A Min-Max Framework of Cascaded Classifier with Multiple Instance Learning for Computer Aided Diagnosis

Dijia Wu[1]*, Jinbo Bi[2], Kim Boyer[1]

[1]Rensselaer Polytechnic Institute, Troy, NY 12180 USA, wud5@rpi.edu
[2]Siemens Medical Solutions, Malvern, PA 19355 USA, jinbo.bi@siemens.com

Abstract

The computer aided diagnosis (CAD) problems of detecting potentially diseased structures from medical images are typically distinguished by the following challenging characteristics: extremely unbalanced data between negative and positive classes; stringent real-time requirement of online execution; multiple positive candidates generated for the same malignant structure that are highly correlated and spatially close to each other. To address all these problems, we propose a novel learning formulation to combine cascade classification and multiple instance learning (MIL) in a unified min-max framework, leading to a joint optimization problem which can be converted to a tractable quadratically constrained quadratic program and efficiently solved by block-coordinate optimization algorithms.

We apply the proposed approach to the CAD problems of detecting pulmonary embolism and colon cancer from computed tomography images. Experimental results show that our approach significantly reduces the computational cost while yielding comparable detection accuracy to the current state-of-the-art MIL or cascaded classifiers. Although not specifically designed for balanced MIL problems, the proposed method achieves superior performance on balanced MIL benchmark data such as MUSK and image data sets.

1. Introduction

Over the years, computer aided diagnosis (CAD) systems have been widely used to assist physicians in interpreting medical images from different modalities such as magnetic resonance imaging (MRI), X-ray, and computed tomography (CT) and to identify potentially diseased regions like lesions or tumors. Most CAD systems comprise of three stages: identify candidate structures, *i.e.*, potentially unhealthy regions, in the image; generate features for each candidate; classify each candidate as normal (negative) or diseased (positive). To maintain high sensitivity, a very large number of candidates are generated in the first stage because any malignant regions missed at this stage can never be recovered later in the CAD system. Consequently, majority of the candidates generated, typically more than 99%, are false positives, which makes the data extremely unbalanced. In this situation, cascaded classifiers can be used to speed up candidate classification by quickly discarding numerous negative samples with low-cost features at early stages and spending more computation on promising disease-like candidates [15].

Moreover, for CAD data, a candidate is labeled as positive if it is sufficiently close to a radiologist's mark (ground truth) and labeled as negative otherwise. Multiple candidates are usually generated corresponding to the same abnormal structure so that if any such candidate is detected, the underlying structure is found. Therefore, CAD problems are better modeled as multiple instance learning (MIL) by enclosing all the candidates within a certain distance to a radiologist's mark into a positive bag [6].

In this paper, we propose a novel approach to combine MIL classifiers in a cascade. In particular, we start out with formulating MIL as an optimization problem in a min-max framework in Section 2. Section 3 reviews the joint optimization principle [5] used to construct all hyperplane classifiers of a cascade in one shot, and describes a new min-max formulation for optimization of the cascade. The two min-max frameworks are fused as discussed in Section 4 to form a unified approach that optimizes a cascade of MIL classifiers simultaneously. Experimental results on two CAD applications and MIL benchmark datasets are given in Section 5 together with some discussion. We conclude with a review of our contributions and potential extensions in Section 6.

*This work was conducted when Dijia Wu was with Siemens Medical Solutions at Malvern PA, USA.

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

2. Multiple Instance Learning

Multiple Instance Learning, originally formalized by [4], is a generalization of supervised learning where the training class labels are associated with sets of instances, *i.e.*, bags, rather than individual instances. A bag is labeled positive if at least one instance in it is positive, and negative if all the instances in it are negative. The objective of an MIL problem is to correctly identify at least one positive instance for each positive bag from all the other negative instances. In CAD problems, multiple positive candidates are generated within a certain distance from a marked lesion. The lesion is detected if only one of these candidates is correctly identified in the classification stage and highlighted to the physicians. Therefore, it is reasonable to assume that MIL algorithms will outperform other traditional classifiers for CAD tasks.

Many MIL algorithms have been proposed over the last decade. Diverse Density (DD), first introduced by [9], and its variant EM Diverse Density in [18], assume that all positive bags intersect at a few points in the feature space. However, this doesn't necessarily hold for all MIL problems. To achieve higher performance, various standard classification algorithms for single instance learning have been extended to adapt to the MIL scenario, such as mi(MI)-SVM [1], MILBoost [16], MILR [12] and MI RVM [10].

We propose a min-max framework for MIL problems that can be applied to many standard learning algorithms, such as Fisher's linear discriminant analysis or support vector machines (SVM), by generalizing the notion of loss functions to bags and minimizing the bag loss directly. In this paper, we use 1-norm SVM as an example to illustrate this approach because 1-norm SVM accomplishes automatic feature selection by shrinking coefficients of irrelevant features to zero [13, **?**], which is particularly desirable in CAD systems for computational efficiency at run time.

