVM$_{SF}$, which confirms the effectiveness of multiple feature fusion. However, there are also some notable exceptions. For example, the FC method gets an accuracy even much lower than the single-view methods on several datasets, inferring that some single features are more discriminative for classification than the simple concatenation of multiple features. As a result, only with proper fusion method can we make full use of multiple features to boost the classification accuracy.
- The proposed multi-view classification algorithm, *i.e.*, top-k SVM$_{MF}$, significantly beats the top-k SVM$_{SF}$ on all the benchmark datasets. Specifically, the improvements of top-1 accuracy are 4.19% on Caltech101, 9.62% on NUSWIDE, 4.20% on



(a) Caltech101

(b) NUSWIDE

(c) Corel 5k

(d) CCV

Fig. 3. The curves of top-$\tilde{k}$ accuracy with respect to different values of $\tilde{k}$ over four datasets.

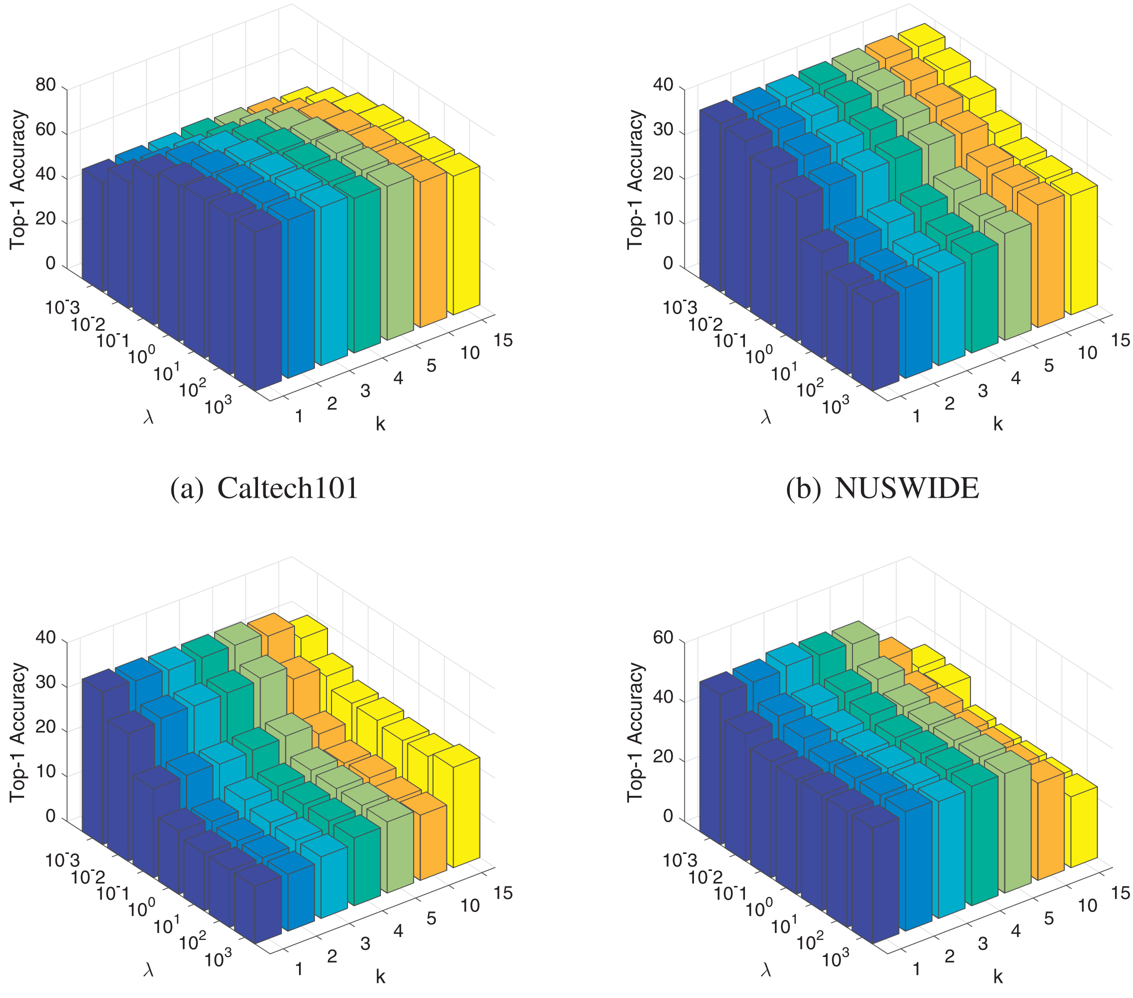
Corel 5k and 11.43% on CCV, respectively. Combining multiple features is an effective method to enhance the performance of top-k multi-class classification.

