Exclusive Lasso for Multi-task Feature Selection

Yang Zhou[1] **Rong Jin**[1] **Steven C.H. Hoi**[2]
[1]Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48910 USA
{zhouyang,rongjin}@msu.edu

[2]School of Computer Engineering
Nanyang Technological University
Singapore 639798
chhoi@ntu.edu.sg

Abstract

We propose a novel group regularization which we call *exclusive lasso*. Unlike the group lasso regularizer that assumes covarying variables in groups, the proposed exclusive lasso regularizer models the scenario when variables in the same group compete with each other. Analysis is presented to illustrate the properties of the proposed regularizer. We present a framework of kernel based multi-task feature selection algorithm based on the proposed exclusive lasso regularizer. An efficient algorithm is derived to solve the related optimization problem. Experiments with document categorization show that our approach outperforms state-of-the-art algorithms for multi-task feature selection.

1 INTRODUCTION

Group regularizers like group lasso (Yuan and Lin 2005) have been extensively studied in both the statistics and machine learning fields. The objective of group lasso is to select a group of features simultaneously for a given task(s). The key assumption behind the group lasso regularizer is that if a few features in a group are important, then most of the features in the same group should also be important. However, in many real-world applications, we may come to the opposite observation. Consider the problem of multi-category document classification. The existing approaches for multi-task feature selection usually assume a positive correlation among the categories, namely, when one keyword is important for several categories, it is also expected to be important for the other categories. This positive correlation is usually captured by a group lasso regularizer, where a group is defined for every word w to include the feature weights of all categories for w. However, when our objective is to differentiate the related categories, we may expect a negative correlation among categories, namely, if word w is deemed to be important for one category, it becomes less likely for w to be an important word for the other categories. It is clear that such a negative correlation violates the assumption made by most of the existing approaches for multi-task feature selection. Another example is visual object recognization where the signature visual patterns of one object class tend to be less useful for identifying objects of the other classes.

In order to capture the negative correlation among categories, we propose the exclusive lasso regularizer. Different from the group lasso regularizer, if one feature in a group is given a large weight, the exclusive lasso regularizer tends to assign small or even zero weights to the other features in the same group. We present a simple analysis to verify the exclusive nature of the proposed regularizer. Based on the proposed exclusive lasso regularizer, we present a framework for kernel based multi-task learning. An efficient algorithm is derived to solve the related optimization problem. Empirical studies with document categorization verify that the proposed regularizer is effective for multi-task feature selection.

2 RELATED WORK

We briefly review the related work in group regularization and multi-task feature selection.

2.1 GROUP REGULARIZATION

Group lasso (Yuan and Lin 2005) has been studied extensively and applied to a number of machine learning problems. It uses the ℓ_1 norm, which is the-

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

oretically proven to generate sparse solutions (Tibshirani 1996), to select groups of variables that are grouped by the ℓ_2 norm. In the same vein, new regularizers have been proposed for grouped and hierarchical selection of variables. Zou and Hastie (2005) and Kim et al. (2006) combined ℓ_1 and ℓ_2 norm to form a more structured regularization. In (Kowalski et al. 2009, Kowalski 2009), the authors generalized the group lasso by exploring the mixed norm for combining groups of variables. Zhao et al. (2009) further extended the idea of group lasso and proposed a general Composite Absolute Penalties (CAP) family, which allows for (i) different norms for combining variables within the same groups, and (ii) overlapping in variables between groups. Let $\beta = (\beta_1, \cdots, \beta_p)^\top$ be the p variables to be regularized. Given the grouping structure $G = \{G_k \subset \{1, \ldots, p\}, k = 1, \cdots, K\}$, and a vector of norm parameters $\gamma = (\gamma_0, \gamma_1, \cdots, \gamma_K) \in \mathbb{R}_+^{K+1}$, the regularizer $T_{G,\gamma}(\beta)$ is defined as follows

$$T_{G,\gamma}(\beta) = \sum_k \left(\sum_{m \in G_k} |\beta_m|^{\gamma_k} \right)^{\gamma_0/\gamma_k}$$

Although our work can be viewed as a special case of mixed norm and the general CAP family, this study is distinguished from the existing ones in two aspects: (i) Unlike the previous studies that only emphasize the sparsity of solutions caused by the regularization, our in depth analysis also reveals that the exclusive lasso is able to introduce competitions among variables within the same group, which is a key property for capturing the negative correlation among tasks; (ii) We apply the exclusive lasso regularization method to kernel based multi-task learning. It results in a mini-max optimization problem that is beyond the capability of the existing algorithms for group regularization. We present an efficient algorithm for solving the related min-max optimization based on the subgradient descent method.

