

A Mathematical Programming Formulation for Sparse Collaborative Computer Aided Diagnosis

Jinbo Bi

CAD & Knowledge Solutions Group
Siemens Medical Solutions Inc.
Malvern, PA 19355
jinbo.bi@siemens.com

Tao Xiong

Department of Electrical and Computer Engineering
University of Minnesota
Twin Cities, MN 55414
txiong@ece.umn.edu

Abstract

A mathematical programming formulation is proposed to eliminate irrelevant and redundant features for collaborative computer aided diagnosis which requires to detect multiple clinically-related malignant structures from medical images. A probabilistic interpretation is described to justify our formulations. The proposed formulation is optimized through an effective alternating optimization algorithm that is easy to implement and relatively fast to solve. This collaborative prediction approach has been implemented and validated on the automatic detection of solid lung nodules by jointly detecting ground glass opacities.

Introduction

Over the last decade, computer-aided diagnosis (CAD) systems have moved from the sole realm of academic publications, to robust commercial systems that are used by physicians in their clinical practice (Roehrig 1999; Buchbinder *et al.* 2004). In many CAD applications, the goal is to detect potentially malignant tumors and lesions in medical images. It is well recognized that the CAD system decreases detection and recognition errors as a second reader and reduces mistakes related to misinterpretation (Armato-III, Giger, & MacMahon 2001; Naidich, Ko, & Stoeckel 2004). However, most CAD systems focus on the diagnosis of a single isolated disease using images taken only for the specific disease. It neglects certain fundamental aspects of physicians diagnosis procedure where physicians examine primary symptoms and tests of the disease in conjunction with other related information, such as symptoms of clinically-related conditions, patient history of other diseases and medical knowledge of highly correlated diseases.

For instance, lung cancer is the leading cause of cancer-related death in western countries with a better survival rate for early diagnosis. An automated CAD system can be built to identify solid nodules or ground glass opacities (GGOs). A patient who has solid nodules can also have GGOs, whereas a patient who has GGOs can later develop calcified GGOs which become solid or partly-solid nodules. Radiologic classification of small adenocarcinoma of

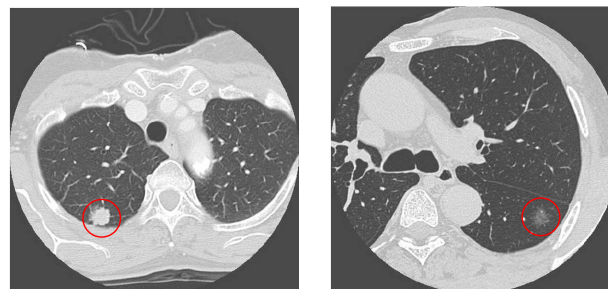


Figure 1: lung CT images: left – Nodule; right – GGO.

lung by means of thoracic thin-section CT discriminates between the GGOs and solid nodules. The solid nodule is defined as an area of increased opacification more than 5mm in diameter, which completely obscures underlying vascular markings. Ground-glass opacity (GGO) is defined as an area of a slight, homogeneous increase in density, which does not obscure underlying vascular markings (Suzuki *et al.* 2006). Figure 1 shows examples of a solid nodule and a GGO. The two detection systems are often constructed independently. Detecting nodules and detecting GGOs are two closely dependent tasks whereas each also has its own respective characteristics, which makes joint learning beneficial when building a specific model for each task for better predictive capacity.

Hence, we introduce a novel concept – “collaborative” computer aided diagnosis – that aims to improve the diagnosis of a single malignant structure by learning the detection process of multiple related abnormal structures from medical images simultaneously. It takes advantage of the opportunity to compare and contrast similar medical conditions in learning to diagnose patients in terms of disease categories.

The collaborative learning problem is often, in machine learning areas, cast as multi-task learning, collaborative filtering or collaborative prediction problems, depending on various applications. Multi-task learning is able to capture the dependencies among tasks when several “related” learning problems are available. The key is how to define task relatedness among tasks. In (Ando & Zhang 2005) a common hidden structure for all related tasks is assumed. One natural way to capture the task relatedness is through hier-

archical Bayesian models (Heskes 2000). From the hierarchical Bayesian viewpoint, multi-task learning is essentially trying to learn a good prior over all tasks to capture task dependencies (Caruana 1997; Evgeniou & Pontil 2004).

