Coarse-to-Fine Classification via Parametric and Nonparametric Models for Computer-Aided Diagnosis Paper ID 787

ABSTRACT

Classification is one of the core problems in Computer-Aided Diagnosis (CAD), targeting for early cancer detection using 3D medical imaging interpretation. High detection sensitivity with desirably low false positive (FP) rate is critical for a CAD system to be accepted as a valuable or even indispensable tool in radiologists' workflow. Given various spurious imagery noises which cause observation uncertainties, this remains a very challenging task. In this paper, we propose a novel, two-tiered coarse-to-fine (CTF) classification cascade framework to tackle this problem. We first obtain classification-critical data samples (e.g., samples on the decision boundary) extracted from the holistic data distributions using a robust parametric model (e.g., [35]); then we build a graph-embedding based nonparametric classifier on sampled data, which can more accurately preserve or formulate the complex classification boundary. These two steps can also be considered as effective "sample pruning" and "feature pursuing + kNN/template matching", respectively. Our approach is validated comprehensively in colorectal polyp detection and lung nodule detection CAD systems, as the top two deadly cancers, using hospital scale, multi-site clinical datasets. The results show that our method achieves overall better classification/detection performance than existing state-of-the-art algorithms using single-layer classifiers, such as the support vector machine variants [45], boosting [40], logistic regression [33], relevance vector machine [35], k-nearest neighbor [30] and sparse projections on graph [6].

Categories and Subject Descriptors

Industrial and Application Paper [**Knowledge Management (KM)**]: Classification and Clustering, Data pre- and post-Processing, Large-scale statistical techniques

Keywords

Cancer lesion classification, coarse-to-fine classification, class regularized graph embedding, feature selection, nearest neighbor voting, template matching

1. INTRODUCTION

Colon cancer and lung cancer are the top two leading causes of cancer deaths in western population. Meanwhile, these two cancers are also highly preventable or "curable" if detected early. Image interpretation based cancer detection via 3D computer tomography has emerged as a common clinical practice, and many computer-aided detection tools for enhancing radiologists' diagnostic performance and effectiveness are developed in the last decade [16, 30, 33, 40, 45]. The key for radiologists to accept the clinical usage of a computer-aided diagnosis (CAD) system is the high detection sensitivity with reasonably low false positive (FP) rate per case.

CAD system generally contains two stages: *Image Processing* as extracting sub-volumes of interest (VOI) by heuristic volume parsing, and informative feature attributes describing the underlying (cancerous) anatomic structures; *Classification* as deciding the class assignment (cancer, or non-cancer) for selected VOIs by analyzing features. VOI selection is also called *candidate generation*, or CG, to rapidly identify possibly anomalous regions with high sensitivity but low specificity, e.g., > 100 candidates per scan with $1 \sim 2$ true positives. Then dozens or hundreds of heterogeneous image features can be computed per VOI, in domains of volumetric shape, intensity, gradient, texture and even context [16, 30, 33, 40, 45]. Last, the essential goal for classification is to achieve the best ROC (Receiver Operating Characteristic) or FROC (Free-Response Receiver Operating Characteristic) analysis on testing dataset, to balance the criteria of sensitivity and specificity, given VOIs and associated features.

This paper mainly focuses on the classification aspect of CAD. We propose and comprehensively evaluate a novel coarse-to-fine classification framework. The method consists of the following two steps, in both training and testing. (1) *Sample Pruning:* Parametric classification models (e.g., logistic regression [33], boosting [40], support/relevance vector machines [3, 35]) are trained on the complexly distributed datasets as coarse, distribution-level classification. The goal is not to assign class labels, but to prune data samples to select more "classification-critical" candidates, which are expected to preserve the decision boundary in the high dimensional feature space (thus vast numbers of samples lying far from classification boundary are discarded [1]). (2) *Feature Pursuing + kNN/Template Matching:* We first apply

[1]This is related with using nearest neighbor analysis to

feature selection and graph embedding methods jointly to find intrinsic lower dimensional feature subspace that preserves group-wise data topology, and then employ nonparametric classifiers for final classification, using kNN or template matching. We argue that more precisely modeling the intrinsic geometric of decision boundary, by graph embedding and nonparametric classifiers in a finer level, can potentially improve the final classification performance. The overall process is illustrated as follows

Samples → **Sample pruning** → **Feature selection** → **Class regularized graph embedding** → **kNN/Template matching**

We applied the proposed framework on colon polyp and lung nodule detection, using two large scale clinical datasets collected from multiple clinical sites across continents. Classification in these two CAD problems is very important, but also challenging due to the large within-class variations (for polyps/nodules in different morphological subcategories, spatial contexts and false positives resulted from various anatomic structures, such as tagged stool, ileo-cecal valve, extra-colonic finding and rectal catheter or balloon for colon polyp detection, and pathology, vessel, vessel junction, fissure, scar tissue and so on for lung nodule detection). The low-level imagery data were extracted and presented as the intermediate-level heterogeneous natured features for the classification task (as special cases of image based object recognition). The results show that the proposed framework significantly outperformed the baseline CAD system using the same set of input image features, and compared favorably with other state-of-the-arts.

The rest of the paper is organized as follows. In Section 2 we present (data) sample pruning using a linear parametric model of Relevance Vector Machine Multiple Instance Learning (RVMMIL) [35]. Section 3 describes the Maximum Relevance Minimum Redundancy (MRMR) based feature selection and our modified graph embedding method for stratified optimization of dimension reduction and manifold projection. The strategy of integrating sparsity into graph embedding is also addressed and compared in section 3. This is followed by k nearest neighbor (kNN) voting and t-center [44] based template matching techniques for classification in Section 4. Then we perform extensive experimental evaluation using our coarse-to-fine classification diagram on both colon polyp and lung nodule classification applications in Section 5. Finally we conclude the paper in Section 6 with discussion.

2. SAMPLE PRUNING USING PARAMETRIC RVMMIL

find data samples either near the decision boundary [41] or in local neighborhoods [48], then training SVM classifiers on reduced or clustered datasets. However we perform sample pruning by selecting data upon their classification scores/confidences of a learned parametric model that is well studied, more robust and stable, compared with nearest neighbor (NN) clustering method, especially in high dimensional space. For example, the neighborhood size selection and defining sensible distance measure problems in NN are non-trivial.

We start by developing a "coarse" classifier for sample pruning using a parametric model. Considering the specific characteristics of CAD classification problems, in this paper we use the RVMMIL approach [35].

Relevance vector machine (RVM) is a supervised Bayesian machine learning approach that estimates the classifier parameters by maximizing the likelihood in a probabilistic setting. A powerful variation/extension has been proposed [35] to integrate feature selection and handle multiple instance learning (MIL) problems which is essential for CAD applications. The output of RVMMIL is a linear logistic regression model on a reduced set of features, and gives a class prediction with probability or confidence for any single instance.

