Adaptive Large Margin Training for Multilabel Classification

Yuhong Guo
Department of Computer & Information Sciences
Temple University
Philadelphia, PA 19122, USA
yuhong@temple.edu

Dale Schuurmans
Department of Computing Science
University of Alberta
Edmonton, AB T6G 2E8, Canada
dale@cs.ualberta.ca

Abstract

Multilabel classification is a central problem in many areas of data analysis, including text and multimedia categorization, where individual data objects need to be assigned multiple labels. A key challenge in these tasks is to learn a classifier that can properly exploit label correlations without requiring exponential enumeration of label subsets during training or testing. We investigate novel loss functions for multilabel training within a large margin framework—identifying a simple alternative that yields improved generalization while still allowing efficient training. We furthermore show how covariances between the label models can be learned simultaneously with the classification model itself, in a jointly convex formulation, without compromising scalability. The resulting combination yields state of the art accuracy in multilabel webpage classification.

Introduction

As machine learning begins to consider the analysis of more complex data types and subtle phenomena, issues such as multilabel classification are of growing importance. In many real world data analysis problems, data objects cannot be naturally categorized into a simple set of disjoint categories. Complex data objects, such as documents, webpages and videos, are complex and have multiple properties that entail multiple labels from overlapping classes. For example, in image labelling (Boutell et al. 2004; Zhou and Zhang 2006; Li et al. 2009) or video annotation (Qi et al. 2007), a given scene will usually exhibit numerous objects of interest. In text categorization, a given newswire article (Joachims 1998; McCallum 1999; Schapire and Singer 2000), webpage (Kazawa et al. 2004; Ueda and Saito 2002), legal document (Mencia and Fuernkranz 2008), or patent (Rousu et al. 2006; Cai and Hofmann 2007) will simultaneously belong to several categories of interest. Important problems in systems biology, such as gene function prediction (Clare and King 2001; Elisseeff and Weston 2001; Lanckriet et al. 2004; Blockeel et al. 2006; Zhang and Zhou 2006) can also be naturally cast as multilabel learning problems.

A key issue in multilabel learning is the structure of the loss function to be minimized during training. The earliest work on this problem adopted a simple reduction where a multilabel problem is expressed as a set of independent binary classification problems, one for each label (Joachims 1998). It was quickly realized, however, that such an approach was unsatisfactory (McCallum 1999), since the different labels occurring in a multilabel classification problem are not independent. On the contrary, they often exhibit strong correlations. Capturing these correlations in an effective yet tractable manner has led to significant research on formulations for multilabel learning. Another important issue, generic to all forms of machine learning, is regularization and prior knowledge, which we will see also play an important role in multilabel learning.

Our main goal in this paper is to report two simple ideas that appear to improve multilabel generalization performance without entailing significant computational overhead. The first idea addresses the problem of capturing dependence structure in the training loss in an efficient yet effective manner. The second idea exploits a recent advance from multi*task* learning to yield an adaptive form of regularization that automatically learns improved covariance structure between individual label models.

Multilabel Training Losses Much research on multilabel classification has focused on devising training losses that properly capture label correlations. We focus on *large margin* formulations in this paper. In the large margin approach the problem of multilabel classification can be most generally cast as a form of *structured output* classification (Tsochantaridis et al. 2005). That is, given an instance, the output can be any of the powerset of possible labels and arbitrary dependence structure between labels can be expressed in the margin loss. Of course, such an approach is impractical unless the label set is small. Making the powerset approach more practical has been a longstanding theme in multilabel classification research (Tsoumakas, Katakis, and Vlahavas 2009; Kazawa et al. 2004). Many authors have considered heuristic schemes to reduce the enumeration of label subsets, via ensemble (Tsoumakas and Vlahavas 2007) or clustering methods (Tenenboim, Rokach, and Shapira 2009) that implicitly introduce conditional independence.

Recently, many authors have begun to explicitly model conditional independence structure in the label set to achieve tractable classification and learning. For example, tree structured, or hierarchical dependencies are amenable to efficient

dynamic programming, and several authors have investigated multilabel learning over label hierarchies (Rousu et al. 2006; Cai and Hofmann 2007). These approaches, however, require that a label hierarchy be provided. Although it is possible to try to learn such a hierarchy as part of the training process, such methods have remained heuristic.

By far the most popular approach to simplifying the dependence structure between labels, without eliminating dependence entirely, is to assume only *pairwise* dependence. Many probabilistic approaches have adopted models that express pairwise dependence over labels (Zhu et al. 2005; Ghamrawi and McCallum 2005), but unfortunately these generally lead to hard classification problems: a pairwise field over labels for a given instance still generally entails an intractable search to recover an optimal label set (although efficient algorithms can be found in some special cases (Kang, Jin, and Sukthankar 2006)). A simpler version of this idea has proved to be somewhat more practical: learn a simple ranking function over labels by training on label pairs (Elisseeff and Weston 2001; Schapire and Singer 2000; Zhang and Zhou 2006; Fuernkranz et al. 2008).

