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

Meizhu Liu Le Lu Xiaojing Ye Shipeng Yu Heng Huang
Siemens Medical Solutions Inc, Malvern, PA 19355, USA
University of Florida, Gainesville, FL 32611, USA
University of Texas, Arlington, TX 76013, USA

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., [13]); 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 [17], boosting [15], logistic regression [11], relevance vector machine [13], k-nearest neighbor [9] or spectral projections on graph [2].

Categories and Subject Descriptors

Industrial and Application Paper [**Knowledge Management (KM)**]:

Keywords

computer-aided diagnosis, coarse-to-fine classification, class regularized graph embedding, total Bregman divergence (tBD) clustering , t-centers

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

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 [4, 9, 11, 15, 17]. 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.

This paper mainly focuses on the classification aspect of CAD. We explain that **Why CAD is an important yet domain-specific classification problem** as follows. 1), In CAD scenario, positive examples (i.e., true cancers) in large population screen are very rare and precious, thus **high sensitivity (recall) is a must-to-have feature to make CAD meaningful.** 2), On the other hand, CAD is legally used as a second-assisted tool for radiologists, which hopes to improve radiologists' overall performance but mostly relies on them to disambiguate false positives from CAD[1]. CAD findings (with only $< 1\%$ true positives) can be very expense burden for radiologists, lowering their confidence, and without effectively guiding them to real detections. **Therefore it is equivalently important to archive sensibly low false positive rate per case (e.g., $2 \sim 5$).** 3), To better address doctor's final decision support on CAD and even his/her own findings, **an ideal setup is to make the system capable of retrieving similar or counterpart lesions when available**. For such consideration, we argue that nearest neighbor (NN) or template matching (TM) type nonparametric classification methodology will be a more sensible choice by precisely modeling all distributional information of rare positives for classification. This is a more complete representation, versus discriminative models (e.g., SVM [17] and RVM [13]) with linear or nonlinear parametric decision boundaries, which is prone to overtraining given small number of positives. However, there are two fundamental problems needed to be solved to make nonparametric classification method robust. 1), **There are dominating numbers of false positives initially** (from a highly sensitive candidate-generation process), which not only can effect the performance negatively also makes the computation slow. 2), It is well known that **NN and TM**

[1]A clinical report stating the reason to dismiss or accept each CAD findings is required in the workflow.

are very sensitive to the feature space or subspace where matching distance or (dis-)similarity metrics are computed.

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 [11], boosting [15], support/relevance vector machines [13]) 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 [2]). (2) *Feature Pursuing + kNN/Template Matching:* We first apply 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 our proposed framework on colon polyp and lung nodule detection, using two large scale clinical datasets collected from multiple clinical sites across continents.

2. SAMPLE PRUNING USING PARAMETRIC RVMMIL

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 [13] which is a powerful extension to integrate feature selection and handle multiple instance learning (MIL).

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 $\mathbf{a}$. For the contents of sparse feature selection and multiple instance learning in RVM, we refer the readers to [13]. From our coarse-to-fine classification model, RVMMIL is adopted 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 will largely impact the final classification accuracy. Note that other classifiers with confidence estimates, as boosting [15] and regularized SVM [17], are also applicable.

[2]This is related with using nearest neighbor analysis to find data samples either near the decision boundary or in local neighborhoods [18], 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.

3. FEATURE SPACE PURSUIT

Our goal of feature pursuit is to estimate intrinsic, lower dimensional feature subspace of data for later sensible nonparametric classification, while preserves 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: feature selection and class regularized graph embedding.

Feature Selection: By applying feature selection, only a compact subset of highly statistical relevant features is retained, to simplify the later graph embedding or feature projection process. There are many feature selection techniques, we use Maximum Relevance Minimum Redundancy (MRMR) feature selection algorithm [10] due to its empirical good performance and computational efficiency. Given a set of features $\mathbb{F} = \{f_i\}$, its MRMR feature subset $\mathbb{H}$ maximizes the following objective κ:

$$\kappa(\mathbb{H}, \mathbf{y}) = \gamma(\mathbb{H}, \mathbf{y}) - \gamma(\mathbb{H}), \tag{1}$$

where

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

$$\gamma(\mathbb{H}, \mathbf{y}) = \frac{1}{m} \sum_{f_i \in \mathbb{H}} \gamma(f_i, \mathbf{y}), \tag{3}$$

and m is the total number of elements in $\mathbb{H}$. Due to space limit, refer the algorithm details to [10].

Class Regularized Graph Embedding: We propose and exploit a new *Class Regularized Graph Embedding* (CRGE) scheme 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. In Graph embedding, feature projections can be learned in many different ways. 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 [1, 3] and Locality Preserving Projection [6].