The following notation is used throughout the paper. We denote $\mathbf{x}$ as a feature vector (an instance) and $y \in \{+1, -1\}$ is the class label assigned to a bag. Keep in mind that class labels are only available to bags. We often implicitly apply the bag label to its instances for notational convenience, although not all instances in positive bags are truly positive instances. In total, there are N instances, N_+ instances in positive bags indexed in C^+ and N_- instances in negative bags indexed in C^-. Index sets B_i^+ and B_j^- contain indices of all instances in the i-th positive and the j-th negative bags, respectively, and N_{B+} and N_{B-} denote the numbers of positive and negative bags.

The regular 1-norm SVM can be formulated as the following optimization problem:

$$\begin{aligned} \underset{\xi,\eta,\omega,b}{\arg\min} \quad & \gamma||\omega||_1 + \sum_{j=1}^{N_+} \xi_j + \sum_{j=1}^{N_-} \eta_j \\ s.t. \quad & \xi_j \geq 1 - (\omega^T \mathbf{x}_j + b), \;\; \xi_j \geq 0, \;\; j \in C^+ \\ & \eta_j \geq 1 + (\omega^T \mathbf{x}_j + b), \;\; \eta_j \geq 0, \;\; j \in C^- \end{aligned} \tag{1}$$

where ω, b and γ represent the weighting vector, offset and regularization parameter; ξ and η measure the hinge loss of instances in C^+ and in C^-, respectively.

The typical objective of MIL problems is to correctly classify the bag rather than each individual instance. For example, in a CAD system, as long as the abnormal structure is found from the medical images, it is not important to detect all candidates corresponding to it. Hence, it is natural to extend the notion of hinge loss from instances to bags. As shown in Eq. (2), for a positive bag the hinge loss ξ is defined by the *minimum* hinge loss incurred on all instances in the bag, *i.e.*, the hinge loss of the *most-likely positive* instance in bag i; on the other hand, the hinge loss η of a negative bag is defined by the *maximum* loss of all instances in this bag, *i.e.*, the hinge loss of the *least negative* instance in the negative bag i. The definition (2) reflects the asymmetric nature between the positive bag where at least one of the instances in it is positive, and the negative bag where all its instances are negative.

$$\xi_i' = \min_{j \in B_i^+} \{\xi_j\}, \qquad \eta_i' = \max_{j \in B_i^-} \{\eta_j\} \tag{2}$$

Using the definition of bag hinge loss, 1-norm SVM can be modified as follows for MIL:

$$\begin{aligned} \underset{\xi,\eta,\omega,b}{\arg\min} \quad & \gamma||\omega||_1 + \sum_{i=1}^{N_{B+}} \min_{j \in B_i^+} \{\xi_j\} + \sum_{i=1}^{N_{B-}} \max_{j \in B_i^-} \{\eta_j\} \\ s.t. \quad & \xi_j \geq 1 - (\omega^T \mathbf{x}_j + b), \;\; \xi_j \geq 0, \;\; j \in C^+ \\ & \eta_j \geq 1 + (\omega^T \mathbf{x}_j + b), \;\; \eta_j \geq 0, \;\; j \in C^- \end{aligned} \tag{3}$$

It is important to point out that the formulation (3) implicitly takes all instances in a positive bag as positive, which is different from a standard MIL problem where instances in a positive bag can be unlabeled or even negative. However, the new formulation aims at detecting at least one instance from each positive bag, which corresponds to the identification of the most-likely positive instances for positive bags. Hence our algorithm finds both the positive bags and corresponding positive instances.

3. Cascade of Hyperplane Classifiers

The cascade classification approaches have long been used in real time object detection [15, 11, 17]. The proposed coarse-to-fine cascade structure, such as cascade AdaBoost,

has been shown effective in speeding up the detection process by discarding many of negative candidates with fewer features before more complex classifiers are called upon to achieve lower false positive rates. However, this structure also has several disadvantages: (1) In each stage of training, a minimum detection rate and maximum false positive rate have to be set *a priori* as a target to achieve, which can be difficult to accomplish especially in later stages as the number of stages increases; (2) To meet specific target detection and false positive rates, a classifier is trained with a fixed rejection threshold at each stage which is *inflexible* to explore the trade-off between accuracy and speed, and even between detection and false positive rate; (3) The classifiers at each stage are sequentially and separately optimized without utilizing the information derived at other stages except the current stage.

Bourdev proposed a soft cascade framework to improve detector accuracy by generalizing each stage to be a scale-valued decision function based on the values of all prior stages rather than just the current stage under consideration, but it still needs to select a rejection distribution vector and fix classifier thresholds at each stage in a separate calibration step [3].