In this paper, rather than consider a pairwise dependence structure, we investigate a simpler notion of *separation dependence*. That is, rather than require "all-pairs" training of a ranking function, we train a ranking function merely to separate the correct from incorrect labels for each instance. The novel form of large margin loss we employ is a modification of the topic-ranking loss of (Crammer and Singer 2003). We show that this simpler dependence structure enables efficient and exact classification in a multilabel setting. Surprisingly, we find that the classification performance is generally superior to that of all-pairs training.

Regularization Independent of the training loss, regularization is always another key aspect of effective training. It has been observed in (Hariharan et al. 2010) that prior knowledge about label covariances can be exploited to achieve state of the art generalization performance, even when using the simple independent binary loss. Others have observed that estimating label covariances as a postprocessing step is also helpful (Boutell et al. 2004; Barutcuoglu, Schapire, and Troyanskaya 2006). We prove that the prior knowledge formulation of (Hariharan et al. 2010) is equivalent to introducing an inverse covariance matrix in the regularizer. Although such covariance information will not always be available, we show that an *optimal* label covariance matrix can be learned simultaneously with the classification model, by expressing the problem in a jointly convex form. Importantly, the resulting training formulation does not entail significant computational overhead.

There are many other approaches to multilabel training that are beyond the scope of this paper. We are not able to discuss the numerous works that have considered nearest neighbor approaches (Zhang and Zhou 2005; 2007; Jin, Wang, and Zhou 2009; Cheng and Huellermeier 2009), or subspace learning approaches (Yu, Yu, and Tresp 2005; Yan, Tesic, and Smith 2007; Park and Lee 2008; Zhang and Zhou 2008; Hsu et al. 2009; Ji et al. 2010). (We note that the approaches we investigate in this paper are directly compatible with subspace learning however.)

Loss Functions for Multilabel Classification

We assume a $t \times n$ matrix of input observations X and a $t \times L$ matrix of boolean label indicators Y are provided, where $Y_{il} = 1$ when instance i has label l, and $Y_{il} = 0$ otherwise. We will find it convenient to work with the ± 1 label indicator matrix $Z = 2Y - 1$, $Y = (Z+1)/2$, at times. For simplicity, we will also assume a linear model for classification that uses an N dimensional feature map $\boldsymbol{\phi}(\cdot)$ over input observations and is parameterized by an $N \times L$ weight matrix W. (This will be reexpressed in a kernel representation later.) Given an input instance $\mathbf{x}$, an L dimensional response vector $\mathbf{s}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})'W$ is recovered using W, giving a "score" for each label. These scores will be compared to a threshold to determine which labels are to be predicted.

We focus on large margin formulations of multilabel training. Large margin training, as in all standard forms of loss minimization, consists of minimizing a loss function over the data subject to a regularizer. That is, given X and Y, a generic training problem can be expressed as

$$\min_W \tfrac{\beta}{2}\|W\|_F^2 + \textstyle\sum_i \ell(Y_{i:}; \mathbf{s}(X_{i:})) \tag{1}$$

where $\|\cdot\|_F$ denotes Frobenius norm, $\ell(\cdot;\cdot)$ is a loss function that is convex in its second argument, $\beta > 0$ is a regularization parameter that controls the magnitude of weights in W, and $\mathbf{s}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})'W$ is the parameterized score function.

Multilabel classification losses are typically contrasted with multi*class* and *binary* classification losses respectively. For multiclass classification, we assume that any training label vector $Y_{i:}$ has only a single 1 entry, which we can indicate by a label l_i such that $Y_{i:} = \mathbf{1}_{l_i}$ (a vector of all 0s except for a 1 in the l_ith position). The standard large margin multiclass loss (Crammer and Singer 2001) is

$$\max_l 1_{(l \neq l_i)} + s_l(X_{i:}) - s_{l_i}(X_{i:}) \tag{2}$$

That is, we want to promote the score $s_{l_i}(X_{i:})$ of the correct label l_i above the score $s_l(X_{i:})$ of any alternative label l for $l \neq l_i$ by a margin of 1.

In the binary classification case, we do not work with a $N \times 2$ model W but an $N \times 1$ weight vector $\mathbf{w}$, yielding a score value that is a *scalar* not a vector. Here one also works with a $t \times 1$ vector of ± 1 valued training labels $\mathbf{z}$ rather than a $t \times 2$ matrix Y. In this case, the standard SVM loss is

$$(1 - z_i s(X_{i:}))_+ \tag{3}$$

where $(1 - z_i s(X_{i:}))_+ = \max(0, 1 - z_i s(X_{i:}))$.