2.2 MULTI-TASK FEATURE SELECTION

Multi-task Learning (MTL) (Caruana 1997) has proven to be useful both theoretically (Baxter 2000, Ben-david and Schuller 2003, Ando and Zhang 2005) and experimentally (Evgeniou et al. 2005, Jebara 2004, Torralba et al. 2004, Chen et al. 2009). Most MTL algorithms assume a positive correlation among tasks. For example, Evgeniou et al. (2005), Bakker and Heskes (2003) assume that functions for different tasks are similar to each other; Baxter (2000), Ben-david and Schuller (2003) and Caruana (1997) assume a common representation of data that is shared by all the tasks.

Many algorithms have been proposed for multi-task feature selection, an important problem in multi-task learning. Xiong et al. (2007) imposed an automatic relevance determination prior on the hypothesis classes associated with individual tasks and regularized the variance of the hypothesis parameters. Argyriou et al. (2006) and Obozinski et al. (2006) used the $\ell_{1,2}$ norm, similar to group lasso, for regularizing features of different tasks. It encourages multiple predictors to have similar parameter sparsity patterns. Jebara (2004) introduced a common vector of binary feature selection switches shared by all the tasks. Lee et al. (2007) introduced meta-features for feature selection in related tasks. Chen et al. (2009) assumed a shared feature space in a linear form of low-dimentional feature map across multiple tasks. All the existing algorithms for multi-task feature selection assume a positive correlation among tasks, and aim to learn a common subset of features for all tasks. In contrast, our proposed exclusive lasso regularizer assumes a negative correlation among tasks, and introduces competition among variables within the same group.

3 EXCLUSIVE LASSO

In this section, we first present the formulation of the exclusive lasso regularizer and its basic properties. Then we apply the exclusive lasso to multi-task learning in which each task is formulated as a multiple kernel learning problem. We derive an efficient algorithm to solve the related optimization problem.

Notations: We use index i for instances, j for features, and k for tasks. We use n to denote the total number of instances, d for the number of features, and m for the number of tasks.

3.1 MULTI-TASK LEARNING WITH LINEAR CLASSIFIERS

We consider a multi-task classification problem. Let $\mathcal{D} = \{(x_i, y_i), i = 1, \ldots, n\}$ be the training data, where $x_i \in \mathbb{R}^d$ is the input pattern and $y_i = (y_i^1, \ldots, y_i^m) \in \{-1, +1\}^m$ is the assigned categories with $y_i^k = 1$ if x_i is assigned to category k and $y_i^k = -1$ otherwise. For simplicity, we assume a linear classifier $f_k(x) = \beta_k^\top x$ where $\beta_k = (\beta_k^1, \ldots, \beta_k^d) \in \mathbb{R}^d$ is the combination weights. We thus have the following optimization problem for multi-task learning:

$$\min_\beta \quad \frac{1}{2} V(\beta) + C \sum_{i=1}^n \sum_{k=1}^m \ell(y_i^k, f_k(x_i)) \qquad (1)$$

where $\ell(z)$ is a loss function that measures the mismatch between y_i^k and the predicted value $f_k(x_i)$; $V(\beta)$ is a regularizer that controls the complexity of combination weights β. We assume a competitive nature among the features shared by all the tasks, i.e., if a very large weight is assigned to the jth feature for

one task, we expect the weights for the same feature to be small or even zero for the other tasks. To this end, we introduce the following regularizer:

$$V(\beta) = \sum_{j=1}^{d} \left(\sum_{k=1}^{m} \left| \beta_k^j \right| \right)^2 \tag{2}$$

As indicated in the above expression, we introduce an ℓ_1 norm to combine the weights for the same feature used by different tasks and an ℓ_2 norm to combine the weights of different features together. Since ℓ_1 norm tends to achieve a sparse solution, the construction in $V(\beta)$ essentially introduces a competition among different tasks for the same feature. We refer to the above regularizer as *exclusive lasso*. Using the exclusive lasso as a regularizer, we have the overall optimization problem written as

$$\min_{\beta} \frac{1}{2} \sum_{j=1}^{d} \Big(\sum_{k=1}^{m} \left| \beta_k^j \right| \Big)^2 + C \sum_{i=1}^{n} \sum_{k=1}^{m} \ell \left(y_i^k, f_k(x_i) \right)$$

An alternative approach to the regularizer shown above is to introduce a constraint for β:

$$\begin{aligned} \min_{\beta} \quad & \sum_{i=1}^{n} \sum_{k=1}^{m} \ell \left(y_i^k, f_k(x_i) \right) \\ s.t. \quad & \sqrt{\sum_{j=1}^{d} \left(\sum_{k=1}^{m} \left| \beta_k^j \right| \right)^2} \leq \gamma \end{aligned}$$

where γ is a predefined constant.

