Clustering-Based Joint Feature Selection for Semantic Attribute Prediction*

Lin Chen and Baoxin Li
Arizona State University, Tempe Arizona
{lin.chen.cs, baoxin.li}@asu.edu

Abstract

Semantic attributes have been proposed to bridge the semantic gap between low-level feature representation and high-level semantic understanding of visual objects. Obtaining a good representation of semantic attributes usually requires learning from high-dimensional low-level features, which not only significantly increases the time and space requirement but also degrades the performance due to numerous irrelevant features. Since multi-attribute prediction can be generalized as an multi-task learning problem, sparse-based multi-task feature selection approaches have been introduced, utilizing the relatedness among multiple attributes. However, such approaches either do not investigate the pattern of the relatedness among attributes, or require prior knowledge about the pattern. In this paper, we propose a novel feature selection approach which embeds attribute correlation modeling in multi-attribute joint feature selection. Experiments on both synthetic dataset and multiple public benchmark datasets demonstrate that the proposed approach effectively captures the correlation among multiple attributes and significantly outperforms the state-of-the-art approaches.

1 Introduction

Recent literature has witnessed fast development of representations using semantic attributes, whose goal is to bridge the semantic gap between low-level feature representation and high-level semantic understanding of visual objects. Attributes refer to visual properties that help describe visual objects or scenes such as "natural" scenes, "fluffy" dogs, or "formal" shoes. Visual attributes exist across object category boundaries and many methods have been employed in applications including object recognition [Farhadi *et al.*, 2010], face verification [Song *et al.*, 2012], image search [Kovashka *et al.*, 2012; Scheirer *et al.*, 2012] and sentiment analysis [Wang *et al.*, 2015].

*The work was supported in part by ONR grant N00014-15-1-2344 and ARO grant W911NF1410371. Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of ONR or ARO.

Figure 1: Illustration of Shoe images with three corresponding attributes "High Heel", "Formal" and "Red".

Good representations of semantic attributes are often built on top of high-dimensional, low-level features. Attribute learning directly based on such raw, high-dimensional features may suffer from the curse of dimensionality curse. Further, often it is reasonable to assume that not all the low-level features would have equal contribution to all the attributes. Feature selection, selecting a subset of most relevant features for a compact and accurate presentation, is proven to be an effective and efficient way to handle high-dimensional data [Tang *et al.*, 2014].

Multi-task joint feature selection has been introduced by [Chen *et al.*, 2014] for attribute ranking by exploring the correlation among attributes. However, this work assumes that all attributes are correlated by sharing the same subset of features, which is not always accurate. For example, as shown in Figure 1, a "high-heel" shoe is usually considered as a "formal" shoe as well. It is reasonable to assume these attributes share the same subset of features, e.g., shape-related descriptors. However, it is hard to identify whether "high heel" or "formal" shoes are in red, which suggests the attribute "color" may not share the same subset of features with the other attributes but is determined by, e.g., color-related descriptors. In other words, attributes are usually related in clustering structures. [Jayaraman *et al.*, 2014] first explores such clustered relatedness on attribute prediction. However, their approach requires manually specified group structure as prior. To our knowledge, there is still lack of a feature selection approach being able to identify grouping/clusering structures among attributes for improved attribute prediction.

In this paper, we propose a regularization-based multi-task feature selection approach that aims at automatically partitioning the attributes into groups while simultaneously utilizing such group information for attribute-dependent feature selection. We employ a clustering regularizer for attribute partition, where strong attribute relatedness is assumed to exist within each cluster. Besides, a group-sparsity regularizer is imposed on the objective function to encourage intra-cluster feature sharing and inter-cluster feature competition. Under this formulation, we propose an alternating structure optimization algorithm, which efficiently solves the relaxed form of the proposed formulation. We verify the effectiveness and generalization capability of our approach on both synthetic and real-world benchmark datasets. The results show that our approach outperforms the state-of-the-art approaches on feature selection, attribute prediction and zero-shot learning.

2 Methodology

Let $\mathcal{F} = \{f_1, f_2, \ldots, f_d\}$ be the set of d features and then we can represent a set of n instances by the feature set $\mathcal{F}$ as $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$. Let $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$ be the set of m attribute labels and $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n] \in \{0,1\}^{m\times n}$ denotes the label matrix where $\boldsymbol{y}_i \in \mathbb{R}^m (i = 1, 2, \ldots, n)$ is the label vector of the i-th instance. We aim to select $K (K \leq d)$ most relevant features from $\mathcal{F}$ by leveraging X, Y and the attribute correlation in $\mathcal{C}$. Let $\mathbf{s} = \pi(\overbrace{0, \ldots, 0}^{d-K}, \overbrace{1, \ldots, 1}^{K})$, where $\pi(\cdot)$ is the permutation function and K is the number of features to select where $\mathbf{s}_i = 1$ indicates that the i-th feature is selected. The original data can be represented as $\text{diag}(\mathbf{s})\mathbf{X}$ with K selected features where $\text{diag}(\mathbf{s})$ is a diagonal matrix. We assume that a linear projection matrix $W = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m] \in \mathbb{R}^{d\times m}$ maps the data X to its label matrix Y where $\boldsymbol{w}_i \in \mathbb{R}^d$ is the projection vector for the i-th class c_i. If we do not consider attribute correlation, we can select K features via solving the following optimization problem:

$$\begin{aligned} \min_{W,\mathbf{s}} \quad & L(W^\top \text{diag}(\mathbf{s}) X, Y) \\ s.t., \quad & \mathbf{s} \in \{0,1\}^n,\ \mathbf{s}^T \mathbf{1}_n = K \end{aligned}$$

where $L(\cdot)$ is the loss function and typical choices of loss functions include least square and logistic regression.

2.1 Modeling Label Correlation

Based on the assumption that correlated attributes would share the same features, we propose to model attribute correlation via learning the clustering structures through k-means. Let E be a permutation partition matrix, then a partition of the projection matrix W into k clusters can be formed as:

$$WE = [W_1, W_2, \ldots, W_k], W_i = [\boldsymbol{w}_1^{(i)}, \boldsymbol{w}_2^{(i)}, \ldots, \boldsymbol{w}_{n_i}^{(i)}];$$

where $W_i \in \mathbb{R}^{d\times n_i} (i = 1, 2, \ldots, k)$ is the i-th partitioned group includes n_i projection vectors (or attribute labels). The associated sum-of-squares cost function for the partition can be formulated as

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \|\boldsymbol{w}_j^{(i)} - \boldsymbol{m}_i\|^2, \boldsymbol{m}_i = \sum_{j=1}^{n_i} \boldsymbol{w}_j^{(i)} / n_i \tag{1}$$

where $\boldsymbol{m}_i$ denotes the mean vector of the i-th cluster. Let $e_i = [1, 1, \ldots, 1]^\top \in \mathbb{R}^{n_i \times 1}$, then Eq. (1) can be derived as

$$\begin{aligned} \sum_{i=1}^{k}\sum_{j=1}^{n_i} \|\boldsymbol{w}_j^{(i)} - \boldsymbol{m}_i\|^2 &= \sum_{i=1}^{k} \|W_i (I_{n_i} - \frac{e_i {e_i}^\top}{n_i})\|_F^2 \\ &= \sum_{i=1}^{k} \mathbf{Tr}(W_i^\top W_i) - (\frac{{e_i}^\top}{\sqrt{n_i}}) W_i^\top W_i (\frac{e_i}{\sqrt{n_i}}) \end{aligned} \tag{2}$$

Let $F = \text{diag}(\frac{e_1}{\sqrt{n_1}}, \frac{e_2}{\sqrt{n_2}}, \ldots, \frac{e_k}{\sqrt{n_k}}) \in \mathbb{R}^{m\times k}$ be an orthonormal matrix, then Eq. (2) can be rewritten as

$$\mathbf{Tr}(W^\top W) - \mathbf{Tr}(F^\top W^\top W F)$$

To make the problem tractable, we ignore the special structure of F and let it be an arbitrary orthonormal matrix. By adding a global penalty $\mathbf{Tr}(W^\top W)$ measuring how large the weight vectors are, capturing label correlation is to partition W into k clusters, which can be achieved by solving the following optimization problem:

$$\min_{F^\top F = I_k} \mathbf{Tr}(W^\top W) - \mathbf{Tr}(F^\top W^\top W F) + \gamma \mathbf{Tr}(W^\top W) \tag{3}$$

2.2 Feature Selection

With the model component to capture attribute correlation in Eq. (3), the proposed feature selection framework is to solve the following optimization problem:

$$\begin{aligned} \min_{W,F,\mathbf{s}} \ & L(W^\top \text{diag}(\mathbf{s}) X, Y) + \gamma \mathbf{Tr}(W^\top W) \\ & + \beta(\mathbf{Tr}(W^\top W) - \mathbf{Tr}(F^\top W^\top W F)) \\ s.t. \quad & F^\top F = I_k,\ \mathbf{s} \in \{0,1\}^n,\ \mathbf{s}^\top \mathbf{1}_n = K \end{aligned} \tag{4}$$

where β controls the contribution from modeling label correlation and γ controls the generalization performance.