In RVMMIL, the probability for an instance $\mathbf{x}_i$ to be positive is $p(y=1|\mathbf{x}_i) = \sigma(\mathbf{a}'\mathbf{x}_i)$, where σ is the logistic function defined as $\sigma(t) = 1/(1+e^{-t})$ and $\mathbf{a}'\mathbf{x}_i$ is the linear dot-product between data feature vector $\mathbf{x}_i$ and model coefficient vector $\mathbf{a}$. Therefore, the probability for a bag or set $\mathcal{X} = \{\mathbf{x}_i\}$ to be positive is $p(y=1|\mathbf{x}) = 1 - \prod_{\mathbf{x}_i \in \mathcal{X}} (1 - p(y = 1|\mathbf{x}_i))$. Given the training dataset $T = (\mathcal{X}, \mathbf{y})$, $\mathcal{X}$ is the set of training bags of multiple instances with label $\mathbf{y}$. The RVMMIL utilizes the *maximum a-posterior* (MAP) estimate based on T to find the optimal parameter a such that

$$\begin{aligned} a &= \arg\max_{\tilde{a}} p(\tilde{a}|T) = \arg\max_{\tilde{a}} p(T|\tilde{a})p(\tilde{a}) \\ &= \arg\max_{\tilde{a}} \sum_i \mathbf{y}_i \log p_i + (1-\mathbf{y}_i)\log(1-p_i) + \log p(\tilde{a}), \end{aligned} \tag{1}$$

where $p_i = p(\mathbf{y}_i = 1|\mathbf{x}_i, \tilde{a})$ and $p(\tilde{a})$ is the prior which can be assumed to be Gaussian. In this case, (1) can be easily solved using Newton-Raphson method [35]. For more details, we refer the readers to [35].

In our coarse-to-fine classification model, RVMMIL is used as the coarse-level cascade classifier for sample pruning, i.e., we will remove samples x_i satisfying $p(y=1|\mathbf{x}_i) < \hat{\rho}$. This step can eliminate massive amount of negatives without effecting much on sensitivity, by choosing a balanced $\hat{\rho}$. The remained data samples $p(y=1|\mathbf{x}_i) \geq \hat{\rho}$ are either true positives (at high recall) or "hard" false positives, lying close to the classification boundary, which largely impact the final classification accuracy. Note that other classifiers with confidence estimates, as boosting [40] and regularized SVM [45], are also applicable.

3. FEATURE PURSUIT VIA SELECTION & GRAPH EMBEDDING

The basic idea of feature pursuit is to estimate intrinsic, lower dimensional feature subspace of data for nonparametric classification, while preserving generative data-graph topology. This is the key to achieve superior classification performance with simple nonparametric classifiers. In the proposed framework it consists of two steps: supervised feature selection and class regularized graph embedding.

3.1 Feature Selection

Feature selection, also as known as variable selection, is a machine learning scheme to search and extract a subset of relevant features so that a desirable objective of model complexity/effectiveness can be optimized. It essentially has exponential combinatorial complexity in feature cardinality, if doing exhaustive search. By applying feature selection,

only a compact subset of highly relevant features is retained, to simplify the later graph embedding or feature projection process and make it more effective. There are many feature selection techniques in the literature [4, 5, 8, 19, 21, 25, 46, 49]. In this work, we use Maximum Relevance Minimum Redundancy (MRMR) feature selection [31], which can give a very good representative feature set with a fixed number of selected features, or the least amount of relevant features to achieve the same accuracy level (as original feature set). Moreover, MRMR is very efficient in computation and storage.

The relevance in MRMR is measured using a variant of *Pearson coefficient* [37]. For any two variables f and $\tilde{f}$, the *Pearson* coefficient γ between them is

$$\gamma(f,\tilde{f}) = \frac{|\mathbf{Cov}(f,\tilde{f})|}{\sqrt{\mathbf{Var}(f)\mathbf{Var}(\tilde{f})}}, \quad \text{where } \mathbf{Cov}(f,\tilde{f}) = \mathbf{E}[(f-\mathbf{E}[f])(\tilde{f}-\mathbf{E}[\tilde{f}])], \tag{2}$$

$\mathbf{E}[\cdot]$ is the expectation and $\mathbf{Var}(\cdot)$ represents the variance. Given a set of features $\mathbb{F} = \{f_i\}$, its MRMR feature subset $\mathbb{H}$ maximizes the following objective κ:

$$\kappa(\mathbb{H},\mathbf{y}) = \gamma(\mathbb{H},\mathbf{y}) - \gamma(\mathbb{H}), \tag{3}$$

where

$$\gamma(\mathbb{H}) = \frac{1}{m^2}\sum_{f_i,f_j\in\mathbb{H}}\gamma(f_i,f_j), \tag{4}$$

$$\gamma(\mathbb{H},\mathbf{y}) = \frac{1}{m}\sum_{f_i\in\mathbb{H}}\gamma(f_i,\mathbf{y}), \tag{5}$$

and m is the total number of elements in $\mathbb{H}$. Starting from $H_0 = \emptyset$, we select f_i by

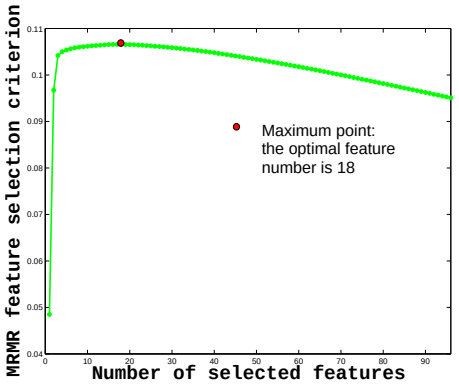
$$f_i = \arg\max_{f\in\mathbb{F}-\mathbb{H}_{i-1}} \gamma(f,\mathbf{y}) - \frac{1}{i-1}\sum_{f_j\in\mathbb{H}_{i-1}}\gamma(f,f_j) \tag{6}$$

Then set $H_i = H_{i-1}\cup f_i$. This is repeated until $\kappa(\mathbb{H}_{i-1},\mathbf{y}) \geq \kappa(\mathbb{H}_i,\mathbf{y})$. where $\mathbb{H}_{i-1}$ reaches optimum. Using this method, we select 18 out of 96 features for the colon dataset, and 23 out of 120 features for the lung nodule dataset. The objective plots are shown in Fig. 1.

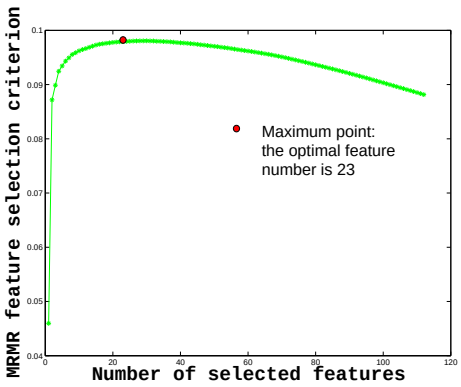
3.2 Class Regularized Graph Embedding

Nonparametric classifiers, such as nearest neighbor (NN) or (t-center [44]) template matching (TM), are flexible and powerful representations for joint classification, clustering and retrieval. However they are very sensitive to high dimensional feature space. In this section, we exploit *Class Regularized Graph Embedding* (CRGE) to project data (after feature selection) into an even lower dimensional subspace, where data samples from the same class getting closer and samples from different classes moving apart, to make NN or TM more robust and semantically interpretable, as shown later.