3.2 UNDERSTANDING THE EXCLUSIVE LASSO REGULARIZER

One of the fundamental questions is how the exclusive lasso regularizer introduces competitions among different tasks for the same feature. To illustrate this point, we consider the following projection problem,

$$\min_{\beta \in \mathcal{G}} |\beta - \bar{\beta}|_2^2 \tag{3}$$

where $\bar{\beta}$ is an existing solution, domain $\mathcal{G}$ is defined as

$$\mathcal{G} = \left\{ \beta = (\beta_1, \ldots, \beta_m)^\top \in \mathbb{R}^{m \times d} : \sqrt{\sum_{j=1}^{d} \Big(\sum_{k=1}^{m} |\beta_k^j| \Big)^2} \leq \gamma \right\}$$

The projection problem in (3) directly demonstrates how the domain $\mathcal{G}$ shapes a solution $\bar{\beta}$, which essentially illustrates the effect of the exclusive lasso regularizer. Projection is an important operation that is used by many optimization algorithms (e.g., subgradient descent). In addition, important problems such as constrained least square regression can be cast into a projection problem (Bishop 2006).

We first convert (3) into a convex-concave problem:

$$\min_{\beta} \max_{\lambda \geq 0} |\beta - \bar{\beta}|_2^2 + 2\lambda \left(\sqrt{\sum_{j=1}^{d} \left(\sum_{k=1}^{m} \left| \beta_k^j \right| \right)^2} - \gamma \right) \tag{4}$$

The following proposition allows us to simplify the problem in (4).

Proposition 1. *Given a vector $\beta = (\beta_1, \ldots, \beta_m)^\top$ where $\beta_k \in \mathbb{R}^d$, we can rewrite $\sqrt{\sum_{i=1}^{d} \left(\sum_{k=1}^{m} |\beta_k^i| \right)^2}$ as*

$$\sqrt{\sum_{j=1}^{d} \left(\sum_{k=1}^{m} \left| \beta_k^j \right| \right)^2} = \max_{\alpha \in \Delta} \alpha^\top \beta$$

where domain Δ is defined as

$$\Delta = \Bigg\{ \alpha = (\alpha_1, \ldots, \alpha_m)^\top : \alpha_k = (\alpha_k^1, \ldots, \alpha_k^d) \in \mathbb{R}^d, \\ k = 1, \ldots, m, \ \sum_{j=1}^{d} \max_{1 \leq k \leq m} [\alpha_k^j]^2 \leq 1 \Bigg\}$$

Using the above proposition, we have the following lemma that simplifies problem (4).

Lemma 1. *Problem (4) is equivalent to the following optimization problem*

$$\min_{\tau} \left\{ 2\gamma |\tau|_2 + \sum_{j=1}^{d} \sum_{k=1}^{m} [|\bar{\beta}_k^j| - \tau_j]_+^2 \right\} \tag{5}$$

where $[x]_+ = \max(x, 0)$. The optimal solution of β is computed as

$$\beta_k^j = [\bar{\beta}_k^j - \tau_j]_+, \quad j = 1, \ldots, d, \quad k = 1, \ldots, m$$

Proof. Using Proposition 1, we rewrite (4) as

$$\max_{\alpha, \lambda} \min_{\beta} |\beta - \bar{\beta}|_2^2 + 2\lambda \left(\alpha^\top \beta - \gamma \right)$$

Taking the minimization over β, we have

$$\max_{\alpha, \lambda} \left\{ -2\lambda\gamma - |\bar{\beta} - \lambda\alpha|_2^2 : \sum_{j=1}^{d} \max_{1 \leq k \leq m} [\alpha_k^j]^2 \leq 1 \right\} \tag{6}$$

with $\beta = \bar{\beta} - \lambda\alpha$. To simplify our analysis, we define $\tau_j = \max_{1 \leq k \leq m} \lambda |\alpha_k^j|$, and Eqn. (6) can be written as

$$\min_{\tau, \lambda} \left\{ 2\lambda\gamma + \sum_{j=1}^{d} \sum_{k=1}^{m} [|\bar{\beta}_k^j| - \tau_j]_+^2 : |\tau|_2 \leq \lambda \right\}$$

or

$$\min_{\tau} \left\{ 2\gamma |\tau|_2 + \sum_{j=1}^{d} \sum_{k=1}^{m} [|\bar{\beta}_k^j| - \tau_j]_+^2 \right\}$$

□

As indicated by the above lemma, whenever $\bar{\beta}_k^j$ is smaller than threshold τ_j, we have β_k^j become zero. The following proposition shows a sufficient condition for $\beta_k^j = 0$.