The original large margin approach to multilabel classification (Joachims 1998) decomposed the training loss into a sum over independent binary training losses for each label

$$\sum_l (1 - Z_{il} s_l(X_{i:}))_+ \tag{4}$$

an approach we refer to as *independent binary* training. Given a test example $\mathbf{x}$, its labelling is determined by $z_l^* = \arg\max_{z_l \in \pm 1} z_l s_l(\mathbf{x})$. Most work since has expressed dissatisfaction with treating the labels as independent, and investigated losses that account for correlations among labels.

As mentioned in the introduction, the most general approach to multilabel classification in the large margin setting

is structured output prediction (Tsochantaridis et al. 2005) that can account for arbitrary label interactions:

$$\max_{\mathbf{y}\in\{0,1\}^L} \Delta(\mathbf{y}; Y_{i:}) + \sum_l (y_l - Y_{il}) s_l(X_{i:}) \quad (5)$$

for some misclassification loss $\Delta(\mathbf{y}; Y_{i:})$. Note that as given, (5) is not easy to compute since it involves a search over 2^L alternative labellings. Nevertheless, this loss can express arbitrary relationships between the behavior of individual labels. Note that if one chooses $\Delta(\mathbf{y}; Y_{i:}) = \sum_l 1_{(y_l \neq Y_{il})}$ then (5) becomes equal to (4), hence structured output training reduces to independent binary training.[1] However, if instead one chooses $\Delta(\mathbf{y}; Y_{i:}) = 1_{(\mathbf{y} \neq Y_{i:})}$, yielding a loss of 0 only if the two vectors are identical across all labels, then a large margin form of "label powerset" training is achieved (Tsoumakas, Katakis, and Vlahavas 2009; Tsoumakas and Katakis 2007).

The best known attempt to capture label correlations while avoiding exponential enumeration is the *pairwise ranking loss* introduced by (Elisseeff and Weston 2001)

$$\frac{1}{|Y_{i:}|} \frac{1}{|\bar{Y}_{i:}|} \sum_{l \in Y_{i:}} \sum_{\bar{l} \in \bar{Y}_{i:}} (1 + s_{\bar{l}}(X_{i:}) - s_l(X_{i:}))_+ \quad (6)$$

where $|Y_{i:}|$ is the number of 1s in $Y_{i:}$ and $\bar{Y}_{i:} = 1 - Y_{i:}$. Although (6) clearly accounts for pairwise correlations during training, classification is difficult with this approach because given a test example $\mathbf{x}$ the resulting vector of scores $\mathbf{s}(\mathbf{x})$ can only be interpreted as a ranking over labels; that is, the cutoff threshold for asserting the presence of a label is not determined by this method. (Elisseeff and Weston (2001) explore a post-hoc heuristic for determining the threshold.) Thus, this approach is usually referred to as a multilabel ranking rather than a multilabel classification method.

A variant of this loss is the *separation ranking loss*

$$\max_{l \in Y_{i:}} \max_{\bar{l} \in \bar{Y}_{i:}} (1 + s_{\bar{l}}(X_{i:}) - s_l(X_{i:}))_+ \quad (7)$$

proposed by (Crammer and Singer 2003). As for the pairwise ranking loss the classification threshold is not determined, and again, the resulting approach is considered a multilabel ranking rather than classification method.

Recently, Fuernkranz et al. (2008) observed that the main weakness with ranking losses, such as those above, is that they are "uncalibrated". In these cases, effective multilabel classification can be recovered by introducing a *reference label* against which all other labels are compared. Effectively, one introduces a new dummy label l_0 and uses the corresponding dummy score $s_0(X_{i:})$ as a reference of comparison between the relevant and irrelevant labels. For example, using this idea (6) can be reformulated as

$$\begin{aligned} &\tfrac{1}{|Y_{i:}|} \textstyle\sum_{l \in Y_{i:}} (1 + s_0(X_{i:}) - s_l(X_{i:}))_+ \\ &+ \tfrac{1}{|\bar{Y}_{i:}|} \textstyle\sum_{\bar{l} \in \bar{Y}_{i:}} (1 + s_{\bar{l}}(X_{i:}) - s_0(X_{i:}))_+ \end{aligned} \quad (8)$$

[1] Proved using the substitution $Y_{i:} = (Z_{i:} - 1)/2$ and $\mathbf{y} = (\mathbf{z} - 1)/2$ and noting that $\sum_l \max_{z_l \in \pm 1} 1_{(z_{\bar{l}} \neq Z_{il})} + (z_l - Z_{il}) s_l(X_{i:})/2 = \sum_l (1 - Z_{il} s_l(X_{i:}))_+$.

The choice of reference s_0 can be left adaptive or fixed to some reference value, such as $s_0(X_{i:}) = 0$. Unfortunately, it can be shown, using the transformation $(Z-1)/2 = Y$ and $s_0(X_{i:}) = 0$ the loss (8) reduces to a form of independent binary training.[2] Consequently, we have not found this loss to be effective in practice.