To tackle a CAD task, researchers often deploy a large amount of experimental features to describe the potential cancerous structures or abnormal structures. It consequently and inevitably introduces irrelevant features or redundant features to the detection or classification problems. Feature selection has been an indispensable and challenging problem in this domain. Moreover, researchers often face a situation where multiple tasks that are related from the physical and medical perspectives are given with a very limited sample size for each. Acquisition of medical data is expensive and time-consuming. For example, in the nodule and GGO detection tasks, often only around 100 patients are available. Commonly, the same set of features are evaluated for candidates of nodules and GGOs. Dimension reduction is required for the purpose of alleviating overfitting. Selecting significant features that are relevant to both tasks or highly relevant to one of the tasks will certainly be desirable and is our main goal to achieve in this article.

In this paper, we model the across-task relatedness with a prior as sharing a common subset of features, and propose a novel algorithmic framework, based on mathematical programming, that eliminates features that are irrelevant or redundant for all tasks, and constructs classifiers for each task by further selecting features from the common set. Although the framework is general enough to be employed in any applications where supervised machine learning problems are involved, our major application domain lies in the area of computer aided diagnosis with medical images.

Formulations

The proposed approach is suitable to be combined with almost any specific existing classification or regression methods that deal with a single task. We take two exemplary methods, one for regression, one for classification, as examples to illustrate how our approach works. Prior to the thorough description of our formulations, we retrospect briefly on the two exemplary methods. Ridge regression has been a successful regression approach while 1-norm SVM has been widely appraised for dealing with the classification problems where feature selection is needed.

Assume that we have T tasks in total and we have sample set $\{(\mathbf{x}_i^t, y_i^t), i = 1, \dots, \ell_t\}$ for the t -th task where $\mathbf{x} \in R^n$. To simplify the notation, we use \mathbf{X}_t to denote the feature matrix where the i -th row corresponds to the transpose of \mathbf{x}_i^t , and \mathbf{y}_t to denote the label vector where the i -th component is y_i^t . Notice that y can take integer numbers $\{-1, 1\}$ for classification or is continuous for regression.

The Ridge regression method for solving the specific task t can be stated as follows:

$$\min_{\alpha_t} \|\mathbf{y}_t - \mathbf{X}_t \alpha_t\|^2 + \mu \|\alpha_t\|^2 \quad (1)$$

where $\|\cdot\|$ denotes 2-norm of a vector and μ is the regularization parameter that controls the balance between the error term (the first one) and the penalty term (the second one).

The 1-norm SVM for solving a single classification task t is stated as follows:

$$\begin{aligned} \min_{\alpha_t} \quad & \|\xi_t\|_1 + \mu \|\alpha_t\|_1 \\ \text{subject to} \quad & \mathbf{y}_t \otimes (\mathbf{X}_t \alpha_t) \geq 1 - \xi_t, \\ & \xi_t \geq 0, \end{aligned} \quad (2)$$

where \otimes denotes the component-wise multiplication between two matrices (or vectors).

The feature selection problem can be formulated as an integer programming problem, or in other words, a very difficult combinatorial optimization problem. Denote a matrix \mathbf{B} as an $n \times n$ diagonal matrix with its j -th diagonal element equal to $\beta_j \in \{0, 1\}$. We call \mathbf{B} an indicator matrix indicating whether or not an according feature is used to build a model. Then for each task, instead of learning a model $y = \mathbf{x}^\top \alpha$, we construct a model $y = \mathbf{x}^\top \mathbf{B} \alpha$ where α is task-specific while the same \mathbf{B} will be used across different tasks. If $\beta_j = 0$, the j -th variable is not used in any model for all tasks regardless of the value of a specific α . Otherwise if $\beta_j = 1$, the j -th variable appears in all models but an appropriate α vector can rule out this feature for a particular task. Then $\mathbf{X} \mathbf{B} \alpha = \tilde{\mathbf{X}} \mathbf{c}$ where $\tilde{\mathbf{X}}$ only contains the selected features and \mathbf{c} corresponds to nonzero components of $\mathbf{B} \alpha$. Then the feature selection approach for learning multiple tasks based on ridge regression is formulated as the following mixed integer program:

$$\begin{aligned} \min_{\beta} \quad & \min_{\alpha_t} \sum_{t=1}^T (\|\mathbf{y}_t - \mathbf{X}_t \mathbf{B} \alpha_t\|^2 + \mu_t \|\mathbf{B} \alpha_t\|^2) \\ \text{subject to} \quad & \mathbf{B} = \text{diag}(\beta), \quad \|\beta\|_0 = m, \\ & \beta_j \in \{0, 1\}, \quad j = 1, \dots, n. \end{aligned} \quad (3)$$

where $\|\cdot\|_0$ denotes the 0-norm (Weston *et al.* 2003) which controls the cardinality of β , (notice the 0-norm is not really a vector norm). This program attempts to choose m important features out of n features for all tasks.