The constraint on $\mathbf{s}$ makes Eq. (4) a mixed integer programming problem, which is difficult to solve. We observe that $\text{diag}(\mathbf{s})$ and W are in the form of $W^T \text{diag}(\mathbf{s})$. Since $\mathbf{s}$ is a binary vector and $d - K$ rows of the $\text{diag}(\mathbf{s})$ are all zeros, $W^T \text{diag}(\mathbf{s})$ is a matrix where the elements of many rows are all zeros. This motivates us to absorb $\text{diag}(\mathbf{s})$ into W as $W = W^T \text{diag}(\mathbf{s})$, and add $\ell_{2,1}$-norm on each grouped W_i to encourage sparse-based group-wise joint feature selection. With this relaxation, Eq. (4) can be rewritten as:

$$\begin{aligned} \min_{W,F;F^\top F = I_k} \ & L(W^\top X, Y) + \alpha \sum_{i=1}^{k} \|W_i\|_{2,1} + \gamma \mathbf{Tr}(W^\top W) \\ & + \beta(\mathbf{Tr}(W^\top W) - \mathbf{Tr}(F^\top W^\top W F)) \end{aligned} \tag{5}$$

where α controls the sparsity of W. The key idea lying here is that we use the clustering regularizer to partition the tasks into groups where strong correlation exists among tasks in the same group; and feature selection based on such group structures would make sure appropriate feature subsets are selected to represent the respective semantic attributes.

3 Algorithm

In this section, we first introduce an optimization algorithm to seek an optimal solution (summarized in Algorithm 1) for Eq. (5). Then we propose an approach to estimate the attribute assignment (summarized in Algorithm 2).

3.1 Optimization

The optimization problem in Eq. (5) is non-convex non-smooth, which makes the formulation difficult to solve in its original form. Thus we adopt several relaxations to make it solvable.

The attribute correlation regularization in Eq. (3) can be rewritten as:

$$\beta\mathbf{Tr}(W((1+\eta)I - FF^\top)W^\top)$$

where $\eta = \gamma/\beta > 0$. Let $M = FF^\top$, according to [Zhou *et al.*, 2011] the previous regularizer can be relaxed into the following convex form:

$$\begin{aligned} &\beta\eta(1+\eta)\mathbf{Tr}(W(\eta I + M)^{-1}W^\top) \\ &s.t.\ \ tr(M) = k,\ M \preceq I,\ M \in \mathbb{S}_+^m \end{aligned} \tag{6}$$

where $\mathbb{S}_+^m$ is the set of $m \times m$ positive semidefinite matrices.

Following a similar idea in [Bach, 2008], we reformulate Eq. (5) by squaring the $\ell_{2,1}$ norm. Since the $\ell_{2,1}$ norm is positive, the squaring represents a smooth monotonic mapping. Without loss of the generality, we adopt the traditional least square loss for demonstration in this paper. Then we get the following jointly convex smooth objective function regarding to W and M.

$$\begin{aligned} \arg\min_{W,M} &\|W^\top X - Y\|_F^2 + \alpha\sum_{i=1}^{k}(\|W_i\|_{2,1})^2 \\ &- \beta\eta(1+\eta)\mathbf{Tr}(W(\eta I + M)^{-1}W^\top) \\ &s.t.\ \ tr(M) = k,\ M \preceq I,\ M \in \mathbb{S}_+^m \end{aligned} \tag{7}$$

Since it is difficult to optimize the linear projection matrix W and attribute correlation matrix M simultaneously, we employ Alternating Structure Optimization (ASO), which has been shown to be effective in many practical applications [Blitzer *et al.*, 2006; Quattoni *et al.*, 2007] and is guaranteed to converge to a global optimal solution.

Optimizing M when fixing W

Given a fixed W, the optimization problem is decoupled into the following optimization problem:

$$\begin{aligned} &\min_{M}\ \mathbf{Tr}(W(\eta I + M)^{-1}W^\top) \\ &s.t.\ \ tr(M) = k,\ M \preceq I,\ M \in \mathbb{S}_+^m \end{aligned} \tag{8}$$

We solve the problem based on the following Lemma due to [Zhou *et al.*, 2011]:

Lemma 1 *For the optimization problem in Eq. (8), let $W = U\Sigma V$ be the singular value decomposition of W where $\Sigma = diag([\sigma_1, \sigma_2, \ldots, \sigma_m])$, $M = Q\Lambda Q^\top$ be the Eigen decomposition of M where $\Lambda = diag([\lambda_1, \lambda_2, \ldots, \lambda_q])$ and q be the rank of Σ. Then the optimal Q^* is given by $Q^* = V$ and the optimal Λ^* is given by solving the following optimization problem:*

$$\begin{aligned} \Lambda^* &= \arg\min_{\Lambda}\sum_{i=1}^{q}\frac{\sigma_i^2}{\eta + \lambda_i} \\ s.t.\ &\sum_{i=1}^{q}\lambda_i = k, 0 \le \lambda_i \le 1 \end{aligned} \tag{9}$$

Eq. (9) can be solved using the similar technology in [Jacob *et al.*, 2009].