However, an interesting and effective new loss is obtained by applying the calibration idea to (7). In particular, we proposed the following *calibrated separation ranking loss* for multilabel classification

$$\max_{l \in Y_{i:}} (1 + s_0(X_{i:}) - s_l(X_{i:}))_+ + \max_{\bar{l} \in \bar{Y}_{i:}} (1 + s_{\bar{l}}(X_{i:}) - s_0(X_{i:}))_+ \quad (9)$$

Given a test example $\mathbf{x}$ its classification is determined by $y_l^* = \arg\max_{y_l \in \{0,1\}} y_l (s_l(\mathbf{x}) - s_0(\mathbf{x}))$. To the best of our knowledge, the loss (9) has not been previously investigated in the literature, yet we find that it gives superior results to the standard margin losses in our experiments below.

Calibrated Separation Ranking: Dual Form

Combining (1) with (9) yields the regularized loss minimization problem we consider for multilabel classification training. Let $\Phi_{i:} = \boldsymbol{\phi}(X_{i:})$, so that the feature representations of the training instances are expressed in a $t \times N$ matrix Φ. Given that the loss (9) is piecewise linear and convex, this problem can be equivalently expressed by a convex quadratic program.

$$\begin{aligned} \min_{W,\mathbf{u}} \tfrac{\beta}{2}(\|W\|_F^2 + \|\mathbf{u}\|_2^2) + \mathbf{1}'\boldsymbol{\xi} + \mathbf{1}'\boldsymbol{\eta} \quad &\text{subject to} \\ \xi_i \geq 1 + \Phi_{i:}\mathbf{u} - \Phi_{i:}W_{:l} \text{ for } l \in Y_{i:}, \forall i \\ \eta_i \geq 1 - \Phi_{i:}\mathbf{u} + \Phi_{i:}W_{:\bar{l}} \text{ for } \bar{l} \in \bar{Y}_{i:}, \forall i \\ \boldsymbol{\xi} \geq 0, \boldsymbol{\eta} \geq 0 \end{aligned} \quad (10)$$

where $\mathbf{1}$ denotes a vector of all 1s. Note, if we wish to enforce a constant dummy score s_0 so that $s_0(X_{i:}) = 0$ we can fix $\mathbf{u} = 0$ in (10).

The dual of this quadratic program is

$$\begin{aligned} \max_{M,N} \mathbf{1}'(M+N)\mathbf{1} - \tfrac{1}{2\beta}\text{tr}((M-N)'K(M-N)\Theta) \quad &\text{s.t.} \\ M \geq 0,\ M\mathbf{1} \leq \mathbf{1},\ M_{i\bar{l}} = 0 \text{ for } \bar{l} \in \bar{Y}_{i:} \\ N \geq 0,\ N\mathbf{1} \leq \mathbf{1},\ N_{i\bar{l}} = 0 \text{ for } l \in Y_{i:} \end{aligned} \quad (11)$$

where $K = \Phi\Phi'$, M and N are both $t \times L$ dual parameter matrices, and Θ is a constant matrix $\Theta = I + \mathbf{1}\mathbf{1}'$. Thus the problem is now expressed entirely with respect to a kernel. In our experiments below, we use a more compact version of this quadratic program, using the fact that M and N are nonzero on complementary entries. That is, by letting $A = M - N$ one obtains the more concise form

$$\begin{aligned} \max_A \text{tr}(A'Z) - \tfrac{1}{2\beta}\text{tr}(A'KA\Theta) \quad &\text{subject to} \\ A \circ Y \geq 0,\ \text{diag}(A'Y) \leq \mathbf{1} \\ A \circ (Y-1) \geq 0,\ \text{diag}(A'(Y-1)) \leq \mathbf{1} \end{aligned} \quad (12)$$

where $\circ$ denotes componentwise multiplication, and diag denotes the main diagonal vector of a square matrix.

[2] In particular, by using $s_0(X_{i:}) = 0$, (8) becomes equal to $\frac{1}{|Y_{i:}|}\sum_{l \in Y_{i:}} (1 - Z_{il} s_l(X_{i:}))_+ + \frac{1}{|\bar{Y}_{i:}|}\sum_{\bar{l} \in \bar{Y}_{i:}} (1 - Z_{i\bar{l}} s_{\bar{l}}(X_{i:}))_+$, which is essentially independent binary with some normalization.

Algorithmic Approach

Unfortunately, a drawback with (12) is still that it requires space quadratic in the number of training instances t simply to store the kernel matrix K. This is not practical in a real multilabel learning scenario, since multilabel classification problems tend to be large (Hariharan et al. 2010; Lewis et al. 2004). To avoid the onerous space requirements of (12) we formulate a scalable approach for tackling the problem in the spirit of (Hariharan et al. 2010; Fan et al. 2008).