Graph embedding is a special class of dimension reduction method that uses linear or nonlinear projections. Feature projections can be learned in different ways: minimizing the reconstruction error as in principal component analysis (PCA) [12, 23]; preserving distances in the original space, e.g. multidimensional scaling (MDS) [11] and ISOMAP [43]; maximizing class-data separation as linear



(a)



(b)

Figure 1: The number of selected features versus the MRMR feature selection criterion in Eq. (3) on colon polyp (a) and lung nodule (b) datasets.

discriminant analysis (LDA) [12], or retaining the linear relationship between local neighbors, e.g., neighborhood component analysis (NCA) [18], locally linear embedding (LLE) [38]. We follow *the principle that keeps the locality of nearby data and maps apart data further*, in the graph-induced subspace, which is similar to Laplacian Eigenmap [2, 7] and Locality Preserving Projection [22].

Given a set of N points $\mathcal{X} = \{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N\} \subset \mathbb{R}^n$, and a symmetric $N\times N$ matrix W which measures the similarity between all pairs of points in $\mathcal{X}$. The set $\mathcal{X}$ and matrix W compose a graph $\mathcal{G}$, with $\mathcal{X}$ as vertices and W as weights of the edges. The conventional graph embedding method will map $\mathcal{X}$ to a much lower dimensional space $\mathcal{Y} = \{\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_N\} \subset \mathbb{R}^{\tilde{n}}$, $\tilde{n} \ll n$. The optimal $\mathcal{Y}$ should minimize the loss function $L(\mathcal{Y})$ which is defined as

$$L(\mathcal{Y}) = \sum_{i,j}\|\mathbf{y}_i - \mathbf{y}_j\|^2 W_{ij}, \tag{7}$$

under some appropriate constraints. This objective function ensures $\mathbf{y}_i$ and $\mathbf{y}_j$ to be close if $\mathbf{x}_i$ and $\mathbf{x}_j$ are close and W_ij is large. Though performed well in many applications [7, 22], the limitation of Eq. (7) is that it does not penalize the similarity between points belonging to different classes. One more comprehensive strategy is to simultaneously maximize the similarity between data pairs of the same class and minimize the similarity between two points rooted from different

classes. In other words, we optimize on mapping the same class data to proximity subspaces, while projecting different class data samples to be far apart, explicitly.

The goal of class regularized graph embedding is to find a mapping $\phi : \mathcal{X} \mapsto \mathcal{Y}$, such that ϕ minimizes the function $E(\mathcal{Y})$ defined as

$$E(\mathcal{Y}) = \sum_{i,j\in\mathcal{S}} \|\mathbf{y}_i - \mathbf{y}_j\|^2 W_{ij} - \sum_{i,j\in\mathcal{D}} \|\mathbf{y}_i - \mathbf{y}_j\|^2 W_{ij}\,, \quad \text{subject to: } \|\mathcal{Y}\|_F = 1. \tag{8}$$

where $i,j \in \mathcal{S}$ means $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same class, and $i,j \in \mathcal{D}$ means $\mathbf{x}_i$ and $\mathbf{x}_j$ are in different classes. $\|\cdot\|_F$ is the Frobenius norm. To avoid notation clutter, we rewrite (8) and get

$$\min \sum_{i,j} \|\mathbf{y}_i - \mathbf{y}_j\|^2 W_{ij} H_{ij}\,, \tag{9}$$

where H_{ij} is the Heaviside function and

$$H_{ij} = \begin{cases} 1, & \text{if } i,j \in \mathcal{S} \\ -1, & \text{if } i,j \in \mathcal{D} \end{cases}.$$

Various choices of the mapping function ϕ have been proposed recently, e.g. linear mapping, kernel mapping and tensor mapping [47]. We use linear mapping because of its simplicity and generality [8]. A linear mapping function ϕ is described as

$$\mathbf{y} = \phi(\mathbf{x}) = M'\mathbf{x},\ M \in \mathbb{R}^{n\times\tilde{n}}\,, \tilde{n} \ll n\,. \tag{10}$$

Plugging (10) into (9), we get

$$\min_M \sum_{i,j} \|M'\mathbf{x}_i - M'\mathbf{x}_j\|^2 W_{ij} H_{ij}\,, \quad \text{subject to: } \|M\|_F = 1\,, \tag{11}$$

where the constraint $\|M\|_F = 1$ eliminates the scaling effect. Eq. (11) can be solved very quickly using gradient descent technique along with iterative projections [36]. The reduced dimension $\tilde{n}$ is determined when the loss function (8) is minimized by varying $\tilde{n}$. Though some other ways are possible.

The computation of W can be done in the following manners, which correspond to different dimension reduction methods as LLE [38], ISOMAP [43], and Laplacian Eigenmap [2, 7].

$$W(i,j) = \begin{cases} 1, & \text{if } i,j \in \mathcal{S} \\ 0, & \text{if } i,j \in \mathcal{D} \end{cases}; \tag{12}$$

$$W(i,j) = \exp\{-\alpha\|\mathbf{x}_i - \mathbf{x}_j\|^2\}\,, \alpha > 0\,; \tag{13}$$

$$W(i,j) = \exp\{-\alpha(\mathbf{x}_i - \mathbf{x}_j)'A(\mathbf{x}_i - \mathbf{x}_j)\}\,, \quad \alpha > 0, A \text{ is a PSD matrix}\,; \tag{14}$$

$$W(i,j) = \mathbf{x}_i'\mathbf{x}_j / \|\mathbf{x}_i\|\|\mathbf{x}_j\|\,. \tag{15}$$

Eq. (12) is the simplest weighting scheme, where $W(i,j) = 1$ if and only if $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same class. However this scheme might lose information about the affinity between the nodes belonging to different classes. Eq. (13) is the heat kernel weighting method, which has an intrinsic connection to the Laplace Beltrami operator on differentiable functions on a manifold [1]. Eq. (14) is related to the Mahalanobis distance between two vectors. Eq. (15) is the dot product weighting scheme, which measures the cosine similarity of the two vectors and is easy to compute. For our CAD purpose of cancer lesion classification, Eq. (12) neglects the similarity between negative and positive samples, which invalidates the penalization about the similarity between samples from different classes; Eq. (13) and (14) are not suitable because they both use Euclidean or Mahalanobis similar distance assumption, which holds when the data samples lie in a (locally) Euclidean space. From our empirical observation, this assumption does not apply to colon polyp or lung nodule dataset. Furthermore, Eq. (13) and (14) bother to tune the parameters α or A which may be sensitive for the similarity calculation. Thus we use (15) for its appropriateness and computation efficiency.