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

$$L(\mathcal{Y}) = \sum_{i,j} \|\mathbf{y}_i - \mathbf{y}_j\|^2 W_{ij}, \tag{4}$$

under some appropriate constraints. Though performed well in many applications [3, 6], the limitation of Eq. (4) is that

it does not penalize the similarity between points belonging to different classes. For this means, we propose class regularized graph embedding (CRGE) to find a mapping $\phi : \mathcal{X} \mapsto \mathcal{Y}$, such that ϕ minimizes the function $E(\mathcal{Y})$ defined as

$$E(\mathcal{Y}) = \sum_{i,j\in\mathcal{S}} \|\mathbf{y}_i - \mathbf{y}_j\|^2 W_{ij} - \sum_{i,j\in\mathcal{D}} \|\mathbf{y}_i - \mathbf{y}_j\|^2 W_{ij}\,, \quad \text{subject to: } \|\mathcal{Y}\|_F = 1. \tag{5}$$

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 (5) and get

$$\min \sum_{i,j} \|\mathbf{y}_i - \mathbf{y}_j\|^2 W_{ij} H_{ij}\,, \tag{6}$$

where H_{ij} is the Heaviside function and

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

The mapping function $\phi(\mathbf{x})$ can be linear or nonlinear, and We use linear mapping because of its simplicity and generality, such as

$$\mathbf{y} = \phi(\mathbf{x}) = M'\mathbf{x},\ M \in \mathbb{R}^{n\times\tilde{n}}\,, \tilde{n} \ll n\,. \tag{7}$$

Plugging (7) into (6), we get

$$\min_M \sum_{i,j} \|M'\mathbf{x}_i - M'\mathbf{x}_j\|^2 W_{ij} H_{ij}\,, \quad \text{subject to: } \|M\|_F = 1\,, \tag{8}$$

where the constraint $\|M\|_F = 1$ eliminates the scaling effect. Eq. (8) can be solved very quickly using gradient descent technique along with iterative projections [14]. The computation of W is chosen in the following manner, to fit our problem specific need.

$$W(i,j) = \mathbf{x}_i'\mathbf{x}_j / \|\mathbf{x}_i\|\|\mathbf{x}_j\|\,. \tag{9}$$

Finally, we argue that our stratified approach which prunes non-informative or redundant features, using feature selection from an information-theoretic aspect, before feature graph embedding or projection, can simplify the optimization process of graph embedding on a reduced feature set. This strategy may achieve better overall results, compared from the joint sparsity-constrained graph embedding (as SPG) [2].

4. NONPARAMETRIC CLASSIFICATION

The naive KNN classification on data instances, in feature space $\mathcal{Y} = \phi(\mathbf{x})$, performs poorly due to unbalanced number of rare positives and dominating negatives. Thus we propose to do KNN voting using learned templates from clustering.

Clustering & Templates: Our previous total Bregman divergence (tBD) clustering algorithm [7] is utilized here. 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, which is naturally robust and leads to efficient algorithms for soft and hard clustering. 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* [12], given by

$$\text{index} = \frac{\text{intra}}{\text{inter}}, \quad \text{intra} = \frac{1}{N}\sum_{i=1}^{c}\sum_{y\in C_i} \|y - z_i\|^2, \quad \text{inter} = \min_{i,j} \|z_i - z_j\|^2, \tag{10}$$

where C_i is the ith cluster with center z_i. Each cluster is represented as the *tBD* center, termed t-center [7], 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

$$z = \arg\min_{\tilde{z}} \sum_{i=1}^{N} \delta_f(\tilde{z}, \mathbf{y}_i), \tag{11}$$

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

$$\delta_f(\mathbf{y}_1, \mathbf{y}_2) = \frac{f(\mathbf{y}_1) - f(\mathbf{y}_2) - \langle \mathbf{y}_1 - \mathbf{y}_2, \nabla f(\mathbf{y}_2)\rangle}{\sqrt{1 + \|\nabla f(\mathbf{y}_2)\|^2}}. \tag{12}$$

Here, we use $f(y) = \|y\|^2$, and hence δ_f becomes the total square loss [7] and the t-center in Eq. (11) becomes

$$z = \sum_{i=1}^{N} a_i \mathbf{y}_i\,, \text{ where } a_i = \frac{1/\sqrt{1 + 4\|\mathbf{y}_i\|^2}}{(\sum_j 1/\sqrt{1 + 4\|\mathbf{y}_j\|^2})}\,. \tag{13}$$

Template Matching via kNN Voting: 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 $\{d_1, d_2, \cdots, d_k\}$. We define the empirical probability of $\mathbf{y}_i$ being positive as p, and

$$p = \frac{\sum_{(z_j \text{ is positive})} 1/d_j}{\sum_{(z_l \text{ is negative})} 1/d_l + \sum_{(z_j \text{ is positive})} 1/d_j}\,. \tag{14}$$

Based on the p value, we can draw the FROC curve of sensitivity and FP rate per case for training and testing datasets. Eq. (14) is a soft kNN voting scheme using the reciprocal of distance $1/d_i$. We found that t-centers are more robust as they lead to preserve 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, by maximizing the partial Area Under FROC Curve (AUC):

$$k = \arg\max_{\tilde{k}}\ \text{AUC}(\text{FPrate} \in [2,\ 4]). \tag{15}$$

5. EXPERIMENTAL RESULTS

Our CTF classification frame is evaluated on representative large scale colon polyp and lung nodule datasets, collected from dozens of hospitals across US, Europe and Asia.