In view of these drawbacks of classic cascade approaches, an AND-OR framework has been proposed in [5] to jointly optimize all classifiers in the cascade. This framework as depicted in Figure 1 minimizes the overall regularized risk of the entire system and provides implicit mutual feedback among all involved classifiers. The new cascade structure differs from classic ones in several perspectives: 1. It is not required to set any target sensitivity and specificity in the training, and thus there are no fixed thresholds, either. 2. Each classifier is optimized using all training data with different feature sets, providing the option for users to group features based on prior knowledge about features. For instance, computationally inexpensive features may be grouped together in earlier stages. 3. Since the classifiers are trained in parallel, the execution order of classifiers in the cascade has no impact on the classification accuracy. The execution order can be determined by minimizing the overall computational cost of the system. For instance, a classifier that removes more false positives than other classifiers may be applied earlier.

In this paper, we present a novel min-max optimization approach to implement the AND-OR framework of building a cascade of hyperplane classifiers, which is in the same spirit as that of MIL in Section 2. Although our algorithm can be extended to construct nonlinear classifiers, we limit our discussion to linear cases here for clarity. The new cascade formulation minimizes the *maximum* hinge loss across all stages on a positive instance since all stages have to classify the positive instance correctly for a final detection, and the *minimum* hinge loss across all stages on a negative instance since the negative instance can be rejected out of hand at any stage where the hinge loss is 0.

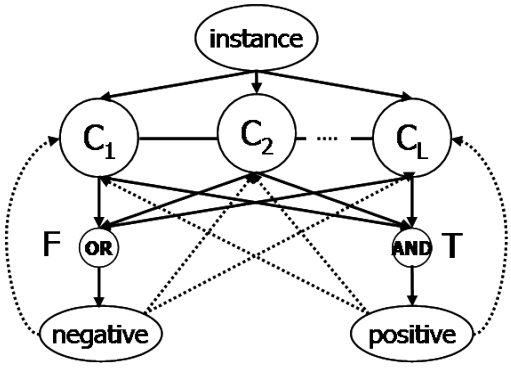


Figure 1. The proposed cascade framework comprising L classifiers: C_1, C_2, ... C_L.

$$\begin{aligned}
\underset{\xi,\eta,\omega,b}{\arg\min} \quad & \gamma \sum_{k=1}^{L} ||\omega_k||_1 + \sum_{j=1}^{N_+} \max_{1\le k\le L} \{\xi_{j,k}\} \\
& + \sum_{j=1}^{N_-} \min_{1\le k\le L} \{\eta_{j,k}\} \qquad (4) \\
s.t. \quad & \xi_{j,k} \ge 1 - (\omega_k^T \mathbf{x}_{j,k} + b_k),\ \xi_{j,k} \ge 0, \\
& \qquad j \in C^+, k = 1, \cdots, L \\
& \eta_{j,k} \ge 1 + (\omega_k^T \mathbf{x}_{j,k} + b_k),\ \eta_{j,k} \ge 0, \\
& \qquad j \in C^-, k = 1, \cdots, L
\end{aligned}$$

where L is the number of classifiers in the cascade, ω_k and b_k represent the weight vector and offset of classifier k, $\mathbf{x}_{j,k}$ is the jth training sample with features in the feature set k.

4. Min-Max Optimization of Cascaded Classifier with MIL

Now, we integrate MIL into the cascaded classification framework. In the test phase, each instance is labeled jointly by all classifiers in the cascade and then the bag is labeled according to labels of the instances in the bag based on the definition of positive and negative bags. Fig. 2 illustrates the idea. The dotted circles represent the positive bags and solid circles stand for the negative bags. The bags can be separated by two simple cascaded classifiers: $y_1 = sgn(-x)$ and $y_2 = sgn(-y)$ where $sgn(x)$ is equal to 1 if $x \ge 0$ and -1 otherwise. The instances with both $y_1 = 1$ and $y_2 = 1$ are labeled as positive which are represented by solid triangles, and a bag containing at least one of these triangles is classified as a positive bag.

Following the same notation, the unified framework of

cascaded MIL classification based on the 1-norm SVM can be formulated as follows:

$$\begin{aligned}\underset{\xi,\eta,\omega,b}{\arg\min}\ & \gamma\sum_{k=1}^{L}||\omega_k||_1+\sum_{i=1}^{N_{B+}}\min_{j\in B_i^+}\max_{1\le k\le L}\{\xi_{j,k}\}\\ &+\sum_{i=1}^{N_{B-}}\max_{j\in B_i^-}\min_{1\le k\le L}\{\eta_{j,k}\} \qquad (5)\\ s.t.\quad & \xi_{j,k}\ge 1-(\omega_k^T\mathbf{x}_{j,k}+b_k),\ \xi_{j,k}\ge 0,\\ & \qquad j\in C^+,k=1,\cdots,L\\ & \eta_{j,k}\ge 1+(\omega_k^T\mathbf{x}_{j,k}+b_k),\ \eta_{j,k}\ge 0,\\ & \qquad j\in C^-,k=1,\cdots,L\end{aligned}$$