The effectiveness of dimension reduction can be evaluated according to several criteria, e.g., information gain [10], Pearson coefficients [37] and Fisher score [14]. We validate the effectiveness of our proposed dimension reduction technique using *Fisher Score* (FS) [14] on both polyp colon and lung nodule datasets. The class separability between negatives and positives is measured via Fisher's linear discriminant [14]. Let the covariance matrices of the negatives and positives be Σ_- and Σ_+, and the means of the negatives and positives be μ_- and μ_+, then the Fisher linear discriminant of the binary classes is

$$s = (\mu_+ - \mu_-)'(\Sigma_+ + \Sigma_-)^{-1}(\mu_+ - \mu_-)\,, \tag{16}$$

where the larger s is, the more statistically distinguishable negative-positive class distributions will be. CRGE is capable to increase the discriminant between positive and negative lesions in the projected feature subspaces, visually and numerically. This is validated on the colon polyp and lung nodule datasets. For comparison, we plot the first three MRMR selected original features and the first three projected dimensions after CRGE, on (testing) colon polyp and lung nodule datasets in Fig. 2. The Fisher (linear discriminant) score for the first three MRMR selected features on the colon polyp dataset is 0.2725, and after CRGE, the score improves to 0.7990. For the lung nodule dataset, the score increases from 0.1083 to 0.6987, reflecting the impact of CRGE. The numerical results demonstrate that our class regularized graph embedding technique indeed enlarges the class separability between negative and positive populations, for both datasets. Note that many dimension reduction methods are tested using image data where each dimension is a pixel or voxel, for classification [7, 22] and registration [20]. As mentioned above, CAD image features are extremely heterogeneous attributes as measuring different nature imaging properties for 3D VOI structures, in different metrics or dimensions.

3.3 Sparse Graph Embedding

As a companion to the above stratified "feature pursuing" strategy of *feature selection + graph embedding*, an integrated approach is Sparse (feature) Projections over Graph (SPG) [6, 8]. SPG utilizes techniques from graph theory [9] to construct an affinity graph over the data and assumes that the affinity graph is usually sparse (e.g. nearest neighbor graph). Thus the embedding results can be efficiently computed. After this, lasso regression [13] is applied to obtain the sparse basis functions. The data in the reduced subspace is represented as a linear combination of a *sparse subset* consisting of the most relevant features, rather than using all features as in PCA, LDA or regular graph embedding. Feature selection and graph embedding based dimen-

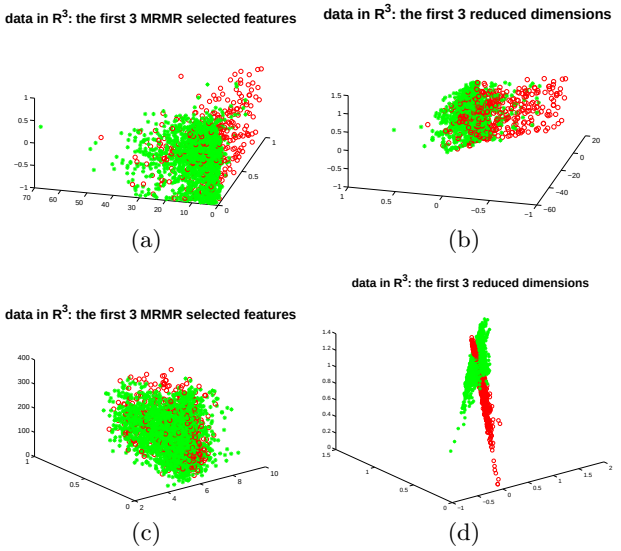


Figure 2: Plot of the data samples (testing) according to the first three features selected by MRMR (a) and the first three dimensions from graph embedding (b) on the colon polyp dataset. Similarly, (c) and (d) are illustrated based on the lung nodule dataset. The dimension coordinates on the figures are not directly comparable.

sion reduction are jointly presented and formulated within the same optimization framework.

The SPG algorithm is described as follows. Given a set of N points $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \subset \mathbb{R}^n$, the goal of SPG is to find a transformation matrix $A = (\mathbf{a}_1, \cdots, \mathbf{a}_{\tilde{n}})$ that maps the N points to a set of lower dimensional points $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N\} \subset \mathbb{R}^{\tilde{n}}$, $\tilde{n} \ll n$. For each i, $\mathbf{y}_i (= A'\mathbf{x}_i)$ is the projection of $\mathbf{x}_i$ onto the lower dimensional space $\mathbb{R}^{\tilde{n}}$. Furthermore, there is a sparsity constraint on each projection $\mathbf{a}_i$, and $\|\mathbf{a}_i\|_0 < k$ ($k < n$), where $\|\mathbf{a}\|_0$ is defined as the number of nonzero entries of $\mathbf{a}$. To obtain the optimal projection, one first needs to create a graph G with affinity matrix W over $\mathcal{X}$, and then minimize the following energy function

$$\begin{aligned} &\min_{\mathbf{a}} \sum_{i,j} (\mathbf{a}'\mathbf{x}_i - \mathbf{a}'\mathbf{x}_j)^2 W_{ij} \\ &\text{subject to: } \mathbf{a}'XDX'\mathbf{a} = 1\,, \\ &\|\mathbf{a}\|_0 \le k\,, \end{aligned} \tag{17}$$

where $X = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$, D is a diagonal matrix and each entry of the diagonal is the sum of the corresponding row of W, i.e., $D_{ii} = \sum_j W_{ij}$. Since Eq. (17) is NP-hard, it is split into two steps. The first step introduces the graph Laplacian matrix [9] $L = D - W$, and the optimization function in Eq. (17) can be reformulated as

$$\frac{1}{2}\sum_{i,j} (\mathbf{a}'\mathbf{x}_i - \mathbf{a}'\mathbf{x}_j)^2 W_{ij} = \mathbf{a}'XLX'\mathbf{a} \tag{18}$$

The solution to (18) with the first constraint in (17) leads to

$$XLX'\mathbf{a} = \lambda XDX'\mathbf{a}. \tag{19}$$

Once obtaining the embedding $y_i = \mathbf{a}'\mathbf{x}_i$, lasso regression can be applied to get the sparse transformation according to the following minimization

$$\min_{\tilde{\mathbf{a}}} \left(\sum_{i=1}^{N} (y_i - \tilde{\mathbf{a}}'\mathbf{x}_i)^2 + \beta \|\tilde{\mathbf{a}}\|_1 \right). \tag{20}$$

After learning the sparse transformation $\mathbf{a}$, we can project all the samples into the lower and more intrinsic dimensional space, in which we can perform classification. SPG, in some sense, integrates the feature selection and dimension reduction processes, which has been shown to be effective in many applications, such as text clustering [6] and classification on many benchmark machine learning datasets [8]. *However, we argue that our stratified approach which prunes non-informative or redundant features from an information-theoretic aspect before graph embedding or feature projection, can simplify the optimization process of graph embedding on a reduced feature set. This strategy may achieve better overall results, compared from the holistic sparsity-constrained graph embedding (as SPG).* The sparse approximation after embedding (i.e., Eq. (20)) is also suboptimal. In practice, superior classification performances over two hospital scale, clinical datasets are demonstrated using our stratified feature pursuit framework, in later experimental section.

4. NONPARAMETRIC CLASSIFICATION

After finding the mapping ϕ and $\mathcal{Y}$, we will perform unsupervised clustering on $\mathcal{Y}$ for training negatives and positives separately. Data samples of the same class are divided into local clusters, where instances from the same cluster are more similar than those from different clusters. Each cluster is then represented using a template. Based on the kNN voting of the cluster templates, each instance in testing is labeled. We explain the details of clustering and template calculation in this section.