Note that the constraints in (12) decompose over training examples and labels. This allows many different strategies for incrementally solving the problem by tackling nearly independent components of A. Note that the entries of A corresponding to different training instances and labels are only coupled in the quadratic objective

$$f(A) \quad = \quad \text{tr}(A'Z) - \tfrac{1}{2\beta}\text{tr}(A'KA\Theta) \qquad (13)$$

We have found it effective to update rows of A across labels, which still leads to an efficient update. Consider an update to row $A_{i:}$, which can be expressed as $A \leftarrow A + \mathbf{1}_i(\mathbf{a} - \mathbf{a}_0)'$; thus we are replacing $\mathbf{a}_0' = A_{i:}$ with a new row $\mathbf{a}'$. Consider the effect of this update on the objective

$$f(A + \mathbf{1}_i(\mathbf{a} - \mathbf{a}_0)') = Z_{i:}\mathbf{a} - \tfrac{1}{\beta}\mathbf{a}'\Theta A'K_{:i} + \tfrac{1}{\beta}\mathbf{a}'\Theta\mathbf{a}_0K_{ii} - \tfrac{1}{2\beta}\mathbf{a}'\Theta\mathbf{a}K_{ii} + const \qquad (14)$$

Note that the update only requires computation of $K_{i:}$ for a single row in K, hence the cost is only linear in t for each update. Also, the update to $A_{j:}$ does not affect the update for another row $A_{i:}$ except through the term $\mathbf{a}'\Theta A'K_{:i}$. Thus, all one needs to do is maintain a $t \times L$ matrix $C = KA\Theta$, which can be updated locally after an update to $A_{i:}$ by $C \leftarrow C + K_{:i}(\mathbf{a} - \mathbf{a}_0)'\Theta$. Otherwise, the updates are independent.

There are many strategies to choose which rows to update and, when selected, what value to update to. We have found it effective simply to visit the training instances in a random order, and locally minimize $f(A)$ with respect to row $A_{i:}$ under the local constraints in (12). This computational approach allows us to solve the quadratic program for large problems in linear space.

Adaptive Regularization: Label Covariance

The work in (Hariharan et al. 2010) provides an interesting extension to the standard independent binary approach, where an invertible $L \times L$ matrix P is introduced to express prior knowledge about the covariance structure of the linear models learned for each label. Despite minimizing a loss as simple as the independent binary loss, Hariharan et al. (2010) observe notable generalization improvements simply by incorporating prior knowledge about the covariance structure. In particular, the score function $\mathbf{s}(\mathbf{x})$ is modified to take P into account: $\mathbf{s}_P(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})'WP$. In this way, the response vector over labels, given $\mathbf{x}$, takes into account prior covariance structure stipulated by P. To understand the effect of P more clearly, observe that

$$\min_{W,\mathbf{u}} \tfrac{\beta}{2}(\|W\|_F^2 + \|\mathbf{u}\|_2^2) + \sum_i \ell(Y_{i:}; \boldsymbol{\phi}(X_{i:})WP) \qquad (15)$$

$$= \min_{U,\mathbf{u}} \tfrac{\beta}{2}(\text{tr}(UR^{-1}U') + \mathbf{u}'\mathbf{u}) + \sum_i \ell(Y_{i:}; \boldsymbol{\phi}(X_{i:})U) \qquad (16)$$

where $R = P'P$. Thus, P influences the training problem by altering the regularization with respect to the inverse covariance $R^{-1} = (P'P)^{-1}$. This is precisely the form of covariance regularization used in current multi*task* learning formulations (Zhang and Yeung 2010) to enforce task relatedness; see also (Argyriou et al. 2008). So in this way (Hariharan et al. 2010) are implicitly imposing the same form of regularization on multilabel learning. One advantage of the alternative formulation (16) is that it is jointly *convex* in both the model parameters U and $\mathbf{u}$ *and* in the covariance matrix R (Boyd and Vandenberghe 2004).

This allows optimal training of the covariance matrix simultaneously with the classification model; yielding an adaptive form of training that appears to be novel in the multilabel learning literature. In particular, we can reexpress the previous quadratic programming formulation (12) as a convex-concave program that does not have local minima.

$$\min_R \max_A \text{tr}(A'Z) - \tfrac{1}{2\beta}\text{tr}(A'KA\Gamma R\Gamma') + \tfrac{\alpha}{2}\|R - I\|_F^2 \quad \text{s.t.}$$
$$R \succeq 0,\ A \circ Y \geq 0,\ \text{diag}(A'Y) \leq \mathbf{1}$$
$$A \circ (Y - 1) \geq 0,\ \text{diag}(A'(Y - 1)) \leq \mathbf{1} \qquad (17)$$

where Γ is a constant matrix $\Gamma = [I, -\mathbf{1}]$ such that $\Gamma\Gamma' = \Theta$. Note that the $\|R - I\|_F^2$ term regularizes the learned covariance matrix, keeping it proximal to the identity matrix to avoid overfitting, while still allowing it to reduce the training loss of the multilabel classification model.