Proposition 2. *For any feature j, we have $\beta_k^j = 0$ if $|\bar{\beta}_k^j| \leq \left(\sum_{k=1}^m |\bar{\beta}_k^j| - \gamma\right)/m$.*

Proof. We consider the first order optimality condition for τ, i.e.,

$$\gamma \frac{|\tau_j|}{|\tau|_2} + \sum_{k=1}^m [|\bar{\beta}_k^j| - \tau_j]_+ \partial_{\tau_j} [|\bar{\beta}_k^j| - \tau_j]_+ = 0$$

where $\partial_x f(x)$ is the subgradient of function $f(x)$. Notice that $\partial_{\tau_j}[|\bar{\beta}_k^j| - \tau_j]_+ \in [-1, 0]$, and is -1 when $|\bar{\beta}_k^j| < \tau_j$. Hence, the above optimality condition implies that

$$\sum_{k=1}^m [|\bar{\beta}_k^j| - \tau_j]_+ \leq \gamma \frac{|\tau_j|}{|\tau|_2} \leq \gamma$$

Since $[|\bar{\beta}_k^j| - \tau_j]_+ \geq |\bar{\beta}_k^j| - \tau_j$, we have $\sum_{k=1}^m |\bar{\beta}_k^j| - m\tau_j \leq \gamma$, which leads to the result in the proposition. □

As indicated in the above proposition, when some tasks take significantly smaller weights for feature j than the other tasks, the regularizer will enforce the weights of feature j to be zero for these tasks, leading to the competition of feature j among tasks. Parameter γ is used to control the degree of domination. A large γ requires a large gap among the weights for the same feature before the small weights can be reduced to zero; similarly, a small γ allows us to reduce small weights to zero even when the gap among the weights for the same feature is still small.

3.3 MULTI-TASK LEARNING WITH KERNEL CLASSIFIERS

We extend the exclusive lasso discussed above to the kernel case. We follow the Multiple Kernel Learning scheme (Lanckriet et al. 2004, Bach et al. 2004, Sonnenburg et al. 2006) and use the proposed regularizer to combine multiple kernels. In particular, we consider there are d kernels at our disposal, denoted by $\mathcal{W} = \{W^j \in \mathbb{S}_+^n, j = 1, \ldots, d\}$. We assume that each kernel matrix in $\mathcal{W}$ is appropriately normalized (e.g., $\mathrm{tr}(W^j) = 1$). For each task k, we assume that its kernel matrix, denoted by K^k, is a linear combination of the kernel matrices in $\mathcal{W}$, i.e., $K^k = \sum_{j=1}^d \lambda_k^j W^j$, where $\lambda_k = (\lambda_k^1, \ldots, \lambda_k^d) \in \mathbb{R}_+^d$ is the combination weights. For each individual task, the learning of combination weights λ_k, often referred to as multiple kernel learning, is cast into the following optimization problem:

$$\min_{\lambda_k \in \mathbb{R}_+^d} \max_{\gamma_k \in [0,C]^n} \left\{ \gamma_k^\top \mathbf{1} - \frac{1}{2}(\gamma_k \circ z_k)^\top \Big(\sum_{j=1}^d W^j \lambda_k^j \Big)(\gamma_k \circ z_k) \right\}$$

where $z_k = (y_1^k, y_2^k, \ldots, y_n^k)$ and $\circ$ is the element-wise dot product. Similar to the linear case, by assuming the exclusive nature among tasks in competing for kernels in $\mathcal{W}$, we introduce the exclusive lasso for regularizing the kernel weights $\lambda = (\lambda_1; \ldots; \lambda_m)$ assigned to different tasks, leading to the following optimization problem:

$$\min_{\lambda_k \in \mathbb{R}_+^d} \max_{\gamma_k \in [0,C]^n} \frac{r}{2} \sum_{j=1}^d \left(\sum_{k=1}^m \lambda_k^j \right)^2 + \sum_{k=1}^m \left(\gamma_k^\top \mathbf{1} - \frac{1}{2}(\gamma_k \circ z_k)^\top \left[\sum_{j=1}^d W^j \lambda_k^j \right] (\gamma_k \circ z_k) \right) \tag{7}$$

where r is a predefined parameter that weights the importance of the regularizer. The following theorem shows the sparsity in the solution of λ and the competition among tasks for kernels caused by the exclusive lasso regularizer.

Theorem 1. *Provided the solution γ, for each kernel W^j, we have $\lambda_k^j > 0$ only if*

$$k = \arg\max_{1 \leq k' \leq m} (\gamma_{k'} \circ z_{k'})^\top W^j (\gamma_{k'} \circ z_{k'})$$

This theorem follows directly from the result in Proposition 4, which will be stated later.