4.1 Clustering & Templates

The clustering process is performed according to a recently introduced clustering algorithm, namely total Bregman divergence clustering [26]. This algorithm utilizes the newly proposed divergence measure first presented in [44]. This divergence measure, called total Bregman divergence (tBD), is based on the orthogonal distance between the convex generating function of the divergence and its tangent approximation at the second argument of the divergence. tBD is naturally robust and leads to efficient algorithms for soft and hard clustering. For more details, we refer the readers to [26, 44].

We employ the total Bregman divergence hard-clustering algorithm [26] to saparate negative or positive data instances, in $\mathcal{Y}$ space. Denote that c_1 clusters, with the cluster centers $\{z_{i-}\}_{i=1}^{c_1}$, are obtained for negatives; and c_2 clusters with centers $\{z_{j+}\}_{j=1}^{c_2}$ for positives. The numbers of clusters c_1,c_2 is chosen to minimize the *intra-inter-validity index* [34], given by

$$\begin{aligned} \text{index} &= \frac{\text{intra}}{\text{inter}}, \\ \text{intra} &= \frac{1}{N}\sum_{i=1}^{c}\sum_{y\in C_i} \|y - z_i\|^2, \\ \text{inter} &= \min_{i,j} \|z_i - z_j\|^2, \end{aligned} \tag{21}$$

where C_i is the ith cluster with center z_i. Each cluster is represented as the tBD center, termed t-center [26, 44], which is the ℓ_1 norm median of all samples in the corresponding cluster. For example, if $\{\mathbf{y}_i\}_{i=1}^N$ is the set of samples, then its t-center z is

$$z = \arg\min_{\tilde{z}} \sum_{i=1}^{N} \delta_f(\tilde{z}, \mathbf{y}_i), \tag{22}$$

where δ_f is the total Bregamn divergence generated by some convex and differentiable generator function f:

$$\delta_f(\mathbf{y}_1, \mathbf{y}_2) = \frac{f(\mathbf{y}_1) - f(\mathbf{y}_2) - \langle \mathbf{y}_1 - \mathbf{y}_2, \nabla f(\mathbf{y}_2) \rangle}{\sqrt{1 + \|\nabla f(\mathbf{y}_2)\|^2}}. \tag{23}$$

Here, we use $f(y) = \|y\|^2$, and hence δ_f becomes the total square loss [26, 44] and the t-center in Eq. (22) becomes

$$z = \sum_{i=1}^{N} a_i \mathbf{y}_i \,, \text{ where } a_i = \frac{1/\sqrt{1 + 4\|\mathbf{y}_i\|^2}}{(\sum_j 1/\sqrt{1 + 4\|\mathbf{y}_j\|^2})} \,. \tag{24}$$

After learning the centers as templates, we can determine whether a given sample is positive or negative, according to the kNN voting on the set of trained positive/negative t-centers.

4.2 Template Matching via kNN Voting

Nearest neighbor voting is a popular nonparametric classifier which has been studied extensively [39]. Given a test sample $\mathbf{y}_i$, we need to find its k nearest neighbors from the t-centers. Suppose the neighbors are $\{z_1, z_2, \cdots, z_k\}$ and the corresponding distance from $\mathbf{y}_i$ to the neighbors are $\{d_1, d_2, \cdots, d_k\}$. The distance d_i can be either Euclidean distance or the vector angle difference (Euclidean distance is used in our experiments). We define the empirical probability of $\mathbf{y}_i$ being positive as p, and

$$p = \frac{\sum_{(z_j \text{ is positive})} 1/d_j}{\sum_{(z_l \text{ is negative})} 1/d_l + \sum_{(z_j \text{ is positive})} 1/d_j} \,. \tag{25}$$

Based on the p value, we can draw the FROC curve of sensitivity and FP rate per case for training and testing datasets. Eq. (25) is a soft kNN voting scheme using the reciprocal of distance $1/d_i$. There are other options to calculate p, e.g., using the counts of positive/negative t-centers. We found that t-centers are more robust as they lead to better sparsity and diversity of CAD lesion data distribution than proximity data samples (as in kNN).

The number of nearest neighbors k is chosen during the training/validation stage. Since the optimal k should lead to the best performance of our algorithm, we set k to be the one that maximizes the Area Under (the FROC) Curve (AUC) on the training dataset. On the other hand, if only a partial range of FROC has more meaningful impacts on clinical practice (e.g., $FP \in [2, 4]$ per case), we can search k to optimize the partial AUC

$$k = \arg\max_{k} \ \text{AUC}(\text{FPrate} \in [2, 4]). \tag{26}$$

5. EXPERIMENTAL RESULTS

Unlike many existing CAD systems [15, 27, 32] where small datasets are often used, our method is evaluated on large scale datasets with representative varieties, collected from dozens of hospitals across US, Europe and Asia. We perform two important clinical tasks of classifying colonic polyps and lung nodules based on 3D CT imagery features. Lung cancer and colon cancer are the two leading deadly cancers in western population.

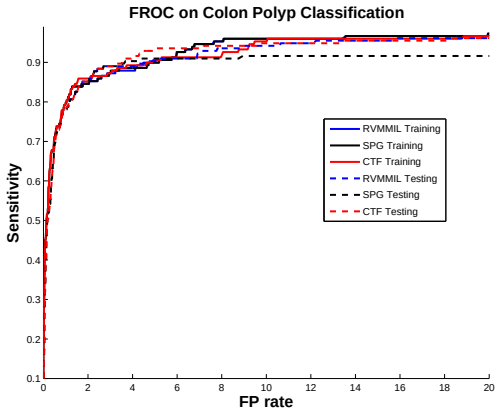
5.1 Colon Polyp Detection & Retrieval

Data: The colon polyp dataset contains 134,116 polyp candidates obtained from an annotated CT colonography (CTC) database of 429 patients. Each sample is represented by a 96-dimensional computer extracted feature vector, describing its shape, intensity pattern, segmented class-conditional likelihood statistics and other higher level features [33, 28, 40, 45]. The patients were examined from 12 hospitals via different scanners from Siemens, GE and Philips, and under various fecal-tagging imaging protocols. Each patient was scanned in two positions, resulting two (prone and supine) scans. There are 1,116 positives out of the 134,116 samples. The CAD sensitivity is calculated at per-polyp level for all actionable polyps $\geq 6mm$ (i.e., polyp is classified correctly at least from one view), and the FP rate counts the sum of two (prone-supine) scans per patient. The colon polyp dataset is split into two parts, namely training and testing dataset, both of which are split at patient level. No data from the same patient is used for both training and testing. Here, we do not employ N-fold cross validation because we intend to keep a portion of data (as our testing dataset) which is always unseen for training. This is practically critical to evaluate the more "true" or trustful performance of a clinical product. As a result, the training dataset contains all the instances detected from 216 patients, and the testing dataset includes the rest 213 patients.