The optimization problem (5) is computationally difficult to solve because it involves a *min-max* of the to-be-determined variables $\xi_{j,k}$, and a *max-min* of $\eta_{j,k}$ in the objective function, which is not differentiable or convex. Hence we convert it to an equivalent and tractable quadratically constrained quadratic program as follows:

$$\underset{\zeta,\mu,\xi,\eta,\lambda,\nu,\omega,b}{\arg\min}\ \gamma\sum_{k=1}^{L}||\omega_k||_1+\sum_{i=1}^{N_{B+}}\sum_{j\in B_i^+}\lambda_j\zeta_j+\sum_{i=1}^{N_{B-}}\mu_i \qquad (6)$$

$$\begin{aligned}s.t.\quad & \text{For a positive bag}\\ & \xi_{j,k}\ge 1-(\omega_k^T\mathbf{x}_{j,k}+b_k),\ \xi_{j,k}\ge 0,\\ & \qquad j\in C^+,\ k=1,\cdots,L,\\ & \zeta_j\ge\xi_{j,k},\ \ \forall 1\le k\le L\\ & \sum_{j\in B_i^+}\lambda_j=1,\ \lambda_j\ge 0,\ i=1,\cdots,N_{B+}\end{aligned} \qquad (7)$$

$$\begin{aligned}& \text{For a negative bag}\\ & \eta_{j,k}\ge 1+(\omega_k^T\mathbf{x}_{j,k}+b_k),\ \eta_{j,k}\ge 0,\\ & \qquad j\in C^-,\ k=1,\cdots,L,\\ & \mu_i\ge\sum_{k=1}^{L}\nu_{j,k}\eta_{j,k},\ \ \forall j\in B_i^-\\ & \sum_{k=1}^{L}\nu_{j,k}=1,\ \nu_{j,k}\ge 0.\end{aligned} \qquad (8)$$

The following theorem characterizes the equivalence between the above two optimization problems.

Theorem 1 *The optimal solution of Problem* (5) *is identical to the optimal solution of Problem* (6) *when* λ_j *and* $\nu_{j,k}$ *are chosen properly.*

Proof. Denote $\mathcal{J}(\zeta,\mu,\xi,\eta,\lambda,\nu,\omega,b)$ as the object function in the optimization problem (6) and assume that

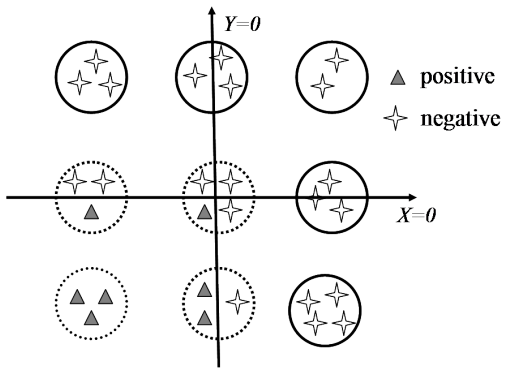


Figure 2. A toy example illustrating the proposed approach. Positive and negative instances are represented by solid triangles and hollow stars. Dotted circles stand for positive bags and solid circles for negative bags.

$(\hat{\zeta},\hat{\mu},\hat{\xi},\hat{\eta},\hat{\lambda},\hat{\nu},\hat{\omega},\hat{b})$ is the optimal solution to this optimization problem with $\hat{\mathcal{J}}$ as the optimal value attained at the optimal solution. Obviously, $(\hat{\xi},\hat{\eta},\hat{\omega},\hat{b})$ is feasible to Problem (5). Notice that the hinge loss $\hat{\xi}$ and $\hat{\eta}$ can be uniquely determined by $\hat{\omega}$ and $\hat{b}$ as

$$\begin{aligned}\hat{\xi}_{j,k}&=max(0,1-\hat{\omega}_k^T\mathbf{x}_{j,k}-\hat{b}_k),\\ \hat{\eta}_{j,k}&=max(0,1+\hat{\omega}_k^T\mathbf{x}_{j,k}+\hat{b}_k),\end{aligned} \qquad (9)$$

and the inequality constraints in (7) and (8) imply that

$$\begin{aligned}\hat{\zeta}_j&\ge\max_{1\le k\le L}\hat{\xi}_{j,k},\\ \hat{\mu}_i&\ge\max_{j\in B_i^-}\textstyle\sum_{k=1}^{L}\hat{\nu}_{j,k}\hat{\eta}_{j,k},\end{aligned} \qquad (10)$$

furthermore, it is straightforward to derive that

$$\begin{aligned}&\textstyle\sum_{j\in B_i^+}\hat{\lambda}_j\hat{\zeta}_j\ \ s.t.\ \sum_{j\in B_i^+}\hat{\lambda}_j=1,\ \hat{\lambda}_j\ge 0\\ &\textstyle\sum_{k=1}^{L}\hat{\nu}_{j,k}\hat{\eta}_{j,k}\ \ s.t.\ \sum_{k=1}^{L}\hat{\nu}_{j,k}=1,\ \hat{\nu}_{j,k}\ge 0\end{aligned} \qquad (11)$$