3.4 ALGORITHM

We focus on solving the problem in (7). A straightforward approach is the subgradient method. Define

$$g(\gamma, \lambda) = \frac{r}{2} \sum_{j=1}^d \left(\sum_{k=1}^m \lambda_k^j \right)^2 + \sum_{k=1}^m \left(\gamma_k^\top \mathbf{1} - \frac{1}{2}(\gamma_k \circ z_k)^\top \Big[\sum_{j=1}^d W^j \lambda_k^j \Big] (\gamma_k \circ z_k) \right)$$

We also define $f(\gamma) = \min_{\lambda_k \in \mathbb{R}_+^d} g(\gamma, \lambda)$. Hence, the problem in (7) can be viewed as a maximization problem:

$$\gamma = \arg\max_{\gamma_k \in [0,C]^n} f(\gamma).$$

We thus can apply the subgradient ascent approach to directly maximizing $f(\gamma)$. In each iteration of the subgradient ascent method, we compute the gradient

of $f(\gamma)$, denoted by $\nabla f(\gamma)$, and the new solution is obtained by moving the existing solution γ along the direction of $\nabla f(\gamma)$, i.e., $\gamma \leftarrow \pi_G(\gamma + s\nabla f(\gamma))$, where $G = \{\gamma = (\gamma_1, \ldots, \gamma_m)^\top \in \mathbb{R}^{m\times n} : \gamma_k \in [0, C]^n, k = 1, \ldots, m\}$ and $\pi_G(x)$ projects solution x onto the domain G. Evidently, there are two key parameters that need to be computed efficiently, i.e., step size s and $\nabla f(\gamma)$. The following proposition allows us to compute $\nabla f(\gamma)$, similar to (Xu et al. 2008).

Proposition 3. *We have the gradient of $f(\gamma)$ computed as*

$$\nabla_{\gamma_k} f(\gamma) = \mathbf{1} - \left[\sum_{j=1}^{d} \lambda_k^j \left(W^j \circ z_k z_k^\top\right)\right] \gamma_k \tag{8}$$

where λ_k^j is the minimizer of $g(\gamma, \lambda)$, i.e.,

$$\lambda = \arg\min_{\lambda_k \in \mathbb{R}_+^d} g(\gamma, \lambda).$$

As indicated in the above proposition, to compute the gradient of $f(\gamma)$, it is important to efficiently compute λ that minimizes $g(\gamma, \lambda)$. To this end, we rewrite $g(\gamma, \lambda)$ to highlight its dependency on λ:

$$g(\gamma, \lambda) = a - \sum_{k=1}^{m}\sum_{j=1}^{d} b_k^j \lambda_k^j + \frac{r}{2}\sum_{j=1}^{d}\left(\sum_{k=1}^{m} \lambda_k^j\right)^2 \tag{9}$$

where

$$a = \sum_{k=1}^{m} \gamma_k^\top \mathbf{1}, \quad b_k^j = \frac{1}{2}(\gamma_k \circ z_k)^\top W^j (\gamma_k \circ z_k) \tag{10}$$

In order to minimize $g(\gamma, \lambda)$ with respect to λ, we define h_j as

$$h_j = -\sum_{k=1}^{m} b_k^j \lambda_k^j + \frac{r}{2}\left(\sum_{k=1}^{m} \lambda_k^j\right)^2 \tag{11}$$

Since $g(\gamma, \lambda) = a + \sum_{j=1}^{d} h_j$ and each h_j only involves variables $\lambda_k^j, k = 1, \ldots, m$, we could optimize h_j separately. The following proposition gives the optimal solution that minimizes h_j.

Proposition 4. *Assume $b_k^j \neq b_{k'}^j$ for any $k \neq k'$ and any j. The optimal $\lambda_k^j, k = 1, \ldots, m$ that minimizes h_j is*

$$\lambda_k^j = \begin{cases} \bar{\lambda}^j & k = \arg\max_{1\le k'\le m} b_{k'}^j \\ 0 & \textit{otherwise} \end{cases}$$

where $\bar{\lambda}^j$ is computed as $\bar{\lambda}^j = \frac{1}{r}\max_{1\le k\le m} b_k^j$.

Proof. For the sake of simplicity, we drop index j and consider a general problem as follows

$$\min_{\lambda \in \mathbf{R}_+^m} \quad -\sum_{k=1}^{m} b_k \lambda_k + \frac{r}{2}\left(\sum_{k=1}^{m} \lambda_k\right)^2$$

We define $\lambda_k = \eta_k + \bar{\lambda}$ and $\bar{\lambda} = \sum_{k=1}^{m} \lambda_k / m$. We therefore have $\eta_k \ge -\bar{\lambda}$ and $\sum_{k=1}^{m} \eta_k = 0$. Thus the original problem can be transformed into a problem of $\bar{\lambda}$ and η, i.e.,

$$\begin{aligned} \min_{\bar{\lambda}, \eta} \quad & \frac{rm^2}{2}\bar{\lambda}^2 - \sum_{k=1}^{m} b_k \eta_k - \bar{\lambda}\sum_{k=1}^{m} b_k \\ \text{s. t.} \quad & \bar{\lambda} \ge 0, \ \sum_{k=1}^{m} \eta_k = 0 \\ & \eta_k \ge -\bar{\lambda}, \ k = 1, \ldots, m \end{aligned}$$