- We also notice that the selection of k has a direct impact on the performance of our model. In each column, the classification accuracy increases gradually along with the increase of k. When reaching the maximum, the accuracy begins to decrease even if the value of k is still increasing. Note that the experiment results on all the datasets follow this rule, achieving the best performance at $k \geq 2$.
- Considering the two single-view methods, top-k SVM$_{SF}$ outperforms LibSVM$_{SF}$ over all the four datasets by 2.66% on Caltech101, 2.87% on NUSWIDE, 3.96% on Corel 5k and 3.57% on CCV respectively. By relaxing the penalty, top-k SVM can enhance the performance of conventional top-1 SVM in multi-class SVM classification. For each dataset, both the two single-view methods achieve the highest accuracy on the same feature of this dataset. In specific, the best features are the 1984-D HOG on Caltech101, 145-D color correlation(CORR) on NUSWIDE, 64-D scalable color on Corel 5k and 5000-D STIP on CCV respectively. To present the comparison results of two single-view methods more clearly, we plot the variation curves of top-$\tilde{k}$ accuracy for each method over the four benchmark datasets in Fig. 3.

5.4. Sensitivity analysis

To demonstrate the robustness of our method, we study the influence of parameters k, λ and r on the performance of visual classification. Since it is not clear to demonstrate all the three parameters in one figure, we separate them into two parts according to their roles in the training process.

Please cite this article as: C. Yan et al., Top-k multi-class SVM using multiple features, Information Sciences (2017), http://dx.doi.org/10.1016/j.ins.2017.08.004



(a) Caltech101

(b) NUSWIDE

(c) Corel 5k

(d) CCV

Fig. 4. The sensitivity of top-1 accuracy with respect to varying values of k and λ on four datasets.

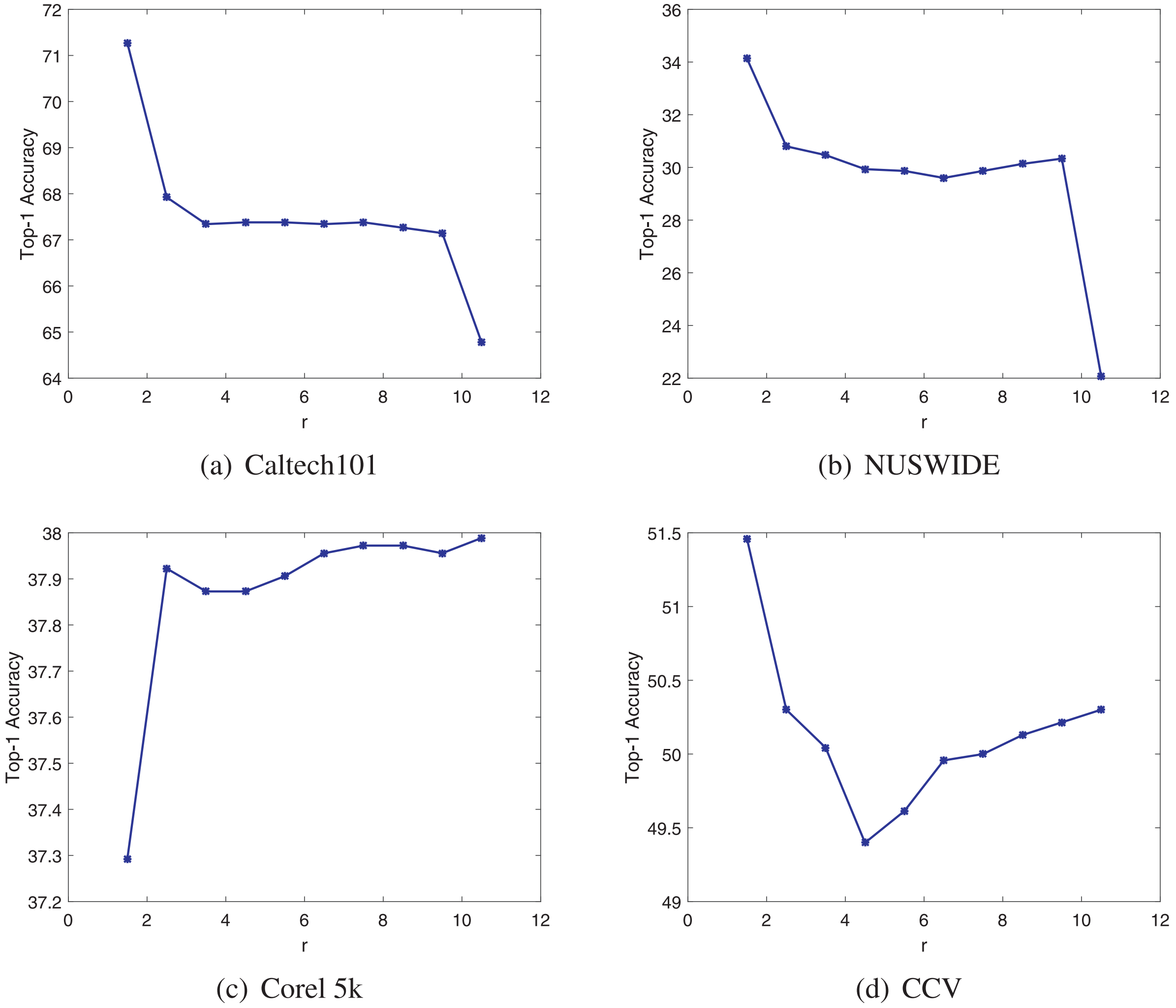
In Fig. 4, we draw the 3D-bar graphs of top-1 accuracy with respect to k and λ in a 3-dimension diagram. In terms of a certain pair of k and λ, we traverse all the optional values of r to obtain the highest top-1 accuracy. As shown in Fig. 4, the top-1 accuracy of Caltech101 is the least sensitive while that of Corel 5k is the most sensitive to the variation of k and λ among all the datasets.