This form of covariance learning can be trivially added to the scalable computational procedure outlined previously. In our implementation, we alternate between updating A in a complete pass over the data, then after each pass minimizing over R optimally given A by the closed form update $R \leftarrow I + \frac{1}{2\alpha\beta}\Gamma'A'KA\Gamma$ which can be computed in linear space. Surprisingly, we find that learning R *accelerates* the convergence of the previous update procedure for A, and we witness both run time and generalization improvements in our experiments below consequently.

Experimental Results

To evaluate the proposed loss and regularization scheme, we conducted experiments on multi-topic web page classification (Ueda and Saito 2002). The data consists of web pages collected from the yahoo.com domain. We preprocessed the data by first removing the largest class label (which covered more than 50% of the instances) and removing class labels that had fewer than 200 instances. We also removed any instances that had no labels or every label. For the instance feature representation, we removed any features that appeared in fewer than 5 instances, and converted the remaining integer features into a standard *tf-idf* encoding. The statistics of the preprocessed data sets are summarized in Table 2.

We compared two versions of our method, CSRL (for *calibrated separation ranking loss*) using an identity covariance matrix, and CSRL+R, which added the convex covariance learning scheme introduced above. We compared our techniques to three other baseline large margin methods for multi-label classification: (1) the standard independent binary SVM, which trains a separate binary classifier

Table 1: Training time of the comparison methods (seconds).

	CSRL	CSRL-R
Arts	1195.0	701.9
Computers	1592.2	687.6
Education	718.8	544.3
Entertain.	1020.0	497.4
Health	1196.2	686.9
Recreation	1369.0	1082.8
Reference	1097.0	660.3
Science	703.6	416.2
Social	796.8	505.4
Society	1197.7	546.4

for each label; (2) the pairwise ranking loss SVM proposed in (Elisseeff and Weston 2001), which first trains a large margin ranking model and then learns the threshold of the multilabel predications using a least-square method; and (3) the large scale max-margin multi-label classification method (M3L) proposed in (Hariharan et al. 2010), which takes prior knowledge about the label correlations into account. In all these experiments, we simply set the regularization parameters $\beta = 1$ and $\alpha = 10$ respectively. The performance of each method is evaluated using three measures: exact match ratio, macro-F1, and micro-F1 (Tang, Rajan, and Narayanan 2009).

We randomly selected 800 instances from each data set for training, and used the remaining data points for testing. This process is repeated five times to generate five random training/test partitions. The average performance and standard deviations of the 5 methods for each of the three evaluation measures is reported in Table 3.

One can see from these results that the novel loss employed by CSRL provides a significant improvement in generalization accuracy over the baseline methods, in all of the data sets considered. Notably, the inclusion of covariance learning in CSRL+R actually *improves* the convergence properties of the iterative training algorithm for solving the convex (convex-concave) program; see Table 1.

Conclusion

We have investigated new large margin loss functions and new adaptive regularization schemes for multilabel classification learning. The new loss we investigate, *calibrated separation ranking*, effectively takes into account label correlations without reducing to independent binary classification under calibration (unlike the pairwise ranking (Elisseeff and Weston 2001) under calibration (Fuernkranz et al. 2008)). We observe significant improvements over baseline training losses, independent binary and pairwise ranking. Furthermore, we show how the method for incorporating prior knowledge proposed by (Hariharan et al. 2010) is equivalent to an inverse covariance regularization of the multilabel model. We exploit this observation to formulate a convex training principle for adapting both the label covariance (regularization) matrix and the classification model, in an efficient joint training scheme that entails minimal overhead over standard training. Experimental results indicate that both the alternative loss function and adaptive regularization scheme yield improved generalization performance in multilabel classification.

References

Argyriou, A., Evgeniou, T. and Pontil, M. 2008. Convex multi-task feature learning. *Machine Learning* 73(3):243–272.

Barutcuoglu, Z.; Schapire, R.; and Troyanskaya, O. 2006. Hierarchical multi-label prediction of gene function. *Bioinformatics* 22:880–836.

Blockeel, H.; Schietgat, L.; Struyf, J.; Dzeroski, S.; and Clare, A. 2006. Decision trees for hierarchical multilabel classification: A case study in functional genomics. In *Proceedings PKDD*, 18–29.

Boutell, M.; Luo, J.; Shen, X.; and Brown, C. 2004. Learning multi-label scene classiffication. *Patt. Recogn.* 37(9):1757–1771.

Boyd, S., and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge U. Press.

Cai, L., and Hofmann, T. 2007. Exploiting known taxonomies in learning overlapping concepts. In *Proceedings IJCAI*, 714–719.

Cheng, W., and Huellermeier, E. 2009. Combining instance-based learning and logistic regression for multilabel classification. In *Proceedings ECML-PKDD*, 28–38.

Clare, A., and King, R. 2001. Knowledge discovery in multi-label phenotype data. In *Proceedings PKDD*, 42–53.