Problem (3) is computationally intractable or expensive since it requires branch-and-bound procedure to optimize integer variables β . Development of mathematically tractable formulations is required for practical applications. We relax the constraints on integer variables β to allow them to take real numbers. Then these β variables correspond to certain scaling factors determining how significantly the corresponding features contribute to the target y . We then enforce the sparsity of β . Sparsity can be enforced by restricting the cardinality of β to exactly m as in (3), or by employing the 1-norm regularization condition on β , which is less stringent than the 0-norm penalty. To derive computationally efficient and scalable formulations, we relax the problem to impose a constraint on the 1-norm of β . Then the relaxation of problem (3) becomes

$$\begin{aligned} \min_{\beta} \quad & \min_{\alpha_t} \sum_{t=1}^T (\|\mathbf{y}_t - \mathbf{X}_t \mathbf{B} \alpha_t\|^2 + \mu_t \|\mathbf{B} \alpha_t\|^2) \\ \text{subject to} \quad & \mathbf{B} = \text{diag}(\beta), \quad \|\beta\|_1 \leq \delta, \\ & \beta_j \geq 0, \quad j = 1, \dots, n. \end{aligned} \quad (4)$$

where μ_t and δ are parameters to be tuned and pre-specified before solving problem (4).

Adding matrix \mathbf{B} to the 1-norm SVM (2) and applying the above relaxation yield a multi-task feature selection ap-

proach for classification which is formulated as follows:

$$\begin{aligned} \min_{\beta} \quad & \min_{\alpha_t} \quad \sum_{t=1}^T (\|\xi_t\|_1 + \mu_t \|\mathbf{B}\alpha_t\|_1) \\ \text{subject to} \quad & \mathbf{y}_t \otimes (\mathbf{X}_t \mathbf{B} \alpha_t) \geq 1 - \xi_t, \\ & \xi_t \geq 0, \quad t = 1, \dots, T, \\ & \mathbf{B} = \text{diag}(\beta), \quad \|\beta\|_1 \leq \delta, \\ & \beta_j \geq 0, \quad j = 1, \dots, n. \end{aligned} \quad (5)$$

Formulations (4) and (5) are non-convex and involve 4th order polynomials in the objective (4) or quadratic forms in constraints (5). We develop efficient algorithms for solving these formulations in later sections.

Probabilistic Interpretation

We derive a probabilistic interpretation using multi-task ridge regression as an example. Note that the probabilistic interpretation could be easily generalized to other loss functions. Consider the following generative framework:

$$\begin{aligned} y_t &= X_t B \alpha_t + \epsilon_t \\ \epsilon_t &\sim \text{Norm}(0, \sigma_t I) \\ p(\beta_i) &\sim \rho^{\beta_i} (1 - \rho)^{1 - \beta_i} \\ p(\alpha_t | B) &= P(B \alpha_t) \sim \text{Norm}(0, \hat{\sigma} I) \\ B &= \text{diag}(\beta) \text{ and } \beta_i \in \{0, 1\} \end{aligned}$$

where we use Bernoulli distribution with parameter ρ for each $\beta_i, i = 1, \dots, d$. The value of ρ will affect the likelihood of including a given feature. For example, setting $\rho = 1$ will preserve all features and smaller ρ values will result in the use of fewer features. The conditional probability $p(\alpha_t | B)$ basically tells that if the feature i is selected, the corresponding α_{ti} for all tasks follows a zero mean Normal distribution; otherwise it follows a noninformative distribution. Furthermore, the noises are assumed independent of each other between different tasks and also the following independency conditions hold:

$$\begin{aligned} p(\beta) &= \prod_{i=1}^d p(\beta_i) \\ p(\alpha_1, \dots, \alpha_T | B) &= \prod_{i=1}^T p(\alpha_i | B). \end{aligned}$$

Then, the posterior conditional distribution of model parameters $(\alpha_1, \dots, \alpha_T, \beta)$ satisfies, in the log form,

$$\begin{aligned} \log P(\alpha_1, \dots, \alpha_T, \beta | X_1, y_1, \dots, X_T, y_T) \\ = \sum_{t=1}^T (\|\mathbf{y}_t - \mathbf{X}_t \mathbf{B} \alpha_t\|^2 + \mu_t \|\mathbf{B} \alpha_t\|^2) + \lambda \sum_{i=1}^d \beta_i + C. \end{aligned}$$

where $\mu_t = \sigma_t^2 / \hat{\sigma}^2$, $\lambda = \log(\rho / (1 - \rho)) \sum \sigma_t^2$ and C is the normalization constant, and can be ignored.