Algorithm 1 Feature Selection Optimization

Input:

1. Multiple attribute data $\{X, Y\}$;
2. Parameters α, β, k(optional) and the number of selected features K;
3. The initial projection matrix W_0;

Procedure:

Set $W = W_0$;
repeat
  Update M according to Eq. (8);
  Update $\boldsymbol{r}$ according to Alg. 2;
  Update $\boldsymbol{\delta}$ according to Eq. (10);
  Update W according to Eq. (11);
until Converges
Sort each feature according to $\|\mathbf{w}^i\|_2$ in descending order of each group;
return The group-wise top-K ranked features;

Algorithm 2 Cluster Assignment Estimation

Input: M;

Procedure:

Approximate F by top-ranked eigenvector of Q;
Calculate R_{11}, R_{12} by applying QR decomposition with column pivoting on F by Eq. (12);
Calculate $\hat{R}$ by Eq. (13);
calculate r by Eq. (14) for each attribute;
return Cluster assignment vector $\boldsymbol{r}$;

Optimizing W When Fixing M

The squared group-wise $\ell_{2,1}$ norm in Eq. (7) is still difficult to derive directly. To alleviate that, we introduce some positive dummy variables $\delta_{ij} \in \mathbb{R}^+$ which satisfies $\sum_i\sum_j \delta_{ij} = 1$. [Argyriou *et al.*, 2008] proves an upper bound of the squared $\ell_{2,1}$ norm in terms of the positive dummy variables

$$\sum_{i=1}^{k}(\|W_i\|_{2,1})^2 = (\sum_{i=1}^{k}\sum_{j=1}^{d}\|\boldsymbol{w}_{i,j}\|_2)^2 \le \sum_{i=1}^{k}\sum_{i=1}^{d}\frac{(\|\boldsymbol{w}_{i,j}\|_2)^2}{\delta_{ij}}$$

where $\boldsymbol{w}_{i,j} \in \mathbb{R}^{1\times m}$ is the row vector of W_i. Thus δ_{ij} can be updated by holding the equality:

$$\delta_{ij} = \|\boldsymbol{w}_{i,j}\|_2 / \sum_{j=1}^{d}\|\boldsymbol{w}_{i,j}\|_2. \tag{10}$$

Given a fixed M, each projection vector $\boldsymbol{w}$ can then be updated by optimize the following problem

$$\begin{aligned} \arg\min_{W} &\|W^T X - Y\|_F^2 + \alpha\sum_{i=1}^{k}\sum_{i=1}^{d}\frac{(\|\boldsymbol{w}_{i,j}\|_2)^2}{\delta_{ij}} \\ &- \beta\eta(1+\eta)\mathbf{Tr}(W(\eta I + M)^{-1}W^\top) \end{aligned} \tag{11}$$

which can be solved by gradient-type approach.

3.2 Estimating Attribute Assignment

The group-wise feature selection is conducted by the clustering structure of the attribute. However, given the M optimized by the previous algorithm, it is not readily possible to observe the cluster assignment of the attributes because M is spectrally relaxed. In this subsection, we propose an approach to acquire the cluster structure.

We first need to obtain a good approximation of the cluster indicator matrix F. Given M, we first apply Eigen decomposition $M = Q\Lambda Q^\top$ where each column of Q is the eigenvector and each diagonal element of Λ is the eigenvalue. Then we rank the columns of Q in decreasing order according to its corresponding eigenvalues, and the top-ranked k columns give an approximation of the cluster assignment matrix F. The number of the cluster k can be either manually specified or automatically explored by setting a threshold ($10e-8$ in our experiment) regarding to the absolute value of the eigenvalue.

After obtaining F, without loss of generality, we assume the optimized $W = [W_1, W_2, \cdots, W_k]^T$ where the submatrix W_i includes all attributes belonging to the i-th cluster. Let $\boldsymbol{t_i} = [t_{i1}, t_{i2}, \ldots, t_{in_i}]^T$ denote the largest eigenvector of ${W_i}^T W_i$, [Zha *et al.*, 2002] showed that F can be reformulated as

$$F^T = [\underbrace{t_{11}\boldsymbol{v}_1, \cdots, t_{1s_1}\boldsymbol{v}_1}_{cluster1}, \cdots, \underbrace{t_{k1}\boldsymbol{v}_k, \cdots, t_{ks_1}\boldsymbol{v}_k}_{clusterk}]$$

where $V^T = [\boldsymbol{v_1}, \boldsymbol{v_2}, \cdots, \boldsymbol{v_k}] \in \mathbb{R}^{k\times k}$ is an orthogonal matrix.