The curves of top-1 accuracy with respect to r are illustrated in Fig. 5. In a similar fashion, the accuracy corresponding to a certain r is obtained by traversing all the possible values of k and λ. The classification accuracy of CCV is more sensitive to the variation of r, whereas the other three datasets have stable performance in a wide range. For a given dataset, the optimal value of r can approximately reflect the correlation relationship among different views. Our model performs the best at around $r = 1.5$ over Caltech101, NUSWIDE and CCV. As for Corel 5k, the highest accuracy is achieved at around $r = 10.5$. We can roughly infer that Corel 5k dataset has richer complementary information among different views than Caltech101, NUSWIDE and CCV. Besides, the variation trend of the accuracy curve corresponding to Caltech101 is in agreement with that of NUSWIDE, having a sharp decline at $r = 9.5$.

In summary, our method shows robustness to different values of the parameters k, λ and r in a wide range. However, there seems no analogous rule to identify the best parameters for all of the applications because of the varying properties related to different datasets.

6. Conclusion

In this paper, we propose a novel method for visual object classification, which integrates multiple feature fusion into the framework of top-k multi-class SVM. Different from the traditional methods, our method takes not only the top-k loss but



(a) Caltech101

(b) NUSWIDE

(c) Corel 5k

(d) CCV

Fig. 5. Top-1 accuracy curves with respect to different values of *r* over four datasets.

also the reasonable combination of multiple features into account. In the proposed model, the largest *k* predictions rather than only the largest one are considered to relax the penalty in multi-class classification. This strategy alleviates the class ambiguity problem to a large extent. By learning adaptive weights for all the single-view classifiers, the weighted sum of their prediction results is regarded as the final decision. Extensive experiments on four benchmark datasets demonstrate that our method performs much better than both the single-view and multi-view baselines in multi-class visual classification.

Acknowledgments

This work was funded by the National Science Foundation of China (Nos. 61502377, 61532004), the National Key Research and Development Program of China (No. 2016YFB1000903); Ministry of Education Innovation Research Team (No. IRT_17R86); Project of China Knowledge Center for Engineering Science and Technology, and China Postdoctoral Science Foundation (No. 2015M582662).