The above derivation requires $\beta_i \in \{0, 1\}$. By relaxing the integer constraint with a nonnegative constraint $\beta_i \geq 0$, maximizing the above posterior distribution of model parameters will give us an equivalent formulation of (4).

Algorithms

The residual term in the objective of problem (4) is bilinear with respect to β and α_t . The 2-norm of the residual introduces to the optimization problem high order polynomials

and thus the problem is still arduous to solve. We propose an alternating optimization approach (Bezdek & Hathaway 2003) to solving formulation (4) by repeating steps depicted in Algorithm 1, which is similar, in spirit, to the principle of Expectation-Maximization (EM) algorithms. Moreover, note that $\|\beta\|_1 = \sum \beta_j$ due to the nonnegativity of β_j .

Algorithm 1

- Fix \mathbf{B} to the current solution (initially to the identity matrix \mathbf{I}), convert $\tilde{\mathbf{X}}_t \leftarrow \mathbf{X}_t \mathbf{B}$, solve the following problem for optimal α_t ,

$$\min_{\alpha_t} \quad \sum_{t=1}^T (\|\mathbf{y}_t - \tilde{\mathbf{X}}_t \alpha_t\|^2 + \mu_t \|\mathbf{B} \alpha_t\|^2) \quad (6)$$

- Fix α_t to the solution obtained at the above step, convert $\hat{\mathbf{X}}_t \leftarrow \mathbf{X}_t \cdot \text{diag}(\alpha_t)$, solve the following problem for optimal $\hat{\beta}$,

$$\begin{aligned} \min_{\beta \geq 0} \quad & \sum_{t=1}^T (\|\mathbf{y}_t - \hat{\mathbf{X}}_t \beta\|^2 + \mu_t \|\beta \otimes \alpha_t\|^2) \\ \text{subject to} \quad & \|\beta\|_1 \leq \delta. \end{aligned} \quad (7)$$

The algorithm can also take a greedy scheme to perform $\mathbf{B} \leftarrow \mathbf{B} \otimes \text{diag}(\hat{\beta})$ after the second step. It assures that features receiving small scaling factors in early iteration will continue receiving small weights. This greedy step speeds up the convergence process but makes the algorithm very likely terminate at sub-optimal solutions.

The first step of Algorithm 1 solves a simple ridge regression problem. Note that the problem (6) can be de-coupled to minimize $(\|\mathbf{y}_t - \tilde{\mathbf{X}}_t \alpha_t\|^2 + \mu_t \|\mathbf{B} \alpha_t\|^2)$ for each individual α_t of task t . Thus, the problem (6) actually has a closed-form solution, which is to solve $\mathbf{B} (\mathbf{X}_t^T \mathbf{X}_t + \mu_t \mathbf{I}) \mathbf{B} \alpha_t = \mathbf{B} \mathbf{X}_t \mathbf{y}_t$ where \mathbf{B} is a diagonal matrix with some diagonal components possibly equal to 0. So the solution is $\hat{\alpha}_t = \mathbf{B}^\dagger (\mathbf{X}_t^T \mathbf{X}_t + \mu_t \mathbf{I})^{-1} \mathbf{X}_t \mathbf{y}_t$ where \mathbf{B}^\dagger denotes the pseudo-inverse of \mathbf{B} , a diagonal matrix whose non-zero diagonal elements equal the inverse of nonzero diagonal components of \mathbf{B} . An advantage of this algorithm is that the matrix inversion $(\mathbf{X}_t^T \mathbf{X}_t + \mu_t \mathbf{I})^{-1}$ only needs to be calculated in the first iteration and can then be reused in later iterations, thus gaining computational efficiency.