Since $\boldsymbol{v_i}$ is orthogonal to each other, the cluster structure can be acquired by picking up a column of F which has the largest norm as the first cluster, and orthogonalizing the other columns against this column. Then the same process is executed on the rest of columns until all clusters are identified. This process is identical to a QR decomposition with column pivoting on F

$$F^T = Q[R_{11}, R_{12}]P^T \tag{12}$$

where $Q \in \mathbb{R}^{k\times k}$ is an orthogonal matrix, $R_{11} \in \mathbb{R}^{k\times k}$ is an upper triangular matrix and $P \in \mathbb{R}^{m\times m}$ is a permutation matrix. Then we calculate the cluster assignment matrix $\hat{R} \in \mathbb{R}^{k\times m}$ by

$$\hat{R} = [I_k, R_{11}^{-1}R_{12}]P^T \tag{13}$$

where $I_k \in \mathbb{R}^{k\times k}$ is an identity matrix. The cluster assignment information can then be inferred from $\hat{R}$. The cluster membership of each attribute (column) is determined by the row index of the largest element (in absolute value) of the corresponding column in $\hat{R}$. Denote $\boldsymbol{r} \in \mathbb{R}^m$ as the cluster identification vector where r_i records which cluster the i-th class belongs to, then $\boldsymbol{r}$ can be calculated by

$$r_i = \arg\max_j \hat{r}_{ij} \tag{14}$$

where $\hat{r}_{ij}$ is the (i,j)-th entry of $\hat{R}$.

4 Experiments

In this section, we first verify the effectiveness of our proposed approach on one synthetic dataset. Since the proposed approach can be generalized to general multi-label problem, we evaluate the feature selection capability on various benchmark datasets. At last we evaluate the attribute prediction and zero-shot learning capabilities on image benchmark datasets. All the datasets are standardized to zero-mean and normalized by the standard deviation. For all approaches, the super parameters are selected via cross-validation. We cannot get the number of cluster k without any prior knowledge for real-world, thus we also select k by the prediction accuracy on a small subset of datasets.

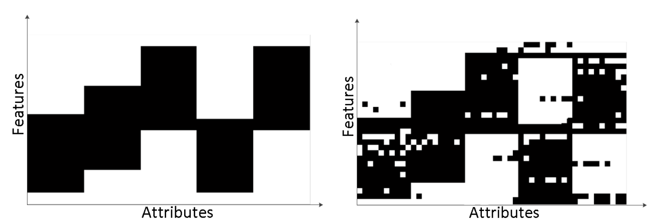
4.1 Simulation Study

Since it is difficult to obtain the groundtruth cluster structure for real applications, we first verify the effectiveness of the proposed approach in obtaining the cluster structures on simulated dataset. Following [Jacob *et al.*, 2009; Zhou *et al.*, 2011], we construct the synthetic data containing 5 clusters with 10 learning tasks in each cluster, generating a total number of 50 tasks. For the i-th task, a dataset $X_i \in \mathbb{R}^{d\times n}$ is randomly drawn from a normal distribution $N(0,1)$ for learning, with the dimension $d = 30$ and the sample size $n = 60$.

The projection model is constructed as follows. For the i-th cluster, we generate a cluster weight vector $\boldsymbol{w}_i^c \in \mathbb{R}^d$ drawn from the normal distribution $N(0, 900)$. Then 15 dimensions of $\boldsymbol{w}_i^c$ are randomly but carefully selected and assigned to zeros, to ensure all $\boldsymbol{w}^c$ are orthogonal to each other. Similarly, for the j-th task belonging to cluster i, we generate a task-specific weight vector $\boldsymbol{w}_j^s \in \mathbb{R}^d$ drawn from the normal distribution $N(0, 16)$ with the same dimensions of $\boldsymbol{w}_i^c$ assigned to zeros. Thus, the ultimate weight vector of the j-th task is the linear combination of the cluster and task-specific weight vector $\boldsymbol{w}_j = \boldsymbol{w}_i^c + \boldsymbol{w}_j^s$.

The corresponding response $\boldsymbol{y}_i$ of the i-th samples $\boldsymbol{x}_i$ of task j is then obtained by $\boldsymbol{y}_i = \boldsymbol{w}_j^T \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i$ where $\boldsymbol{\varepsilon}$ is the noise vector drawn from $N(0, 0.1)$. We choose 0.5 as the threshold to assign binary label to each sample.

We verify the effectiveness of our proposed approach by comparing the learned cluster structure and the selected features with the groundtruth. Based on the prior knowledge implied by the construction of the groundtruth, We set $k = 5$ and the number of selected features as $K = 15$. Figure 2 shows one example of the learned projection matrix 2(b) with the comparison of the groundtruth 2(a) where the white part represents zeros and the black part represents non-zeros. The result shows that our approach is able to roughly capture the correct group sparse structures.



(a) Groundtruth model (b) Learned model

Figure 2: The learned projection matrix and the corresponding groundtruth in the simulation experiments. The white parts are zeros and the black parts are non-zeros.