We consider the solution for η when $\bar{\lambda}$ is fixed, which leads to the following linear programming problem:

$$\begin{aligned} \min_{\eta} \quad & -\sum_{k=1}^{m} b_k \eta_k \\ \text{s. t.} \quad & \sum_{k=1}^{m} \eta_k = 0, \quad \eta_k \ge -\bar{\lambda}, \ k = 1, \ldots, m \end{aligned}$$

Since $b_k \ge 0$, it is clear that the optimal solution for the above linear programming problem is

$$\eta_k = \begin{cases} (m-1)\bar{\lambda} & k = \arg\max_{1\le k'\le m} b_{k'} \\ -\bar{\lambda} & \text{otherwise} \end{cases}$$

Using the solution for η, we have the following problem for $\bar{\lambda}$

$$\min_{\bar{\lambda} \ge 0} \quad \frac{rm^2}{2}\bar{\lambda}^2 - m\bar{\lambda}\max_{1\le k\le m} b_k$$

It is obvious that $\bar{\lambda} = \max_{1\le k\le m} b_k/(rm)$. □

Note that Proposition 4 only addresses the situation when there is a unique element for $k = \arg\max_{1\le k'\le m} b_{k'}^j$. Similar results can be easily derived when multiple elements tie for the maximum value of b_k^j. This proposition clearly demonstrates the competition of kernels among tasks resulting from the exclusive lasso regularizer. Using the result from Proposition 4, we can efficiently compute the optimal λ for a given γ, which allows us to efficiently compute the gradient of $f(\gamma)$ in Eqn. (8).

We determine the step size s by the backtracking line search (Boyd and Vandenberghe 2004). Finally, the duality gap is used to check the convergence. Given the solution λ^* and γ^*, the duality gap is defined as

$$\delta = \min_{\lambda_k \in \mathbb{R}_+^d} g(\gamma^*, \lambda) - \max_{\gamma_k \in [0, C]^n} g(\gamma, \lambda^*), \tag{12}$$

Table 1: Metadata of the Yahoo datasets. m, N, "MaxNPI" and "MinNPI" denote the number of subcategories, the total number of instances, the maximum and minimum number of positive instances for each subcategory respectively.

Dataset	m	N	MaxNPI	MinNPI
Arts	19	7441	1838	104
Business	17	11182	9723	110
Computers	23	12371	6559	108
Education	14	11817	3738	127
Entertainment	14	12691	3687	221
Health	14	9109	4703	114
Recreation	18	12797	2534	169
Reference	15	7929	3782	156
Science	22	6345	1548	102
Social	21	11914	5148	104
Society	21	14507	7193	113

where $\min_{\lambda \in \mathbb{R}^d_+} g(\gamma^*, \lambda)$ can be computed efficiently using Proposition 4, and $\max_{\gamma_k \in [0,C]^n} g(\gamma, \lambda^*)$ is solved by a kernel SVM.

4 EXPERIMENTS

We evaluate the efficacy of the proposed exclusive lasso regularizer by multi-task feature selection. We use the Yahoo dataset (Ueda and Saito 2003) in our experiments. This multi-topic web page categorization dataset was collected from 11 top-level categories ("Arts", "Business", "Computers", etc.) in the "yahoo.com" domain. Each top-level category is further divided into a number of second-level subcategories. Each subcategory is an individual task in our multi-task classification algorithm. We preprocessed the datasets by removing topics with less than 100 documents and documents with no topics. 300 keywords are randomly sampled for each dataset after the high-frequency and low-frequency terms are removed. Metadata of the datasets can be found in Table 1. By constructing a kernel for each individual keyword, we apply the proposed method for kernel based multi-task learning to document categorization. Throughout this study, a linear kernel is used by all the methods and for all the experiments because it is proven to be effective for document categorization.

4.1 EVALUATION

We use the following two algorithms as baselines in our experiments to compare with the proposed exclusive lasso algorithm:

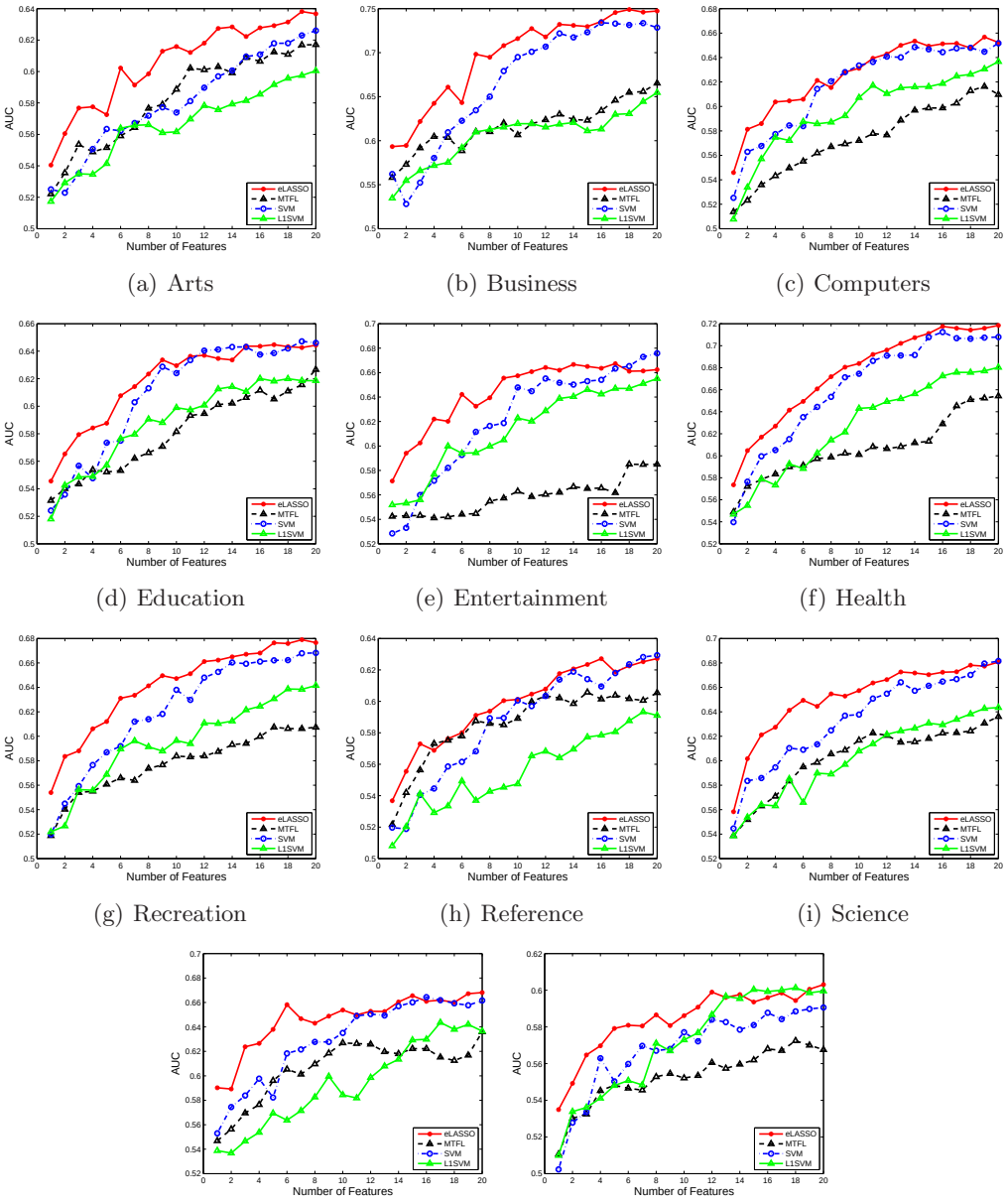
- SVM feature selection (Bradley and Mangasarian 1998). We train a linear SVM classifier for each category and select the features that have the largest absolute values in their coefficients. Note that the SVM classifiers are trained independently in this case, and therefore features are selected independently for each task. We used two kinds of SVMs: the L2-regularized SVM which is most commonly used, and the L1-regularized SVM (Zhu et al. 2004) which enforces sparsity of the classifiers.
- Multi-task Feature Learning (MTFL) (Argyriou et al. 2006). MTFL used the group lasso to jointly penalize the features used by different tasks. It encourages multiple predictors to have similar parameter sparsity patterns, and aims to learn a subset of features common to all the tasks. We use the hinge loss function in the MTFL algorithm because our work follows directly the SVM framework.

To evaluate the efficacy of feature selection, we randomly sample 10 examples from each subcategory for training and use the remaining documents for testing. We use a small number of training examples because it is well known that in document categorization, with sufficient numbers of training documents, any feature selection method works well. After training the classification models, we choose the top features for each subcategory that have the largest weights. An SVM classifier is constructed for each subcategory by using the selected features, and its classification accuracy computed over the test documents is used to evaluate the efficacy of feature selection algorithms. The hypothesis is that the more effective the feature selection algorithm is, the more accurate the SVM classifier will be. The area under the receiver operating characteristic curve (AUC) (Egan 1975) is used in our study as performance metric. We vary the number of selected features from one to twenty, and repeat each experiment ten times. The reported AUC for each dataset is averaged over ten random trials.

The regularization constant C of SVM is set to be 10 for all the SVM classifiers in the experiments according to our experience. The regularizer parameter r in Eqn. (7) is set to be 1 in all the experiments. Note that we did not employ cross validation to determine the parameters because of the small number of training samples.