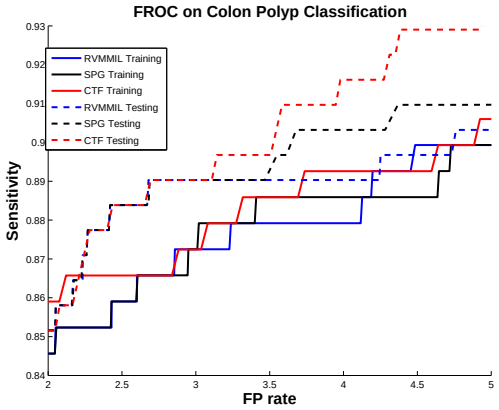
After estimating the parametric RVMMIL model [35], we get the probability (classification score) of each candidate being positive. Then we perform thresholding according to the classification scores. Let the condition on classification scores $p(y = 1|\mathbf{x}_i) \geq \hat{\rho} = 0.0157$ as a cascade with high-recall, we obtain a total of 3,466 data samples, pruned from 134,116 polyp candidates on the training dataset. All the 554 true positive lesion instances are contained, along with other "harder" negatives, having higher classification scores. For fine-level classification, we learn the mapping function $\phi : \mathcal{X} \mapsto \mathcal{Y}$ after feature selection using the pruned dataset, and the t-centers are fitted in the reduced $\mathcal{Y}$ feature space for the soft kNN classifier. We plot the FROC curves comparing using RVMMIL as a single classifier, using SPG [2] as a integrated dimension reduction approach, and our two-tiered coarse-to-fine classifier, on training and testing datasets, as shown in Fig. 3. Fig. 3(a) shows the whole FROC curve. Since reasonably small FP rates are clinically more meaningful, we highlight in the partial-FROC with FP rate $\in$ [2, 5], as shown it in Fig. 3(b). For validation, the testing results demonstrate that our CTF method can increase the sensitivity of RVMMIL by 2.58% (from 0.8903 to 0.9161) at the FP rate = 4, or reduce the FP rate by 1.754 (from 5.338 to 3.584) when sensitivity is 0.9097, which are statistically significant improvements for colorectal cancer detection. It also clearly outperforms other state-of-the-arts, e.g. SPG [6] as shown in Fig. 3, as well as [33, 35, 40, 45].

To fully leverage the topology-preserving property of learned $\mathcal{Y}$, we also use it for polyp retrieval, which is defined as giving a query polyp in one prone/supine scan, to retrieve its

[2]We use the code implemented by Dr. Deng Cai http://www.zjucadcg.cn/dengcai/SR/index.html



(a)



(b)

Figure 3: (a) FROC comparison of using our proposed CTF method, single-layer RVMMIL [35] classifier and spectral projection on graph (SPG) [6] on classifying the training and testing datasets of colon polyps. (b) Zoom in of (a) for the part of FP rate $\in [2, 5]$.

counterparts in the other view. To achieve this, we find the k nearest neighbors (kNN) of a query $\mathbf{y}_i \in \mathcal{Y}$ using the classified polyps, and check whether the true match is inside the neighborhood of kNN. If the true matched polyp is in the kNN, a 'hit' will occur. We record the retrieval rate, as the ratio of the number of 'hit' polyp divided by the query polyp number, at different k levels. Especially, high retrieval rate with small k can greatly alleviate radiologists' manual efforts on finding the counterpart same polyp, with better accuracy. To show its advantage, we employ a traditional geometric feature based polyp retrieval scheme, namely geodesic distance that measures the geodesic length of a polyp to a fixed anatomical point (e.g., rectum), along the colon centerline curve. The retrieval rate comparison is illustrated in Fig. 4, for training and testing datasets. The results indicate that the retrieval accuracy can achieve 80% when only 2 to 4 neighbors are necessary. This shows that nonparametric kNN in $\mathcal{Y}$ subspace based retrieval significantly improves the conventional polyp matching scheme, contingent on geometric computation of geodesic distance and the SPG based retrieval.

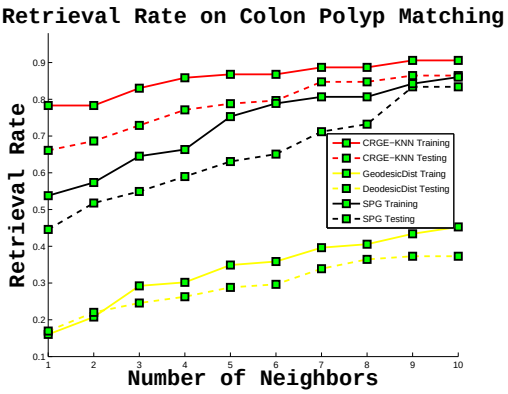


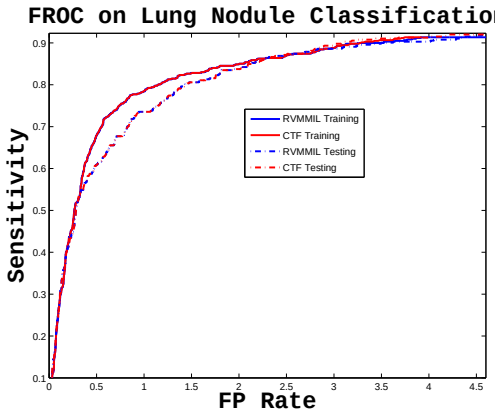
Figure 4: Retrieval comparison of using our proposed CTF method, the single-layer RVMMIL [35] classifier, and spectral projection on graph (SPG) [6] on colon polyp retrieval, in training and testing.

5.2 Lung Nodule Classification

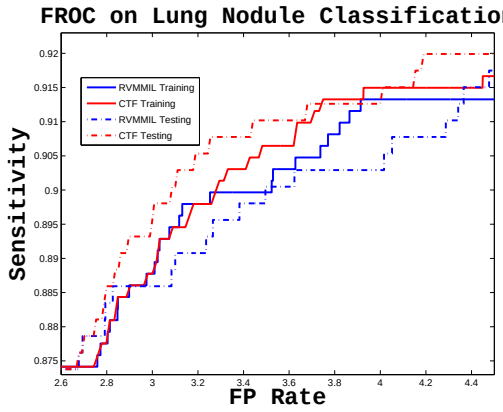
Data: The lung nodule dataset is collected from $1,000$ patients from multiple hospitals in different countries, using multi-vendor scanners. Before sample pruning, there are $28,804$ samples of which $27,334$ are negatives and $1,470$ are true nodule instances from 588 patients in training dataset. The testing dataset contains $20,288$ candidates, with $19,227$ are negatives and $1,061$ are positives of 412 patients. Several instances may correspond to the same lung nodule in one volume. All types of *solid, partial-solid and Ground Glass Nodules* with a diameter range of 4-30mm are considered. Each sample has 112 informative features, including texture appearance features (e.g. as the moments of responses to a multiscale filter bank, [17, 29]), shape (e.g. width, height, volume, number of voxels), location context (e.g. distance to the wall, at the right or left of the wall), gray value, and morphological features (e.g., obtained using the edge-guided wavelet snake model as in [24]).

First, FROC analysis by using our proposed coarse-to-fine classification framework, compared to single-layer RVMMIL classifier, for the lung nodule classification in training and testing is shown in Fig. 5. From the figure we can see that the testing FROC of CTF dominates the RVMMIL FROC, when the FP rate $\in [3, 4]$, with $1.0 \sim 1.5\%$ consistent sensitivity improvements. We also compared with the SPG framework, and the FROC analysis is shown in Fig. 6. The comparison also shows the higher classification accuracy of our proposed method. Furthermore, our CTF classification performance compares favorably with other recent developments in lung CAD [16].