4.2 Feature Selection

We verify the feature selection capability on general multi-label datasets in this section. The experiment is conducted on 6 public benchmark feature selection datasets including one object image dataset **COIL100** [COI, 1996], one handwritten digit image dataset **USPS** [Hull, 1994], one spoken letter speech dataset **Isolet** [Fanty and Cole, 1991], three face image dataset **YaleB** [Georghiades *et al.*, 2001], **ORL** [Samaria and Harter, 1994] and **PIX10P**[1]. The statistics of the datasets are summarized in Table 2. We compare the proposed approach with the following representative feature selection algorithms: Fisher Score [Duda *et al.*, 2001], mRMR [Peng, 2005], Relief-F [Liu and Motoda, 2008], Information Gain [Cover and Thomas, 1991], MTFS [Argyriou *et al.*, 2008].

Following the common way to evaluate supervised feature selection, we assess the quality of selected features in terms of the classification performance [Han *et al.*, 2013; Cai *et al.*, 2013]. The larger classification accuracy is, the better performance the corresponding feature selection approach achieves. In our experiments, we employ linear Support Vector Machine (**SVM**) and k-nearest neighbors (k**NN**) classifier with $k = 3$ for evaluation. How to determine the optimal number of selected features is still an open question for feature selection; hence we vary the number of selected features as $\{10,30, 50 \ldots,90\}$ in this work. In each setup 50% samples are randomly selected for training and the remaining is for testing. Specific constrains are imposed to make sure the class labels of the training set are balanced. The whole experiment is conducted 10 rounds and average accuracies are reported.

Figure 1 shows the comparison results for **SVM** and kNN on the 6 benchmark datasets when 50 features are selected. The result shows that MTFS and the proposed framework outperform Fisher Score, mRMR and Information Gain. The performance gain comes from that Fisher Score, mRMR and Information Gain select features one by one while MTFS and FSMC select features in a batch model. It is consistent with what was suggested in [Tang and Liu, 2012] that it is better to analyze features jointly for feature selection. Besides, in most cases, the proposed framework outperforms MTFS. Better performance gain is usually achieved when fewer number of features are selected. This performance gain suggests that modeling label correlation can significantly improve feature selection performance for multi-class data.

4.3 Attribute Prediction

We then compare our approach with state-of-the-art attribute learning work [Chen *et al.*, 2014] (referred as MTAL) and [Jayaraman *et al.*, 2014] (referred as DSVA). Since MTAL is initially proposed for attribute ranking, we replace the original loss function with the one adopted in this paper for fair comparison. DSVA requires attribute groups as prior, thus we run k-means offline to obtain the clusters for datasets do not have such information.

[1]PIX10P is publicly available from https://featureselection.asu.edu/datasets.php

The experiments are conducted on three benchmark datasets: aYahoo [Farhadi *et al.*, 2009], Animals with Attributes (AwA) [Lampert *et al.*, 2009] and SUN attribute [Patterson and Hays, 2012] and the statistics of the datasets are summarized in Table 4. To obtain a good representation of the high-level attributes, we require that the features can capture both the spatial and context information. Thus, we constructed the features by pooling a variety types of feature histograms including GIST, HoG, SSIM. For **aPascal/aYahoo** and **AwA** datasets we use predefined seen/unseen split published with the datasets. For **SUN** dataset, 60% of categories are randomly split out as “seen” categories in each round with the rest as “unseen” categories. During training 50% of samples are randomly and carefully drawn from each seen categories to ensure the balance of the positive and negative attribute labels. The rest samples from “seen” classes and all samples from “unseen” classes are used for testing.

Table 3 shows the average prediction accuracy of each approach over all attributes by running the experiment 10 rounds. The result shows that for both “seen” and “unseen” categories, DSVA outperforms MTAL in prediction accuracy and our proposed approach further outperforms DSVA by $2\%\sim4\%$. DSVA decorrelates low-correlated attributes compared with MTAL thus achieves better prediction performance. However, the manually specified or off-line learned group structures are not able to achieve the optimal result. Our approach iteratively optimizes the clustering structure and the projection model, which achieves the best performance.

4.4 Zero-shot Learning

We also experiment on the zero-shot learning problem on all three datasets. Zero-shot learning aims to learn a classifier based on training samples from some seen categories, and classify some new samples to a new unseen category. We adopt the Direct Attribute Prediction (DAP) framework proposed in [Lampert *et al.*, 2009] with attribute prediction probability from each approaches as input. Since only continuous image level attribute labels are provided on the **SUN** dataset, we construct the class level attribute labels by thresholding the average attribute label values of all samples from the class. Same “Seen”\“Unseen” categories splits are adopted as previous experiments.

The Average classification accuracies of 10 rounds experiment are reported in Table 5. The result shows that on aYahoo and **AwA**, our approach achieves significant performance gains than the baseline approaches. The large number of categories in **SUN** dataset make the classification problem very hard which leads to all low performance of all approaches. Our approach still works better than the baseline approaches.

4.5 On Choosing the Parameters

The proposed framework has three important parameters - α controlling the sparsity of W, β controlling the contribution of modeling label correlation and $gamma$ controls the global penalty. We study the effect of each parameter by fixing the other to see how the performance of the proposed approach varies with the number of selected features. Due to the page

Table 1: Classification results (ACC%±std) of different feature slection algorithm on different datasets. (the higher the better).