are minimized when $\hat{\lambda}_j$ and $\hat{\nu}_{j,k}$ are all zero except for

$$j^*=\underset{j\in B_i^+}{\arg\min}\,\hat{\zeta}_j \qquad (12)$$

and

$$k^*=\underset{1\le k\le L}{\arg\min}\,\hat{\eta}_{j,k}, \qquad (13)$$

respectively. Accordingly, at the optimal solution to problem (6), the objective value $\hat{\mathcal{J}}$ equates to

$$\begin{aligned}\gamma\sum_{k=1}^{L}||\hat{\omega}_k||_1+&\sum_{i=1}^{N_{B+}}\min_{j\in B_i^+}\max_{1\le k\le L}\{\hat{\xi}_{j,k}\}\\ &+\sum_{i=1}^{N_{B-}}\max_{j\in B_i^-}\min_{1\le k\le L}\{\hat{\eta}_{j,k}\}\end{aligned} \qquad (14)$$

which is exactly the same as the objective function of Problem (5) which implies the optimal value of Problem (5) is attained at the solution $(\hat{\xi},\hat{\eta},\hat{\omega},\hat{b})$. ∎

To solve the optimization problem (6) by applying the block-coordinate optimization algorithms [14], we exploit the fact that problem (6) reduces to a linear program that can be solved exactly and quickly if parameters λ_j and $\nu_{k,j}$ are fixed to constants. The scheme of the optimization procedure alternating over two subsets of the parameters is described as follows:
Our algorithm takes training data $\{x_j\}_{j=1}^{N}$, bag index sets $\{B_i^+\}_{i=1}^{N_{B+}}$, $\{B_i^-\}_{i=1}^{N_{B-}}$, and regularization const γ as inputs, and automatically computes the weighting vectors $\{\omega_k\}_{k=1}^{L}$ and offsets $\{b_k\}_{k=1}^{L}$ of the cascaded MIL classifier as outputs. The algorithm repeatedly alternates between optimizing the classifier parameters $\{\omega_k\}, \{b_k\}$ and the combination coefficients used in the MIL and cascade, i.e. parameters $\{\lambda_j\}$ and $\{\nu_{j,k}\}$ until a termination rule is satisfied.

At each iteration, classifier parameters $\{\omega_k\}, \{b_k\}$ are optimized by solving Problem (6) with fixed λ and ν. Notice Problem (6) becomes a linear program when λ and ν are fixed. Then the bag-level hinge loss can be evaluated as in Eqs. (9) and (10) for the obtained $\{\omega_k\}, \{b_k\}$. We then compute j^* and k^* as in Eqs. (12) and (13), and update $\lambda_{j^*} = 1$, other λ's to 0, and $\nu_{j,k^*} = 1$, other ν's to 0.

There are many possibilities to refine the above optimization heuristic strategy, for instance, by starting from different initial parameter values, or by performing simulated annealing to avoid unfavorable local minima. However, we have been able to achieve competitive results even with this simple optimization heuristic as shown in the next section, which validates the cascaded MIL algorithm. Further algorithmic improvements will be addressed in future work.

5. Experimental Results and Discussion

For the experiments in this paper, we compare our algorithm with three other techniques: regular 1-norm SVM without MIL and cascading, MI RVM which is reported to achieve the most competitive classification accuracy compared with other popular MIL classification methods [10], and cascade AdaBoost well known for its high computational efficiency [15]. The proposed cascade MIL classification algorithm is validated with respect to the generalization performance and computational efficiency.

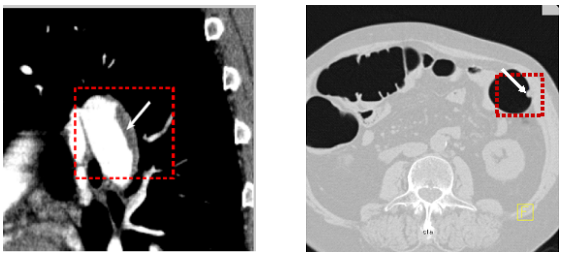
5.1. CAD applications: Pulmonary embolism and Colon cancer detection

Pulmonary embolism (PE) is a blockage of the pulmonary artery or one of its branches usually from a blood clot that breaks off from a large vein and travels to the lungs [2], see Fig. 3(a). As the third most common cause of death in the US, PE is a highly lethal condition with at least 650,000 cases occurring annually. Most patients who die of PE do so within 30 to 60 minutes after symptoms start; on the other hand, anti-clotting medications which are highly effective in treating PEs may lead to hemorrhage and bleeding sometimes. Therefore, an accurate and quick diagnosis is the key to survival. Computed tomography angiography (CTA) has emerged as an accurate diagnostic tool for PE. However, hundreds of CT slices are generated for each CTA study and manual reading is laborious, time consuming and complicated by many PE look-alikes such as respiratory motion artifacts, lymph nodes, and vascular bifurcation, among many others. Hence, CAD systems have been developed to assist radiologists in detecting and characterizing emboli. In this experiment, we collected the PE CAD data from 193 image volumes comprising 863 positive candidates referring to 435 pulmonary emboli, which are marked and checked by expert radiologists as ground truth, and 7,722 negative candidates. They were randomly split into two datasets: training (255 PE with 520 positive candidates and 4,514 negative candidates) and testing (180 PE with 343 positive candidates and 3,208 negative candidates). Each candidate was represented by 90 features [8].