The second step of Algorithm 1 solves a quadratic programming problem. Denote $\Lambda_t = \text{diag}(\alpha_t)$. The problem (7) can be rewritten in the following canonical form of a quadratic program:

$$\begin{aligned} \min_{\beta \geq 0} \quad & \beta^\top \sum_{t=1}^T (\Lambda_t (\mathbf{X}_t^\top \mathbf{X}_t + \mu_t \mathbf{I}) \Lambda_t) \beta \\ & - 2 \left(\sum_{t=1}^T \mathbf{y}_t^\top \mathbf{X}_t \Lambda_t \right) \beta \\ \text{subject to} \quad & \mathbf{e}^\top \beta \leq \delta. \end{aligned} \quad (8)$$

Problem (8) is a simple quadratic program where \mathbf{e} is the vector of ones of proper dimension, but a quadratic problem is less scalable than a linear program from the optimization perspective.

The proposed multi-task feature selection approach seeks a subset of m (or few) features of \mathbf{X} so that the best ridge regression models for each of the tasks can be attained using the selected features. The standard formulation (1) minimizes the quadratic loss and is subject to capacity control determined by the 2-norm penalty. Many recent studies (Zhu *et al.* 2004; Bi *et al.* 2003) have shown that the 1-norm regularization together with the absolute deviation loss is equally suitable to learning regression or classification models, and often produces sparse solutions for better approximation. Our algorithms can easily generalize to other loss functions and various regularization penalties.

Following the derivation of Algorithm 1 for multi-task ridge regression, we design an alternating algorithm for the multi-task 1-norm SVM (5) in Algorithm 2.

Algorithm 2

- Fix \mathbf{B} to the current solution, convert $\tilde{\mathbf{X}}_t \leftarrow \mathbf{X}_t \mathbf{B}$, solve the following problem for optimal α_t ,

$$\begin{aligned} \min_{\alpha_t, \xi_t, \mathbf{v}_t} \quad & \sum_{t=1}^T \left(\mathbf{e}^\top \xi_t + \mu_t \beta^\top \mathbf{v}_t \right) \\ \text{subject to} \quad & \mathbf{y}_t \otimes (\tilde{\mathbf{X}}_t \alpha_t) \geq 1 - \xi_t, \\ & \xi_t \geq 0, \quad t = 1, \dots, T, \\ & -\mathbf{v}_t \leq \alpha_t \leq \mathbf{v}_t, \quad t = 1, \dots, T. \end{aligned} \quad (9)$$

- Fix α_t to the solution obtained at the above step, convert $\hat{\mathbf{X}}_t \leftarrow \mathbf{X}_t \cdot \text{diag}(\alpha_t)$, solve the following problem for optimal $\hat{\beta}$,

$$\begin{aligned} \min_{\beta \geq 0} \quad & \sum_{t=1}^T \mathbf{e}^\top \xi_t + \left(\sum_{t=1}^T \mu_t \alpha_t \right)^\top \beta \\ \text{subject to} \quad & \mathbf{y}_t \otimes (\hat{\mathbf{X}}_t \beta) \geq 1 - \xi_t, \\ & \xi_t \geq 0, \quad t = 1, \dots, T, \\ & \mathbf{e}^\top \beta \leq \delta. \end{aligned} \quad (10)$$

Note that both problems (9) and (10) are linear programs, and can be solved efficiently. Further, problem (9) can be decoupled as well to optimize each individual α_t separately by minimizing $\mathbf{e}^\top \xi_t + \mu_t \beta^\top \mathbf{v}_t$ with constraints $\mathbf{y}_t \otimes (\tilde{\mathbf{X}}_t \alpha_t) \geq 1 - \xi_t, \xi_t \geq 0, -\mathbf{v}_t \leq \alpha_t \leq \mathbf{v}_t$. These T subproblems are small and thus the overall algorithm is scalable.

Experiments

We validate the proposed approach on classification tasks by comparing it to standard approaches where tasks are solved independently using the 1-norm SVM, and comparing it to the pooling method where a single model is constructed using available data from all tasks. These methods represent two extreme cases: the former one treats multiple tasks completely independently assuming no relatedness; the latter one treats all tasks identically. Our results clearly show that the multi-task learning approach as proposed is superior to these extreme cases. We also implemented another multi-task learning approach (Evegniou & Pontil 2004) that is derived based on the regularization principle and we compared it to the proposed approach in terms of performance.

Synthetic data

We generated synthetic data to verify the behavior of the proposed algorithms regarding the selected features and the accuracy in comparison with single-task 1-norm SVM. The synthetic data was generated as follows.