Crammer, K., and Singer, Y. 2001. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR* 265–292.

Crammer, K., and Singer, Y. 2003. A family of additive online algorithms for category ranking. *JMLR* 3:1025–1058.

Elisseeff, A., and Weston, J. 2001. A kernel method for multilabelled classification. In *Proceedings NIPS*.

Fan, R.; Hsieh, K. C. C.; Wang, X.; and Lin, C. 2008. LIBLINEAR: A library for large linear classification. *JMLR* 9:1871–1874.

Fuernkranz, J.; Huellermeier, E.; Mencia, E.; and Brinker, K. 2008. Multilabel classification via calibrated label ranking. *Machine Learning* 73(2).

Ghamrawi, N., and McCallum, A. 2005. Collective multi-label classification. In *Proceedings CIKM*.

Hariharan, B.; Zelnik-Manor, L.; Vishwanathan, S.; and Varma, M. 2010. Large scale max-margin multi-label classification with priors. In *Proceedings ICML*.

Hsu, D.; Kakade, S.; Langford, J.; and Zhang, T. 2009. Multi-label prediction via compressed sensing. In *Proceedings NIPS*.

Ji, S.; Tang, L.; Yu, S.; and Ye, J. 2010. A shared-subspace learning framework for multi-label classification. *ACM Transactions on Knowledge Discovery in Data* 4(2):1–29.

Jin, R.; Wang, S.; and Zhou, Z. 2009. Learning a distance metric from multi-instance multi-label data. In *Proc. CVPR*, 896–902.

Joachims, T. 1998. Text categorization with support vector machines: learn with many relevant features. In *Proceedings ECML*.

Kang, F.; Jin, R.; and Sukthankar, R. 2006. Correlated label propagation with application to multi-label learning. In *Proceedings CVPR*, 1719–1726.

Kazawa, H.; Izumitani, T.; Taira, H.; and Maeda, E. 2004. Maximal margin labeling for multi-topic text categorization. In *Proc. NIPS*.

Lanckriet, G.; Deng, M.; Cristianini, N.; Jordan, M.; and Noble, W. 2004. Kernel-based data fusion and its application to protein function prediction in yeast. In *Proc. Pacific Sympo. on Biocomput.*

Table 2: Statistics of the multi-label data sets: k–the number of classes, d–the number of features, T–the number of instances.

Params	Arts	Computers	Education	Entertain.	Health	Recreation	Reference	Science	Social	Society
k	15	19	11	13	13	14	11	15	11	18
d	12011	14026	11320	16506	13909	15007	8341	11942	13651	22973
T	3298	3716	3920	3512	4425	3881	1132	1865	2426	5824

Table 3: Summary of the performance (%) for the compared methods in terms of exact match ratio (top section), macro F1 (middle section), and micro F1 (bottom section).

Methods	Arts	Computers	Education	Entertain.	Health	Recreation	Reference	Science	Social	Society
SVM	0±0.0	0±0.0	0.1±0.1	3.8±0.1	0±0.0	0±0.0	13.8±0.2	0±0.0	0±0.0	2.2±0.0
RankPair	3.4±0.1	0.3±0.1	1.9±0.4	3.6±2.5	1.1±0.5	0.3±0.2	10.7±2.7	0.6±0.2	3.8±1.5	1.1±0.5
M3L	2.6±0.5	12.1±0.3	13.2±0.6	4.7±2.4	7.4±1.6	0.0±0.0	13.9±0.5	8.4±1.6	16.5±1.0	2.2±0.0
CSRL	8.0±0.3	12.4±0.3	17.9±0.3	16.3±0.5	17.1±0.1	11.4±0.6	16.2±0.6	10.8±0.3	17.9±0.6	2.2±0.1
CSRL+R	8.2±0.3	13.5±0.3	18.0±0.4	17.3±0.4	17.9±0.1	11.6±0.6	16.4±0.8	11.2±0.4	18.1±0.6	2.2±0.1
SVM	5.5±0.0	4.0±0.0	8.1±0.1	12.5±0.1	0±0.0	6.6±0.0	16.0±0.0	5.4±0.0	8.2±0.0	8.7±0.0
RankPair	10.2±1.8	6.1±0.2	23.9±1.4	23.2±0.5	13.8±1.5	13.7±0.3	16.0±1.9	11.2±1.9	15.9±1.2	11.4±0.8
M3L	9.6±0.2	6.8±0.6	26.0±0.5	13.0±0.1	19.5±1.0	6.6±0.0	16.5±0.2	14.2±1.2	26.7±0.8	8.7±0.0
CSRL	41.4±0.2	40.7±0.3	49.5±0.3	57.5±0.5	55.3±0.2	45.2±0.5	47.8±0.4	46.4±0.5	49.5±0.5	37.2±0.1
CSRL+R	41.5±0.2	40.7±0.3	49.2±0.3	57.4±0.5	55.3±0.3	45.1±0.5	47.9±0.4	46.3±0.5	49.3±0.5	37.3±0.1
SVM	34.5±0.1	31.6±0.1	39.2±0.1	53.2±0.1	0±0.0	43.9±0.1	67.4±0.3	34.4±0.3	46.7±0.1	48.3±0.1
RankPair	42.1±0.1	33.7±0.3	52.4±1.3	61.1±0.3	39.2±5.3	50.7±1.2	61.0±5.8	40.3±2.2	53.7±1.4	45.9±1.4
M3L	42.1±0.4	36.5±1.1	54.9±0.4	53.6±0.1	47.7±0.4	44.0±0.1	67.5±0.3	47.1±1.1	62.7±0.5	48.3±0.1
CSRL	52.3±0.1	50.0±0.1	64.1±0.2	68.3±0.2	61.7±0.3	60.1±0.3	68.7±0.1	56.9±0.3	66.1±0.2	50.7±0.1
CSRL+R	52.3±0.1	50.2±0.2	64.0±0.2	68.3±0.2	61.8±0.3	60.2±0.3	68.8±0.1	57.0±0.3	66.1±0.2	50.8±0.1