Colorectal cancer is the third most common cancer in both men and women. It is estimated that over 140,000 cases of colon and rectal cancer are diagnosed in the US, with more than 50,000 people dying from colon cancer annually [7]. It is critical to identify and remove lesions, or polyps (see Fig. 3(b)) when colon cancer is still in an early stage. If it progresses to an advanced stage, survival rates are very low. CT colonography is emerging as a new procedure to help in early diagnosis of colon polyps. However, hundreds of CT slices for each patient makes manual reading demanding and time-consuming. Here, we have randomly partitioned training data from 340 image volumes comprising 243 positive candidates associated with 198 colon polyps and 67,145 negative candidates; and testing data from 395 image volumes comprising 265 positive candidates corresponding to 197 colon polyps and 79,057 negative candidates. A total of 236 features are extracted for each candidate.

5.1.1 Classifier Design

We randomly set aside 20% training data as validation data for parameter tuning, such as regularization parameter γ in 1-norm SVM and thresholds at each stage in cascade AdaBoost. When the tuning process is over, *i.e.*, all the parameters are set, the entire training data is used to train the classifier again. For the cascade AdaBoost, each stage classifier is trained in the presence of all features and the number of features used is increased in a greedy manner un-



(a) (b)

Figure 3. (a) The pulmonary emboli appears as dark region residing in bright vessel lumen; (b) The colon polyp is attached to the surface of large intestine like a bulb.

til no further improvement in the classification performance on the validation data is observed. The decision thresholds are set to satisfy the target false positive rate of 70%, *i.e.*, to discard at least 30% negative candidates at each stage. On the other hand, for the proposed cascaded MIL SVM, the features are grouped into different tiers based on their computational cost and each stage classifier is trained with feature groups from the simplest to the most complex tier. As shown above, no fixed thresholds need to be fixed in the training stage. In accordance with our AND-OR cascade framework, the minimum output of all stage classifiers is used in testing which implies that all these classifiers take the same decision threshold.

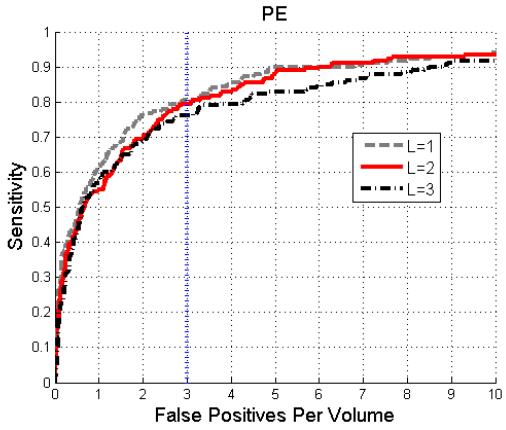


Figure 4. The ROC curves of the proposed approach with different number of layers L for PE CAD.

5.1.2 Results

First, we experiment with different numbers of stages, L, in our cascade framework to examine its impact on the overall classification performance. Fig. 4 shows the ROC curves of PE CAD with L from 1 to 3, where the performance drops clearly with the increase of L. Although no theoretical justification is derived, it is more likely to harm the system robustness when the number of classifiers in the cascade grows and accordingly the learning model becomes more complicated. To simplify the training and achieve reasonable trade-off between performance and speed, we choose to use two layers in our cascade MIL framework and cascade AdaBoost as well for fair comparison in the following experiments.

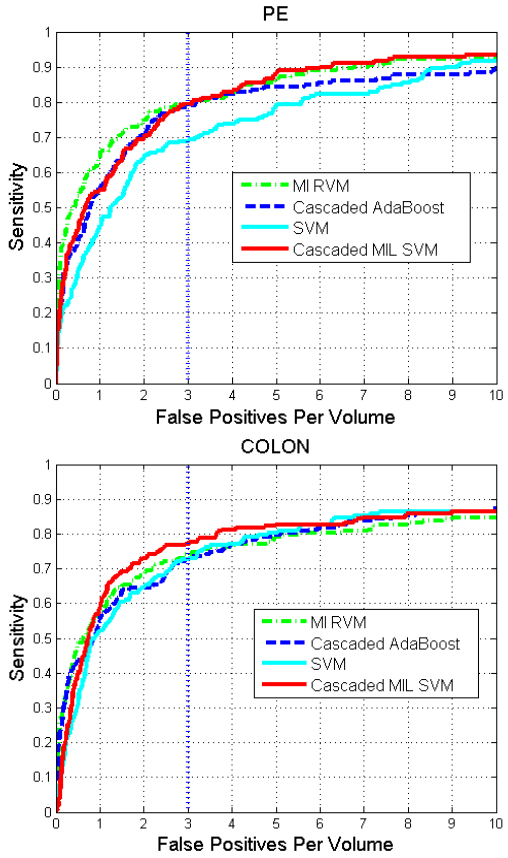


Figure 5. The ROC curves of the different approaches for (*top*) PE CAD, (*bottom*) COLON CAD.