Synthetic Data Generation

1. Set number of features $d = 10$, number of tasks $T = 3$.
2. Generate $\mathbf{x} \in R^{10}$ with each component $x_i \sim \text{Uniform}[-1, 1], i = 1, \dots, 10$.
3. The coefficient vectors of three tasks are specified as:

$$\beta_1 = [1, 1, 0, 0, 0, 0, 0, 0, 0, 0]$$

$$\beta_2 = [1, 1, 1, 0, 0, 0, 0, 0, 0, 0]$$

$$\beta_3 = [0, 1, 1, 1, 0, 0, 0, 0, 0, 0]$$

4. For each task and each data vector, $y = \text{sign}(\beta^\top \mathbf{x})$.

For each task, we generated training sets of sizes from 20 to 90, each used in a different trial, 150 samples for validation and 1000 samples for testing, and repeated each trial 20 times. In figure 2, we show a bar plot of the averaged estimated coefficient vectors by our approach and the single-task 1-norm SVM. Clearly, our approach successfully removed all irrelevant features. Since linear classifiers were used, re-scaling the classifier by a constant did not have effect on predictions. Each coefficient vector was normalized by its norm, then averaged over all runs in all trials, and shown on figure 2. Although single-task learning also produced reasonable classifiers for each task, it could not remove all irrelevant features using data available for each individual task in every trial.

Figure 2(right) shows the prediction results. For lucid presentation, we have averaged the classification errors of the three tasks over 20 runs and drawn them in figure 2 with error bars proportional to error standard deviation. It shows that our approach outperforms the single-task approach and as expected, the difference of the two approaches becomes smaller as the sample size of each task becomes larger.

Lung CAD data

The standard paradigm for computer aided diagnosis of medical images follows a sequence of three stages: identification of potentially unhealthy candidate regions of interest (ROI) from the image volume, computation of descriptive features for each candidate, and classification of each candidate (eg normal or diseased) based on its features.

Data preparation A prototype version of our lung CAD system (not commercially available) was applied on a proprietary de-identified patient data set. The nodule dataset consisted of 176 high-resolution CT images (collected from multiple sites) that were randomly partitioned into two groups : a training set of 90 volumes and a test set of 86 volumes. The GGO dataset consisted of 60 CT images. Since there were only a limited number of GGO cases, they were not partitioned beforehand to have a test set. The original goal was to use the additional GGO cases to improve the

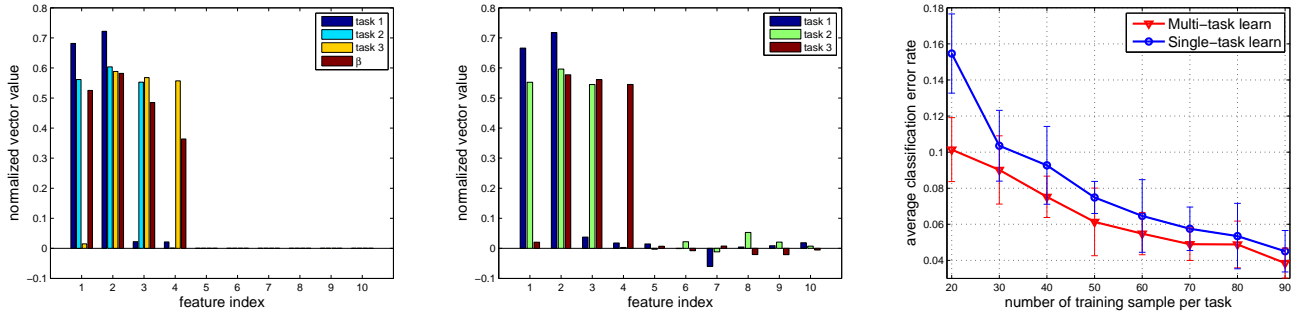


Figure 2: Performance on synthetic data, left: coefficient vectors by our approach; middle: coefficient vectors by single-task 1-norm SVM; right: accuracy comparison with various training sizes.

	Nodule train	Nodule test	GGO
# patients	90	86	60
# cand.	11056	13985	10265
# cancer	81	48	53
# positives	131	81	87
# FP/vol	121	161	169
# feature	86	86	86

Table 1: Specifications of lung CAD data sets.

nodule detection performance. In total, 129 nodules and 53 GGOs were identified and labeled by radiologists. Among the marked nodules, 81 appeared in the training set and 48 in the test set. The training set was then used to optimize the classification parameters, and construct the final classifier which was tested on the independent test set of 86 volumes.