Lewis, D.; Yang, Y.; Rose, T.; and Li, F. 2004. RCV1: A new benchmark collection for text categorization research. *JMLR* 5:361–397.

Li, Y.; Ji, S.; Ye, J.; Kumar, S.; and Zhou, Z. 2009. Drosophila gene expression pattern annotation through multi-instance multi-label learning. In *Proceedings IJCAI*, 1445–1450.

McCallum, A. 1999. Multi-label text classification with a mixture model trained by em. In *AAAI Workshop on Text Learning*.

Mencia, E., and Fuernkranz, J. 2008. Efficient pairwise multilabel classification for large scale problems in the legal domain. In *Proceedings PKDD*.

Park, C., and Lee, M. 2008. On applying linear discriminant analysis for multi-labeled problems. *Pattern Recognition Letters* 29:878–887.

Qi, G.; Hua, X.; Rui, Y.; Tang, J.; Mei, T.; and Zhang, H. 2007. Correlative multi-label video annotation. In *Proc. Multimedia*.

Rousu, J.; Saunders, C.; Szedmak, S.; and Shawe-Taylor, J. 2006. Kernel-based learning of hierarchical multilabel classification models. *JMLR* 7:1601–1626.

Schapire, R., and Singer, Y. 2000. Boostexter: A boosting-based system for text categorization. *Machine Learning* 39:135–168.

Tang, L.; Rajan, S.; and Narayanan, V. 2009. Large scale multi-label classification via metalabeler. In *Proceedings WWW*.

Tenenboim, L.; Rokach, L.; and Shapira, B. 2009. Multi-label classification by analyzing labels dependencies. In *International Workshop on Learning from Multi-Label Data*.

Tsochantaridis, I.; Hofmann, T.; Joachims, T.; and Altun, Y. 2005. Large margin methods for structured and interdependent output variables. *JMLR* 6:1453–1484.

Tsoumakas, G., and Katakis, I. 2007. Multi-label classification an overview. *International Journal of Data Warehousing and Mining*.

Tsoumakas, G., and Vlahavas, I. 2007. Random k-labelsets: An ensemble method for multilabel classification. In *Proc. ECML*.

Tsoumakas, G.; Katakis, I.; and Vlahavas, I. 2009. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook, 2nd edition. Springer*.

Ueda, N., and Saito, K. 2002. Parametric mixture models for multi-labeled text. In *Proceedings NIPS 15*, 721–728.

Yan, R.; Tesic, J.; and Smith, J. 2007. Model-shared subspace boosting for multi-label classification. In *Proceedings SIGKDD*.

Yu, K.; Yu, S.; and Tresp, V. 2005. Multi-label informed latent semantic indexing. In *Proceedings SIGIR*.

Zhang, Y., and Yeung, D. 2010. A convex formulation for learning task relationships in multi-task learning. In *Proceedings UAI*.

Zhang, M., and Zhou, Z. 2005. A k-nearest neighbor based algorithm for multi-label classification. In *IEEE International Conference on Granular Computing*, 718–721.

Zhang, M., and Zhou, Z. 2006. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Trans. Knowledge and Data Eng.* 18(10):1338–1351.

Zhang, M., and Zhou, Z. 2007. Multi-label learning by instance differentiation. In *Proceedings AAAI*, 669–674.

Zhang, M., and Zhou, Z. 2008. Multi-label dimensionality reduction via dependency maximization. In *Proc. AAAI*, 1503–1505.

Zhou, Z., and Zhang, M. 2006. Multi-instance multi-label learning with application to scene classification. In *Proceedings NIPS 19*.

Zhu, S.; Ji, X.; Xu, W.; and Gong., Y. 2005. Multi-labelled classification using maximum entropy method. In *Proceedings SIGIR*.