Algorithm	DataSet	Fisher	mRMR	Relief-F	Information Gain	MTFS	Proposed
SVM	COIL100	60.66±3.54	55.72±3.34	62.80±2.56	62.00±2.84	78.77±2.35	**79.08±2.12**
	USPS	86.30±2.81	58.44±4.02	86.83±2.83	70.25±3.16	86.25±2.52	**93.15±2.18**
	Isolet	75.64±3.01	70.92±3.72	82.30±2.81	76.51±2.56	84.05±2.24	**87.06±1.98**
	YaleB	66.85±3.65	56.91±4.21	71.91±2.24	71.74±2.11	76.08±2.14	**78.17±2.18**
	ORL	46.50±4.21	84.51±2.32	67.18± 3.01	53.24±2.96	85.62±1.94	**90.51±1.78**
	PIX10P	93.56±2.01	90.45±3.32	96.00±1.77	92.01±1.97	96.81±1.54	**99.54±1.68**
kNN	COIL100	63.33±3.21	54.86±4.32	65.11±2.01	63.44±2.76	81.86±1.94	**82.48±1.68**
	USPS	89.39±2.11	59.17±3.72	89.61±2.01	74.70±2.76	90.44±1.54	**95.53±1.18**
	Isolet	75.38±2.45	57.56±3.42	79.87±2.21	73.71±2.42	77.01±2.14	**83.21±2.18**
	YaleB	69.17±3.24	58.41±3.72	65.53±2.81	65.37±2.42	77.08±2.45	**78.96±2.28**
	ORL	53.01±3.44	72.56±2.42	60.38±2.71	52.44±2.76	85.86±2.24	**88.10±2.10**
	PIX10P	94.56±1.91	86.45±2.22	96.00±1.81	86.04±2.04	97.81±1.54	**99.34±1.22**

Table 2: Statistics of the Feature Selection datasets

Dataset	# of Samples	# of Features	# of Classes
COIL100	7200	1024	100
YaleB	2414	1024	38
ORL	400	4096	40
PIX10P	100	10000	10
USPS	9298	256	10
Isolet	7797	617	150

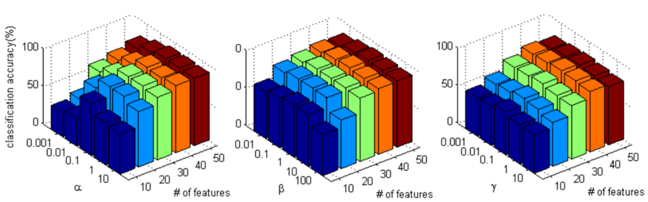


Figure 3: Parameter Analysis on SVM.

limitation, we only report the result on the **Isolet** dataset with SVM but we have similar observations in other datasets.

Figure 3 demonstrates the performance variance w.r.t. different parameters and the number of selected features. With the increase of β, the performance first increases, demonstrating the importance of modeling label correlation, and then decreases. This property is practically useful because we can use this pattern to set β. When α increases, the performance also increases dramatically, which suggests the capability of $\ell_{2,1}$-norm for feature selection. The performance also increases with γ and then decrease, but relatively stable. The best performance is achieved around 0.1.

5 Conclusions

In this paper, we proposed a clustering-base multi-task joint feature selection framework for semantic attribute prediction. Our approach employs both clustering and group-sparsity regularizers for feature selection. The clustering regularizer partitions the attributes into different groups where strong correlation lies among attributes in the same group while weak correlation exists between groups. The group-sparsity regularizer encourages intra-group feature-sharing and inter-group feature competition. With an efficient alternating optimization algorithm, the proposed approach is able to obtain a good group structure and select appropriate features to represent semantic attributes. The proposed approach was verified on both synthetic and real-world benchmark datasets with comparison with state-of-the-art approaches. The result shows effective group structure identification capability of our method, as well as its significant performance gains on feature selection, attribute prediction and zero-shot learning.

References

[Argyriou *et al.*, 2008] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. *J. Mach. Learn. Res.*, 73(3):243–272, December 2008.

[Bach, 2008] Francis R. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, June 2008.

[Blitzer *et al.*, 2006] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. EMNLP ’06, pages 120–128, 2006.

[Cai *et al.*, 2013] Xiao Cai, Feiping Nie, and Heng Huang. Exact top-k feature selection via l2,0-norm constraint. In *IJCAI ’13*, pages 1240–1246, 2013.

[Chen *et al.*, 2014] Lin Chen, Qiang Zhang, and Baoxin Li. Predicting multiple attributes via relative multi-task learning. In *Proc. of CVPR’14*, pages 1027–1034, June 2014.

[COI, 1996] Columbia Object Image Library (COIL-100). Technical report, Columbia University, 1996.

[Cover and Thomas, 1991] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[Duda *et al.*, 2001] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley & Sons, New York, 2 edition, 2001.