The candidate identification algorithm was independently applied to the training, test nodule sets and the GGO set, achieving 98.8% detection rate on the training set at 121 FPs per volume, 93.6% detection rate on the test set at 161 FPs per volume and 90.6% detection rate on the GGO set at 169 FPs per volume, resulting in totally 11056, 13985 and 10265 candidates in the respective nodule training, nodule test and GGO sets. There can exist multiple candidates pointing to one nodule or one GGO, so 131, 81 and 87 candidates were labeled as positive in the training set, test set and GGO set, respectively. A total of 86 numerical image features were designed to depict both nodules and GGOs. The feature set contained some low-level image features, such as size, shape, intensity, template matching features, and some high-level features, such as multi-scale statistical features depicting sophisticated higher-order properties of nodules and GGOs. The specifications of all the related data sets are summarized in Table for clarity.

Experimental setting and performance The first set of experiments were conducted as follows. We randomly sampled 50% (45 volumes) of the nodule patient data from the training set, 50% (30 volumes) of the GGO patient data. These samples were used in the training phase. Notice that the random sampling can only take place at the patient level

rather than the candidate level since otherwise information from a single patient may appear in both training and test sets, making the testing not independent. The nodule classifiers obtained by our approach and three other approaches were tested on the unseen test set of 86 patient cases.

We compared Algorithm 2 to the single task 1-norm SVM, the pooling method with 1-norm SVM, and the regularized MTL (Evegniou & Pontil 2004). In the first trial, we tuned the model parameters such as μ_1, μ_2, δ in Algorithm 2 and the regularized parameters in (Evegniou & Pontil 2004) according to a 3-fold cross validation performance, and $\mu_1 = 0.2$ for GGOs, $\mu_2 = 1$ for nodules were the best choice for single task learning. Then we fixed them for other trials, and used the same μ s in the proposed multi-task learning formulation (5) for a fair comparison since the multi-task learning had the same parameter settings as a start, and then tuned δ ($=10$) to improve performance. Note that the proposed Algorithm 2 may produce better performance if we tuned μ according to its own cross validation performance.

Figure 3(left) shows ROC curves averaged over the 15 trials together with test error bars as the standard deviation of detection rates of the 15 trials. Clearly, Algorithm 2 generates a curve that dominates the ROC curves corresponding to other approaches. It also had a relatively small model variance by referencing the error bars. The classifier test error variance of the regularized MTL varied significantly with variations of samples as shown in Figure 3.

We also report the performance comparisons with area-under-the-ROC-curve (AUC) measure since AUC is independent of the selected decision criterion and prior probabilities. We randomly sampled $p\%$ of training nodule set and the GGO set where $p = 10, 25, 50, 75, 100$. Obviously, when more and more data for a specific task is available, the resulting model achieves better performance, and accurate models can be learned with less help from other related tasks. We averaged the AUC numbers over 15 trials for each sample size choice p . Figure 3(right) illustrates the averaged AUC values and associated error bars. Our method presents relatively small model variance in comparison with the regularized MTL as shown in the error bars.

A recent paper (Argyriou, Evgeniou, & Pontil 2007) proposes a method to learn sparse feature representation from