Next we evaluate the effects of using t-center (default), mean or median as estimated templates in CTF. The comparisons are shown in Fig. 8 and Fig. 9 on the training and testing parts of the lung dataset. The comparison validates that t-center outperforms the templates formed by typical mean or median method. Last, we compare our method to a related locality-classification framework, SVM-kNN [48] which shows highly competitive results on image based multiclass object recognition problems. SVM-kNN uses kNN to find data clusters as nearest neighbors and train a support vector machine (SVM) on each locality group for "divide-



(a)



(b)

Figure 5: (a) FROC analysis using our proposed CTF method and RVMMIL classifier, in training and testing of the lung nodule dataset. (b) Zoom in of (a) for the part of FP rate $\in [2.6, 4.5]$.

and-conquer" classification [48]. The comparison results are illustrated in Fig. 7, which shows that our method outperforms the SVM-kNN method on both training and testing datasets.

6. CONCLUSIONS & FUTURE WORK

Our main contributions are summarized in three folds. First, we introduce a new coarse-to-fine classification framework for computer-aided (cancer) detection problems by robustly pruning data samples and mining their heterogeneous imaging features. Second, we propose a new objective function to integrate the between-class dissimilarity information into embedding method. Third, two challenging large scale clinical datasets on colon polyp and lung nodule classification are employed for performance evaluation, which show that we outperform, in both tasks, the state-of-the-art CAD systems [16, 30, 33, 40, 45] where a variety of single parametric classifiers were used. For future work, we plan to investigate optimizing the fine-level classification in an associate Markov network [42] setting, which integrates structured prediction among data samples (i.e., graph parameters are jointly learned with classification).

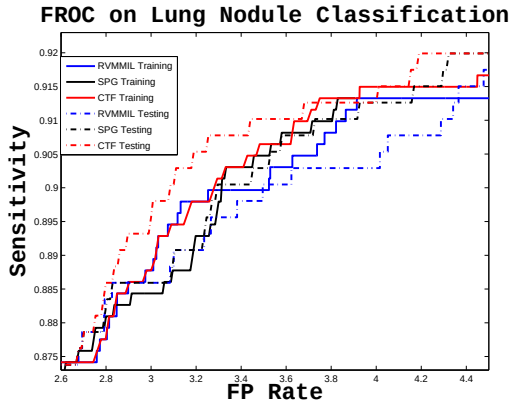


Figure 6: FROC analysis using our proposed CTF method, RVMMIL classifier and SPG in training and testing of the lung nodule dataset.

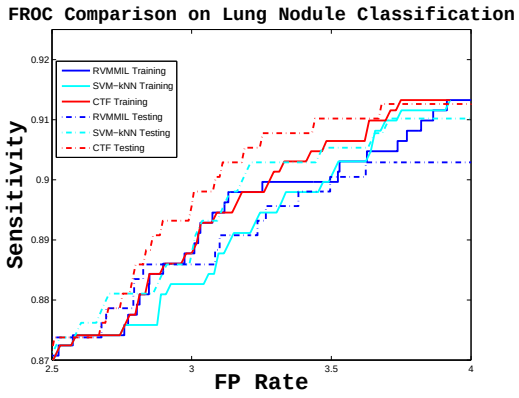


Figure 7: FFROC analysis using our proposed CTF method, RVMMIL classifier and SVM-kNN classification scheme, in training and testing.

7. REFERENCES

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *In Advances in Neural Information Processing Systems*, pages 585–591, 2001.

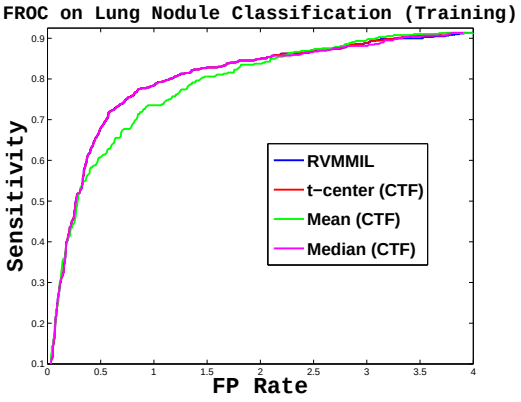
[2] M. Belkin and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, pages 1373–1396, 2003.

[3] J. Bi, K. Bennett, M. Embrechts, C. Breneman, and M. Song. Dimensionality Reduction via Sparse Support Vector Machines. *Journal of Machine Learning Research*, pages 633–42, 2003.

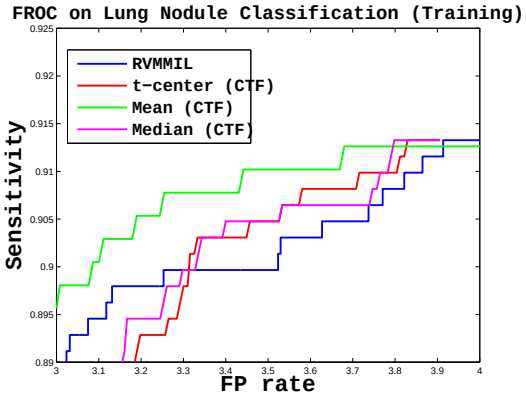
[4] S. Boutemedjet, N. Bouguila, and D. Ziou. A hybrid feature extraction selection approach for high-dimensional non-gaussian data clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(8):1429–1443, 2009.

[5] C. Boutsidis, M. W. Mahoney, and P. Drineas. Unsupervised feature selection for principal components analysis. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 61–69, 2008.

[6] D. Cai, X. He, and J. Han. Sparse Projections over

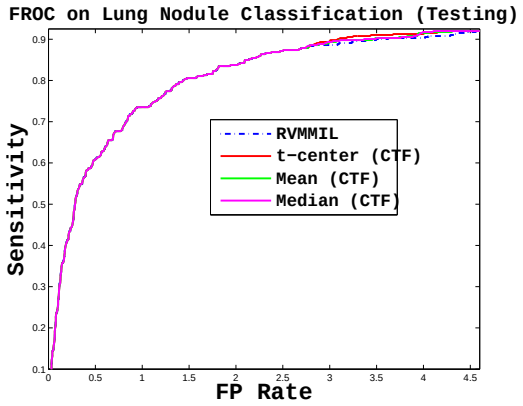


(a)

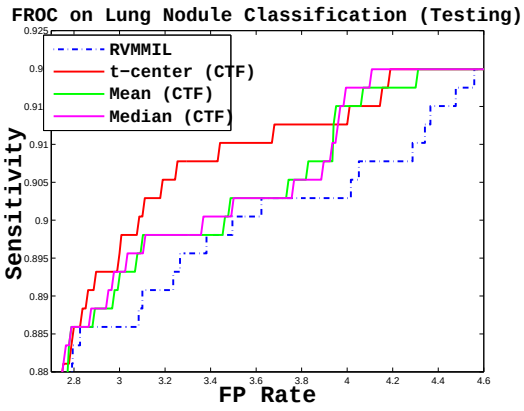


(b)

Figure 8: FROC analysis using t-center, mean or median as estimated templates in CTF, compared with RVMMIL classifier in training (a): original comparison and (b): after zooming in.