References

[1] J. Borwein, A.S. Lewis, Convex Analysis and Nonlinear Optimization: Theory and Examples, Springer Science & Business Media, 2010.
[2] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 27.
[3] X. Chang, Z. Ma, Y. Yang, Z. Zeng, A.G. Hauptmann, Bi-level semantic representation analysis for multimedia event detection, IEEE Trans. Cybern. 47 (5) (2017a) 1180–1197.
[4] X. Chang, Y. Yu, Y. Yang, E.P. Xing, Semantic pooling for complex event analysis in untrimmed videos, IEEE Trans. Pattern Anal. Mach. Intell. 39 (8) (2017b) 1617–1632.
[5] X. Chang, Y.-L. Yu, Y. Yang, Robust top-k multi-class svm for visual category recognition, ACM KDD, 2017c.
[6] S. Chaudhuri, A. Tewari, Online learning to rank with top-k feedback, arXiv:1608.06408, (2016).

Please cite this article as: C. Yan et al., Top-*k* multi-class SVM using multiple features, Information Sciences (2017), http://dx.doi.org/10.1016/j.ins.2017.08.004

[7] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, Nus-wide: a real-world web image database from national university of singapore, CIVR, 2009.
[8] K. Crammer, Y. Singer, On the algorithmic implementation of multiclass kernel-based vector machines, J. Mach. Learn. Res. 2 (Dec) (2001) 265–292.
[9] J. Farquhar, D. Hardoon, H. Meng, J.S. Shawe-taylor, S. Szedmak, Two view learning: Svm-2k, theory and practice, NIPS, 2005.
[10] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories, Comput. Vision Image Underst. 106 (1) (2007) 59–70.
[11] P. Gehler, S. Nowozin, On feature combination for multiclass object classification, ICCV, 2009.
[12] M.R. Gupta, S. Bengio, J. Weston, Training highly multiclass classifiers., J. Mach. Learn. Res. 15 (1) (2014) 1461–1492.
[13] C. He, J. Shao, X. Xu, D. Ouyang, L. Gao, Exploiting score distribution for heterogenous feature fusion in image classification, Neurocomputing 253 (30) (2017) 70–76.
[14] X. Huang, L. Zhang, An svm ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery, IEEE Trans. Geosci. Remote Sens. 51 (1) (2013) 257–272.
[15] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, A.C. Loui, Consumer video understanding: a benchmark database and an evaluation of human and machine performance, ICMR, 2011.
[16] T.-K. Kim, J. Kittler, R. Cipolla, Discriminative learning and recognition of image set classes using canonical correlations, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 1005–1018.
[17] K.C. Kiwiel, Variable fixing algorithms for the continuous quadratic knapsack problem, J. Optim. Theory Appl. 136 (3) (2008) 445–458.
[18] D. Lahat, T. Adali, C. Jutten, Multimodal data fusion: an overview of methods, challenges, and prospects, Proc. IEEE 103 (9) (2015) 1449–1477.
[19] G.R. Lanckriet, N. Cristianini, P. Bartlett, L.E. Ghaoui, M.I. Jordan, Learning the kernel matrix with semidefinite programming, J. Mach. Learn. Res. 5 (January) (2004) 27–72.
[20] M. Lapin, M. Hein, B. Schiele, Top-k multiclass svm, NIPS, 2015.
[21] M. Lapin, M. Hein, B. Schiele, Loss functions for top-k error: analysis and insights, CVPR, 2016.
[22] J. Li, X. Huang, P. Gamba, J.M. Bioucas-Dias, L. Zhang, J.A. Benediktsson, A. Plaza, Multiple feature learning for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 53 (3) (2015) 1592–1606.
[23] S. Li, T. Lu, L. Fang, X. Jia, J.A. Benediktsson, Probabilistic fusion of pixel-level and superpixel-level hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 54 (12) (2016) 7416–7430.
[24] G. Lin, C. Fan, H. Zhu, Y. Miu, X. Kang, Visual feature coding based on heterogeneous structure fusion for image classification, Inf. Fusion (2017), doi:10.1016/j.inffus.2016.12.010.
[25] H. Liu, D. Guo, F. Sun, Object recognition using tactile measurements: Kernel sparse coding methods, IEEE Trans. Instrum. Meas. 65 (3) (2016a) 656–665.