[Fanty and Cole, 1991] Mark Fanty and Ronald Cole. Spoken letter recognition. In *NIPS ’91*, pages 220–226. 1991.

[Farhadi *et al.*, 2009] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Com-*

Table 3: Average prediction accuracies of all attributes on Seen and Unseen categories (the higher the better).

DataSet	aPascal/aYahoo		AwA		SUN	
Methods	Seen	Unseen	Seen	Unseen	Seen	Unseen
MTAL	0.5967±0.020	0.5663±0.022	0.5976±0.011	0.5587±0.012	0.6326±0.021	0.6020±0.022
DSVA	0.6105±0.018	0.5826±0.019	0.6053±0.015	0.5622±0.018	0.6469±0.025	0.6165±0.027
Porposed	**0.6363±0.014**	**0.6011±0.015**	**0.6254±0.007**	**0.5837±0.008**	**0.6682±0.011**	**0.6324±0.013**

Table 4: Statistics of Attribute Prediction Image Datasets.

Dataset	aPascal/aYahoo	AwA	SUN
# of images	15339	30475	14340
# of attributes	64	85	102
# of classes	32	50	611
# of features	2429	1200	1112

Table 5: Zero-shot learning accuracy on both real dataset.

	aYahoo	AwA	SUN
MTAL	0.1834	0.2953	0.1842
DSVA	0.2052	0.3085	0.2010
Proposed	**0.2262**	**0.3258**	**0.2133**

puter Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 1778–1785, June 2009.

[Farhadi *et al.*, 2010] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *Proc. of CVPR'10*, pages 2352–2359, June 2010.

[Georghiades *et al.*, 2001] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):643–660, 2001.

[Han *et al.*, 2013] Yahong Han, Yi Yang, and Xiaofang Zhou. Co-regularized ensemble for feature selection. In *IJCAI '13*, pages 1380–1386, 2013.

[Hull, 1994] J. J. Hull. A database for handwritten text recognition research. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(5):550–554, May 1994.

[Jacob *et al.*, 2009] Laurent Jacob, Jean philippe Vert, and Francis R. Bach. Clustered multi-task learning: A convex formulation. In *Proc. of NIPS'09*, pages 745–752. 2009.

[Jayaraman *et al.*, 2014] D. Jayaraman, Fei Sha, and K. Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *Proc. of CVPR'14*, pages 1629–1636, June 2014.

[Kovashka *et al.*, 2012] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *Proc. of CVPR'12*, pages 2973–2980, June 2012.

[Lampert *et al.*, 2009] C.H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proc. of CVPR'09*, pages 951–958, June 2009.

[Liu and Motoda, 2008] H. Liu and H. Motoda, editors. *Computational Methods of Feature Selection*. Chapman & Hall, 2008.

[Patterson and Hays, 2012] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Proc. of CVPR'12*, 2012.

[Peng, 2005] F. Ding C. Peng, H. Long. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(8):1226–1238, 2005.

[Quattoni *et al.*, 2007] A. Quattoni, M. Collins, and T. Darrell. Learning visual representations using images with captions. In *CVPR '07*, pages 1–8, June 2007.

[Samaria and Harter, 1994] F.S. Samaria and A.C. Harter. Parameterisation of a stochastic model for human face identification. In *Applications of Computer Vision' 94*, pages 138–142, Dec 1994.

[Scheirer *et al.*, 2012] W.J. Scheirer, N. Kumar, P.N. Belhumeur, and T.E. Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *Proc. of CVPR'12*, pages 2933–2940, June 2012.

[Song *et al.*, 2012] Fengyi Song, Xiaoyang Tan, and Songcan Chen. Exploiting relationship between attributes for improved face verification. In *Proc. of BMVC'12*, pages 27.1–27.11, 2012.

[Tang and Liu, 2012] Jiliang Tang and Huan Liu. Feature selection with linked data in social media. In *SDM '12*, pages 118–128. SIAM, 2012.

[Tang *et al.*, 2014] Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data Classification: Algorithms and Applications. Editor: Charu Aggarwal, CRC Press In Chapman & Hall/CRC Data Mining and Knowledge Discovery Series*, 2014.

[Wang *et al.*, 2015] Yilin Wang, Suhang Wang, Jiliang Tang, Huan Liu, and Baoxin Li. Unsupervised sentiment analysis for social media images. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 2378–2379, 2015.

[Zha *et al.*, 2002] Hongyuan Zha, Xiaofeng He, Chris Ding, Ming Gu, and Horst D. Simon. Spectral relaxation for k-means clustering. In *Proc. of NIPS'02*, pages 1057–1064. 2002.

[Zhou *et al.*, 2011] Jiayu Zhou, Jianhui Chen, and Jieping Ye. Clustered multi-task learning via alternating structure optimization. In *Proc. of NIPS'11*, pages 702–710. 2011.