The overall classification performance obtained by the different classifiers are compared in Fig. 5. In both CAD experiments, the proposed classifier achieves better or comparable performance than the other methods. In particular, it reaches higher or equivalent detection rates at three false positives per volume, which is an important measure of clinical interest. However, our classifier uses far fewer features, resulting in a system that runs faster than all the other methods including the cascade AdaBoost, as shown in Table 1. Taking PE CAD for instance, our approach only uses 7 features in the first stage and an additional 21 features in the second stage, which are only required for a small portion (1,108) of the total 3,551 testing samples that pass through the first stage, compared with 76 features by SVM, 40 features by MI RVM, 24 and 29 features by two stages of cascade AdaBoost with more samples (1,377) progressing to the second stage.

Table 1. The results obtained by the different approaches on (a) PE CAD, (b) COLON CAD. Numbers are given for every stage if the approach is cascaded.

(a)

ALGORITHM	NUMBER OF SELECTED FEATURES (OUT OF 90)		NUMBER OF CLASSIFIED CANDIDATES		AVERAGE CPU TIME PER VOLUME (MIN.)	
	L=1	L=2	L=1	L=2	L=1	L=2
SVM	76	/	3,551	/	1.57	/
MI RVM	40	/	3,551	/	0.83	/
CAS ADABOOST	24	29	3,551	1,377	0.50	0.23
CAS MIL SVM	7	21	3,551	1,108	0.14	0.14

(b)

ALGORITHM	NUMBER OF SELECTED FEATURES (OUT OF 236)		NUMBER OF CLASSIFIED CANDIDATES		AVERAGE CPU TIME PER VOLUME (MIN.)	
	L=1	L=2	L=1	L=2	L=1	L=2
SVM	27	/	79,322	/	26.64	/
MI RVM	53	/	79,322	/	46.99	/
CAS ADABOOST	21	16	79,322	26,466	16.17	4.12
CAS MIL SVM	4	18	79,322	64,512	1.34	15.11

5.2. MIL benchmark data

The musk datasets, MUSK1 and MUSK2, consist of descriptions of molecules (bags) using multiple low energy conformations (instances). The goal is to predict whether a molecule will bind to a target protein. Each conformation is represented by a 166-dimensional feature vector derived from surface properties. MUSK1 and MUSK2 contain around 6 and 60 conformations, respectively, per molecule on average. However, only one conformation can actually bind with the target for each molecule.

The image datasets, FOX, TIGER and ELEPHANT, contain images (bags) where each image is decomposed into a set of segments (instances), each characterized by color, texture and shape descriptors. The task is to identify images containing the objects of interest, such as fox, tiger or elephant, which might be contained in more than one segment.

5.2.1 Classifier Design

In this experiment, the classifiers are evaluated using 10-fold cross validation: the training data is randomly divided into 10 subsets on bag level, $i.e.$, the instances in the same bag are kept in one subset, then we combine the classification results of 10 runs to plot the complete ROC curves. For the proposed cascaded MIL SVM, because there's no prior knowledge of the computational cost or exact definition of these features, we simply split the features into two groups with the first half features in one group and the second half in the other group. The other settings are the same as Subsection 5.1.1.

5.2.2 Results

Fig. 6 shows the ROC curves of different approaches with benchmark datasets. Note that in order to construct a complete ROC curve for the cascade AdaBoost, we have to decrease the fixed thresholds at each stage one by one from the last stage to the first, as described in [15]. The results show that our approach outperforms the other approaches significantly in most of the datasets, even though it is not specifically designed for balanced MIL problems where the numbers of positive and negative bags are equivalent.

6. Conclusion

Motivated by CAD applications which can be modeled as MIL problems with extremely unbalanced data and stringent real time requirements, we propose a novel approach to integrate MIL and cascade classification in one framework to improve both computational efficiency and classification performance. The experiments show that the proposed algorithm can significantly reduce the computational cost with comparable prediction accuracy on real world CAD problems, and yield higher accuracy on benchmark datasets, compared with other competitive MIL or cascade classification methods. Ongoing work will include the study of how to separate features into different groups for each classifier in the cascade to gain better classification performance, and we also conjecture that better optimization techniques, *e.g.*, to avoid local minima, may further improve the classification accuracy.