4.2 RESULTS

Figure 1 shows the average AUC of the 11 datasets of the Yahoo data collection for the three feature selection methods in comparison. We observe that the pro-



(a) Arts (b) Business (c) Computers

(d) Education (e) Entertainment (f) Health

(g) Recreation (h) Reference (i) Science

(j) Social (k) Society

Figure 1: AUC of exclusive lasso (eLASSO), feature selection with SVM and L1-regularized SVM (L1SVM), and MTFL on the 11 datasets of Yahoo data collection. The x axis is the number of selected features from each category used in the testing phase, and the y axis is the corresponding AUC measure. All performances are averaged for 10 runs each with a random sampling of training instances. When there are not expected number of features due to sparsity, some features with 0 weight are randomly sampled.

posed algorithm for multi-task feature selection outperforms the other three baseline algorithms. This is not surprising given the topic structure in the Yahoo data collection. Although documents within each dataset belong to a common topic and therefore are expected to share many common terms, our goal is to classify documents in each dataset further into subcategories. As a result, we need to select discriminative terms that are sufficient to differentiate the subcategories, not the terms that are commonly shared among subcategories. These discriminative terms are more likely to be discovered by the proposed exclusive lasso algorithm since a discriminative term for a given subcategory is unlikely to be also discriminative for another subcategory. Finally, we observe that the advantage of the proposed algorithm over the other

comparative methods tends to diminish as the selected number of features is increased. This is within our expectation, as any feature selection method will work well if we aim to select most of the features.

5 CONCLUSIONS

We introduce a new regularization which we call exclusive lasso in this paper. We give detailed theoretical analysis to illustrate that the proposed exclusive lasso regularizer is able to introduce competitions among variables and thus generate sparse solutions. This regularizer is applied to a multi-task feature selection setting and an efficient algorithm is given to solve the related optimization problem. Empirical study shows that our proposed algorithm outperforms the baseline algorithms on benchmark datasets.

Acknowledgements

This work was supported in part by National Science Foundation (IIS-0643494) and National Institute of Health (1R01GM079688-01). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF and NIH.

References

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68:49–67, 2005.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288, 1996.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67:301–320, 2005.

Yuwon Kim, Jinseog Kim, and Yongdai Kim. Blockwise sparse regression. *Statistica Sinica*, 16:375–390, 2006.

Matthieu Kowalski, Marie Szafranski, and Liva Ralaivola. Multiple indefinite kernel learning with mixed norm regularization. In *ICML*, 2009.

Matthieu Kowalski. Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis*, 27(3): 303–324, 2009.

Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37:3468–3497, 2009.

Rich Caruana. Multi-task learning. *Machine Learning*, 28: 41–75, 1997.

Jonathan Baxter. A model for inductive bias learning. *J. Artificial Intelligence Research*, 12:149–198, 2000.

Shai Ben-david and Reba Schuller. Exploiting task relatedness for multiple task learning. In *COLT*, 2003.

R.K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

Theodoros Evgeniou, Charles A. Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *J. Mach. Learn. Res.*, 6:615–637, 2005.

Tony Jebara. Multi-task feature and kernel selection for SVMs. In *ICML*, 2004.

Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *CVPR*, 2004.

J. Chen, L. Tang, J. Liu, and J. Ye. A convex formulation for learning shared structures from multiple tasks. In *ICML*, 2009.

Bart Bakker and Tom Heskes. Task clustering and gating for Bayesian multitask learning. *J. Mach. Learn. Res.*, 4:83–99, 2003.

Tao Xiong, Jinbo Bi, Bharat Rao, and Vladimir Cherkassky. Probabilistic joint feature selection for multi-task learning. In *SDM*, 2007.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *NIPS*, 2006.

G. Obozinski, B. Taskar, and M. I. Jordan. Multi-task feature selection. In *ICML-06 Workshop on Structural Knowledge Transfer for Machine Learning*, 2006.

Su-In Lee, Vassil Chatalbashev, David Vickrey, and Daphne Koller. Learning a meta-level prior for feature relevance from multiple related tasks. In *ICML*, 2007.

Christopher Bishop. *Pattern recognition and machine learning*. Springer New York, 2006.

Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72, 2004.

Francis R. Bach, Gert R. G. Lanckriet, and Michael I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *ICML*, 2004.

Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *J. Mach. Learn. Res.*, 7:1531–1565, 2006.

Zenglin Xu, Rong Jin, Irwin King, and Michael R. Lyu. An extended level method for multiple kernel learning. In *NIPS*, 2008.

Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

N. Ueda and K. Saito. Parametric mixture models for multi-labeled text. In *NIPS*, 2003.

P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In *ICML*, 1998.

J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm support vector machines. In *NIPS*, page 49, 2004.

James P. Egan. *Signal detection theory and ROC-analysis*. Academic Press, 1975.