(a)



(b)

Figure 9: FROC analysis using t-center, mean or median as estimated templates in CTF, compared with RVMMIL classifier in testing (a): original comparison and (b): after zooming in.

Graph. *Proceedings AAAI Conference on Artificial Intelligence*, pages 610–615, 2008.

[7] D. Cai, X. He, Y. Hu, J. Han, and T. Huang. Learning a Spatially Smooth Subspace for Face Recognition. *IEEE Computer Vision and Pattern Recognition*, pages 1–7, 2007.

[8] D. Cai, C. Zhang, and X. He. Unsupervised Feature Selection for Multi-Cluster Data. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2010.

[9] F. R. K. Chung. Spectral Graph Theory. *Regional Conference Series in Mathematics. AMS*, 1997.

[10] T. M. Cover and J. A. Thomas. Elements of Information Theory. *Wiley-Interscience, 2nd edition*, 2006.

[11] M. Cox and T. Cox. *Multidimensional Scaling*, pages 315–347. Springer Handbooks Comp. Statistics, 2008.

[12] R. Duda, P. Hart, and D. Stork. Pattern Classification. *Wiley-Interscience, Hoboken, NJ, 2nd edition*, 2000.

[13] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[14] R. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7:179–188, 1936.

[15] D. Furlan and et al. Hierarchical clustering analysis of pathologic and molecular data identifies prognostically and biologically distinct groups of colorectal carcinomas. *Modern Pathology*, 24:126–137, 2011.

[16] B. V. Ginneken and et al. Comparing and Combining Algorithms for Computer-aided Detection of Pulmonary Nodules in Computed Tomography Scans: The ANODE09 Study. *Medical Image Analysis*, pages 707–22, 2010.

[17] B. V. Ginneken, B. M. T. H. Romeny, M. A. Viergever, and M. Ieee. Computer-aided diagnosis in chest radiography: A survey. *IEEE Transactions on Medical Imaging*, 20(12):1228–1241, 2001.

[18] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighborhood Component Analysis. *In Advances in Neural Information Processing Systems*, 2004.

[19] I. Guyon and A. Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, pages 1157–1182, 2003.

[20] J. Hamm, D. Ye, R. Verma, and C. Davatzikos.

GRAM: A Framework for Geodesic Registration on Anatomical Manifolds. *Medical Image Analysis*, pages 633–42, 2010.

[21] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. *In Advances in Neural Information Processing Systems 18*, pages 61–69, 2005.

[22] X. He and P. Niyogi. Locality Preserving Projections. *In Advances in Neural Information Processing Systems*, 2003.

[23] I. Jolliffe. Principal Component Analysis. *Springer-Verlag*, 1986.

[24] B. Keserci and H. Yoshida. Computerized detection of pulmonary nodules in chest radiographs based on morphological features and wavelet snake model. *Medical Image Analysis*, 6:431–447, 2002.

[25] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491–502, 2005.

[26] M. Liu, B. vemuri, S. Amari, and F. Nielsen. Total Bregman Divergence and its Applications to Shape Retrieval. *IEEE Computer Vision and Pattern Recognition*, 2010.

[27] S.-C. B. Lo, L.-Y. Hsu, M. T. Freedman, Y. M. F. Lure, and H. Zhao. Classification of lung nodules in diagnostic CT: an approach based on 3D vascular features, nodule density distribution, and shape features. *Proceedings of the SPIE*, 5032:183–189, 2003.

[28] L. Lu, J. Bi, M. Wolf, and M. Salganicoff. Effective 3D Object Detection and Regression Using Probabilistic Segmentation Features in CT Images. *IEEE Computer Vision and Pattern Recognition*, 2011.

[29] M. N. Muhammad, D. S. Raicu, J. D. Furst, and E. Varutbangkul. Texture versus shape analysis for lung nodule similarity in computed tomography studies. *Proceedings of the SPIE*, 2008.

[30] K. Murphy and et al. A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification. *Medical Image Analysis*, pages 670–70, 2009.

[31] H. Peng, F. Long, and C. Ding. Feature Selection Based on Mutual Information: Criteria of Max-dependency, Max-relevance, and Min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1226–1238, 2005.

[32] C. S. Pereira, L. A. Alex, A. M. Mendona, A. Campilho, and R. R. Frias. A Multiclassifier Approach for Lung Nodule Classification. *International Conference on Image Analysis and Recognition*, 2006.

[33] V. F. V. Ravesteijn, C. V. Wijk, F. M. Vos, R. Truyen, J. F. Peters, J. Stoker, and L. J. V. Vliet. Computer Aided Detection of Polyps in CT Colonography Using Logistic Regression. *IEEE Transactions on Medical Imaging*, 2010.

[34] S. Ray and R. H. Turi. Determination of Number of Clusters in K-means Clustering and Application in Colour Image Segmentation. *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques*, pages 137–143, 1999.

[35] V. Raykar and et al. Bayesian Multiple Instance Learning: Automatic Feature Selection and Inductive Transfer. *Proceedings of International Conference on Machine Learning*, pages 808–815, 2008.

[36] R. Rockafellar. Convex Analysis. *Princeton University Press*, 1970.

[37] J. L. Rodgers and W. A. Nicewander. Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 42(1):59–66, 1988.

[38] S. Roweis and L. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290:2323–2326, 2000.

[39] S. Salem, K. Seridi, L. Seridi, J. Wu, and M. Zaki. VOKNN: Voting-based Nearest Neighbor Approach for Scalable SVM Training. *The 2nd Workshop on Large-scale Data Mining: Theory and Applications in conjunction with the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010.

[40] G. Slabaugh and et al. A Robust and Fast System for CTC Computer-Aided Detection of Colorectal Lesions. *Algorithms*, 3(1):21–43, 2010.

[41] B. Tang and D. Mazzoni. Multiclass Reduced-set Support Vector Machines. *Proceedings of International Conference on Machine Learning*, pages 921–928, 2006.

[42] B. Taskar, V. Chatalbashev, and D. Koller. Learning Associative Markov Networks. *Proceedings of International Conference on Machine Learning*, 2004.

[43] J. Tenenbaum, V. Silva, and J. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290:2319–2323, 2000.

[44] B. Vemuri, M. Liu, S. Amari, and F. Nielsen. Total Bregman Divergence and its Applications to DTI Analysis. *IEEE Transactions on Medical Imaging*, 30(2):475–483, 2011.

[45] S. Wang, J. Yao, and R. Summers. Improved Classifier for Computer-aided Polyp Detection in CT Colonography by Nonlinear Dimensionality Reduction. *Medical Physics*, pages 35:1377–1386, 2008.

[46] L. Wolf and A. Shashua. Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach. *Journal of Machine Learning Research*, pages 1855–1887, 2005.

[47] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin. Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 40–51, 2006.

[48] H. Zhang, A. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. *IEEE Computer Vision and Pattern Recognition*, 2:2126–2136, 2006.

[49] Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. *In Proceedings of the 24th Annual International Conference on Machine Learning*, pages 1151–1157, 2007.