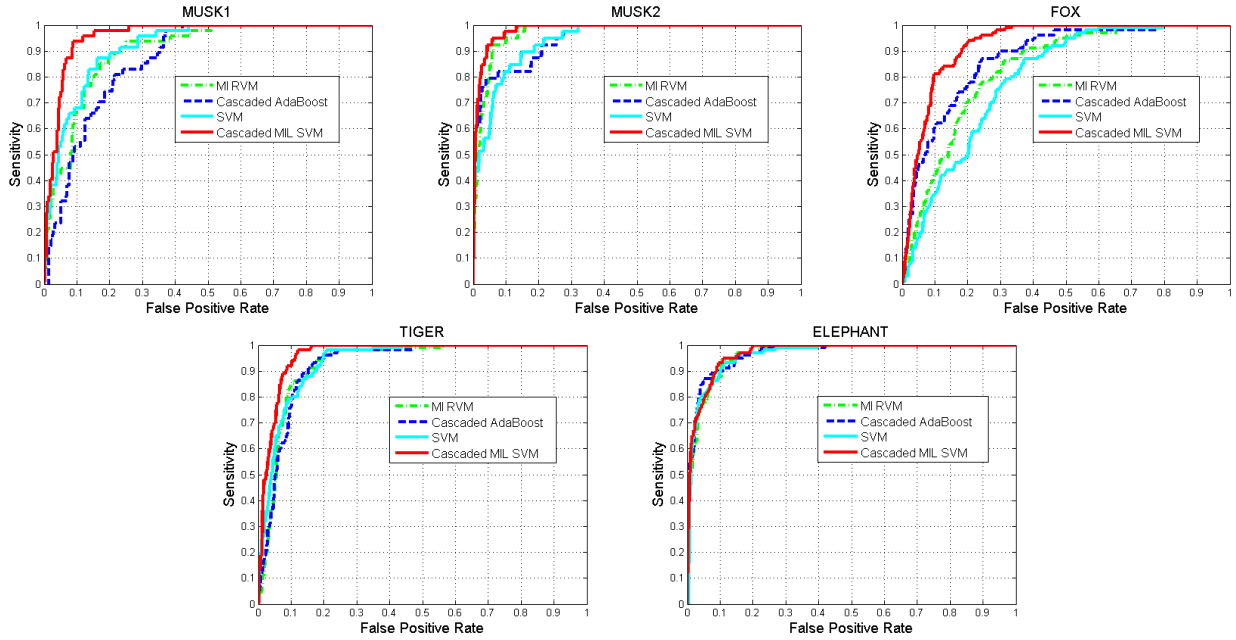


Figure 6. The ROC curves of the different approaches applied on benchmark MIL data.

References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. *Advances in neural information processing systems*, 15, 2002.

[2] J. Bi and J. Liang. Multiple instance learning of pulmonary embolism detection with geodesic distance along vascular structure. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[3] L. Bourdev and J. Brandt. Robust object detection via soft cascade. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:236–243, 2005.

[4] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence Journal*, 89:31C71, 1997.

[5] M. Dundar and J. Bi. Joint optimization of cascaded classifiers for computer aided detection. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2007*, 2007.

[6] G. Fung, M. Dundar, B. Krishnapuram, and R. B. Rao. Multiple instance learning for computer aided diagnosis. *Advances in neural information processing systems*, 19:425–432, 2007.

[7] D. Jemal, R. Tiwari, T. Murray, A. Ghafoor, A. Saumuels, E. Ward, E. Feuer, and M. Thun. Cancer statistics, 2004.

[8] J. Liang and J. Bi. Local characteristic features for computer aided detection of pulmonary embolism in ct angiography. *Proceedings of Pulmonary Image Analysis at Annual Conference on Medical Image Computing and Computer-Assisted Intervention*, 2008.

[9] O. Maron and T. Lozano-Perez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1998.

[10] V. C. Raykar, B. Krishnapuram, J. Bi, M. Dundar, and B. Rao. Bayesian multiple instance learning: automatic feature selection and inductive transfer. *Proceedings of the 25th International Conference on Machine Learning*, pages 808–815, 2008.

[11] H. Schneiderman. Feature-centric evaluation for efficient cascaded object detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 2:29–36, 2004.

[12] B. Settles, M. Craven, and S. Ray. Multiple instance active learning. *Advances in neural information processing systems*, 20:1289–1296, 2008.

[13] R. Tibshirani. Regression selection and shrinkage via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288, 1996.

[14] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, pages 475–494, 2001.

[15] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, May 2004.

[16] P. Viola, J. C. Platt, and C. Zhang. Multiple instance boosting for object detection. *Advances in neural information processing systems*, 18:1417–1424, 2006.

[17] R. Xiao, L. Zhu, and H. J. Zhang. Boosting chain learning for object detection. *Ninth IEEE Conference on Computer Vision and Pattern Recognition*, 1:709–715, 2003.

[18] Q. Zhang and S. A. Goldman. Em-dd: An improved multiple-instance learning technique. *Advances in Neural Information Processing Systems*, 14, 2002.