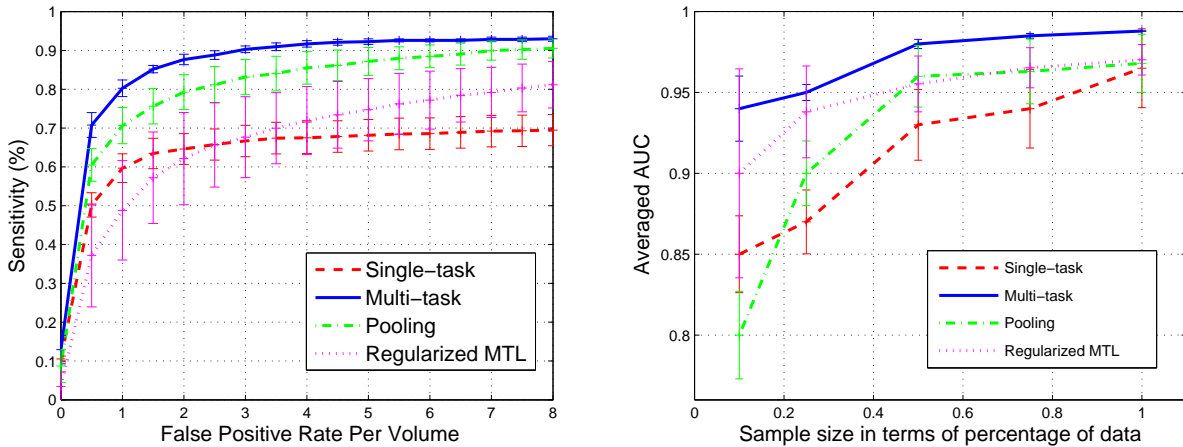


Figure 3: Left: ROC plot on 50% of nodule and GGO training patient volumes; right: the AUC plot versus sample size.

multiple tasks. It does not directly enforce sparsity on original feature set if orthonormal transformation is applied to features since the orthonormal matrix \mathbf{U} is not in general sparse. We implemented this method using $\mathbf{U} = \mathbf{I}$ for comparison. Our method provides more sparse solutions $\mathbf{B}\alpha_t$.

Conclusions

We have discussed the challenges of collaborative computer aided diagnosis which motivated the investigation of a mathematical-programming based multi-task learning framework. By applying an indicator vector β to the feature sets across different tasks and regularizing on the 1-norm of β , similar feature patterns across different tasks are encouraged and features that are irrelevant to any of the tasks are eliminated. Efficient algorithms have been devised to solve our formulations. Experimental results on detecting solid nodules from CT images show that the proposed approach outperforms the regularized multi-task learning approach and traditional single-task-learning and pooling methods. Due to limited available medical data, more extensive evaluation of our system on three or more related CAD tasks remains for further research.

References

Ando, R. K., and Zhang, T. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research* 6:1855–1887.

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2007. Multi-task feature learning. In Schölkopf, B.; Platt, J.; and Hoffman, T., eds., *Advances in Neural Information Processing Systems 19*. Cambridge, MA: MIT Press.

Armato-III, S. G.; Giger, M. L.; and MacMahon, H. 2001. Automated detection of lung nodules in CT scans: preliminary results. *Medical Physics* 28(8):1552 – 1561.

Bezdek, J. C., and Hathaway, R. J. 2003. Convergence of alternating optimization. *Neural, Parallel Sci. Comput.* 11:351–368.

Bi, J.; Bennett, K.; Embrechts, M.; Breneman, C.; and Song, M. 2003. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research* 3:1229–1243.

Buchbinder, S.; Leichter, I.; Lederman, R.; Novak, B.; Bamberger, P.; Sklair-Levy, M.; Yarmish, G.; and Fields, S. 2004. Computer-aided classification of BI-RADS category 3 breast lesions. *Radiology* 230:820 – 823.

Caruana, R. 1997. Multitask learning. *Machine Learning* 28(1):41–75.

Evgeniou, T., and Pontil, M. 2004. Regularized multi-task learning. In *Proc. of 17-th SIGKDD Conf. on Knowledge Discovery and Data Mining*.

Heskes, T. 2000. Empirical bayes for learning to learn. In Langley, P., ed., *Proceedings of the 17th International Conference on Machine Learning*, 367–374.

Naidich, D. P.; Ko, J. P.; and Stoeckel, J. 2004. Computer aided diagnosis: Impact on nodule detection amongst community level radiologist. A multi-reader study. In *Proceedings of CARS 2004 Computer Assisted Radiology and Surgery*, 902 – 907.

Roehrig, J. 1999. The promise of CAD in digital mammography. *European Journal of Radiology* 31:35 – 39.

Suzuki, K.; Kusumoto, M.; Watanabe, S.; Tsuchiya, R.; and Asamura, H. 2006. Radiologic classification of small adenocarcinoma of the lung: Radiologic-pathologic correlation and its prognostic impact. *The Annals of Thoracic Surgery CME Program* 81:413–20.

Weston, J.; Elisseeff, A.; Schölkopf, B.; and Tipping, M. 2003. Use of the zero-norm with linear models and kernel methods. *Journal of Machine Learning Research* 3:1439–1461.

Zhu, J.; Rosset, S.; Hastie, T.; and Tibshirani, R. 2004. 1-norm support vector machines. In Thrun, S.; Saul, L.; and Schölkopf, B., eds., *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press.