[26] H. Liu, J. Qin, F. Sun, D. Guo, Extreme kernel sparse learning for tactile object recognition, IEEE Trans. Cybern. (2016b), doi:10.1109/TCYB.2016.2614809.
[27] H. Liu, F. Sun, B. Fang, X. Zhang, Robotic room-level localization using multiple sets of sonar measurements, IEEE Trans. Instrum. Meas. 66 (1) (2017a) 2–13.
[28] H. Liu, Y. Wu, F. Sun, B. Fang, D. Guo, Weakly paired multimodal fusion for object recognition, IEEE Trans. Autom. Sci. Eng. (2017b), doi:10.1109/TASE.2017.2692271.
[29] H. Liu, Y. Yu, F. Sun, J. Gu, Visual–tactile fusion for object recognition, IEEE Trans. Autom. Sci. Eng. 14 (2) (2017c) 996–1008.
[30] P. Liu, J.-M. Guo, K. Chamnongthai, H. Prasetyo, Fusion of color histogram and lbp-based features for texture image retrieval and classification, Inf. Sci. (2017d), doi:10.1016/j.ins.2017.01.025.
[31] M. Luo, X. Chang, L. Nie, Y. Yi, A.G. Hauptmann, Q. Zheng, An adaptive semisupervised feature analysis for video semantic recognition, IEEE Trans. Cybern. (2017a), doi:10.1109/TCYB.2017.2647904.
[32] M. Luo, F. Nie, X. Chang, Y. Yang, A. Hauptmann, Q. Zheng, Avoiding optimal mean robust pca/2dpca with non-greedy l1-norm maximization, IJCAI, 2016.
[33] M. Luo, F. Nie, X. Chang, Y. Yang, A.G. Hauptmann, Q. Zheng, Adaptive unsupervised feature selection with structure regularization, IEEE Trans. Neural Netw. Learn. Syst. (2017b), doi:10.1109/TNNLS.2017.2650978.
[34] Z. Ma, Y. Yang, N. Sebe, A.G. Hauptmann, Multiple features but few labels? A symbiotic solution exemplified for video analysis, MM, 2014.
[35] H. Nguyen, J. Cao, Trustworthy answers for top-k queries on uncertain big data in decision making, Inf. Sci. 318 (10) (2015) 73–90.
[36] M. Ozay, F.T.Y. Vural, A new fuzzy stacked generalization technique and analysis of its performance, arXiv:1204.0171, (2012).
[37] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet, Simplemkl, J. Mach. Learn. Res. 9 (November) (2008) 2491–2521.
[38] S. Shalev-Shwartz, T. Zhang, Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization, Math. Program. 155 (1) (2016) 105–145.
[39] K. Swersky, D. Tarlow, R.P. Adams, R. Zemel, B.J. Frey, Probabilistic n-choose-k models for classification and ranking, NIPS, 2012.
[40] N. Usunier, D. Buffoni, P. Gallinari, Ranking with ordered weighted pairwise classification, ICML, 2009.
[41] A. Vinokourov, N. Cristianini, J.S. Shawe-Taylor, Inferring a semantic representation of text via cross-language correlation analysis, NIPS, 2002.
[42] J. Wei, H. Liu, G. Yan, F. Sun, Robotic grasping recognition using multi-modal deep extreme learning machine, Multidimens. Syst. Signal Process. 28 (3) (2016) 817–833.
[43] J. Weston, S. Bengio, N. Usunier, Wsabie: Scaling up to large vocabulary image annotation, IJCAI, 2011.
[44] C. Xu, D. Tao, C. Xu, A survey on multi-view learning, arXiv:1304.5634, (2013).
[45] Z. Xu, I. Tsang, Y. Yang, Z. Ma, A. Hauptmann, Event detection using multi-level relevance labels and multiple features, CVPR, 2014.
[46] F. Yang, M. Ding, X. Zhang, W. Hou, C. Zhong, Non-rigid multi-modal medical image registration by combining l-bfgs-b with cat swarm optimization, Inf. Sci. 316 (C) (2015) 440–456.
[47] Y. Yang, Y. Que, S. Huang, P. Lin, Multimodal sensor medical image fusion based on type-2 fuzzy logic in nsct domain, IEEE Sensors J. 16 (10) (2016) 3735–3745.
[48] Y. Yu, O. Aslan, D. Schuurmans, A polynomial-time form of robust regression, NIPS, 2012.
[49] Z. Zhong, B. Fan, K. Ding, H. Li, S. Xiang, C. Pan, Efficient multiple feature fusion with hashing for hyperspectral imagery classification: a comparative study, IEEE Trans. Geosci. Remote Sens. 54 (8) (2016) 4461–4478.
[50] A. Zien, C.S. Ong, Multiclass multiple kernel learning, ICML, 2007.

Please cite this article as: C. Yan et al., Top-*k* multi-class SVM using multiple features, Information Sciences (2017), http://dx.doi.org/10.1016/j.ins.2017.08.004