Colon Polyp Detection & Retrieval: 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 imagery feature vector, describing its shape, intensity pattern, segmented class-conditional likelihood statistics and other higher level features [11, 8, 15, 17]. 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 training and testing datasets with no overlapping.

Upon obtaining the parametric RVMMIL model [13], we get the probability (or classification score) of each instance being positive. Then we perform thresholding and only consider candidates with $p(y = 1|\mathbf{x}_i) \geq \hat{\rho} = 0.0157$ as a classification cascade with high-recall. A total of 3,466 data samples are retained, 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 as a integrated sparsity and dimension reduction approach, and our two-tiered coarse-to-fine classifier, on training and testing datasets, as shown in Fig. 1 **Left**. 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 [2] as shown in Fig. 1, as well as [11, 13, 15, 17].

To fully leverage the feature space topology-preserving property of the lower-dimensional embedding $\mathcal{Y}$, we also test it for polyp retrieval, which is defined as giving a query polyp in one prone/supine scan, to retrieve its 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 demonstrate its advantage, we employ a conventional geometric distance 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. The retrieval rate comparison is illustrated in Fig. 1 **Middle**, for both 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.

Lung Nodule Classification: 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 imagery features, including texture appearance features (e.g. as the moments of responses to a multiscale filter bank [5]), 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. 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. 1 **Right**. 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. 2 **Left**. 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 [4]. Next, we compare our method to a related locality-classification framework, SVM-kNN [18] 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-and-conquer" classification [18]. The comparison results are illustrated in 2 **Middle**, which shows that our method outperforms the SVM-kNN method on both lung training and testing datasets. Finally, we evaluate the effects of using t-center (default), mean or median as estimated templates in CTF. The comparison in testing using the lung dataset is shown in Fig. 2 **Right**. The comparison validates that t-center outperforms the templates formed by typical mean or median method.

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 [4, 9, 11, 15, 17] 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 [16] setting, which integrates structured prediction among data samples (i.e., graph parameters are jointly learned with classification).

7. REFERENCES

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

[2] D. Cai, X. He, and J. Han. Sparse Projections over Graph. *Proceedings AAAI Conference on Artificial Intelligence*, pages 610–615, 2008.

[3] 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.

[4] B. V. Ginneken and et al. Comparing and Combining Algorithms for Computer-aided Detection of Pulmonary

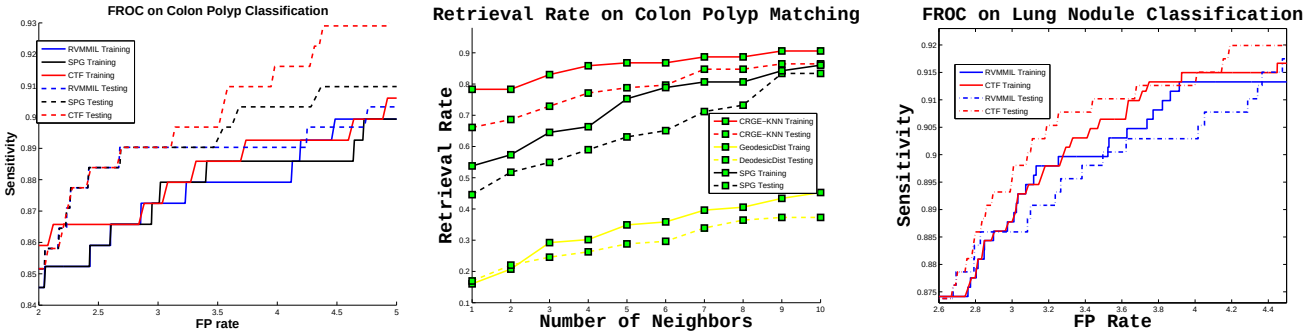

Figure 1: Left: FROC comparison of using our proposed CTF method, single-layer RVMMIL [13] classifier and spectral projection on graph (SPG) [2] on classifying the training and testing datasets of colon polyps, with FP rate $\in [2, 5]$. Middle: Nearest neighbor based retrieval comparison of using our proposed CTF method, Geodesic distance, and spectral projection on graph (SPG) [2] on colon polyp retrieval, in training and testing. Right: Partial FROC analysis using our proposed CTF method and RVMMIL classifier, in training and testing of the lung nodule dataset, with FP rate $\in [2.6, 4.5]$

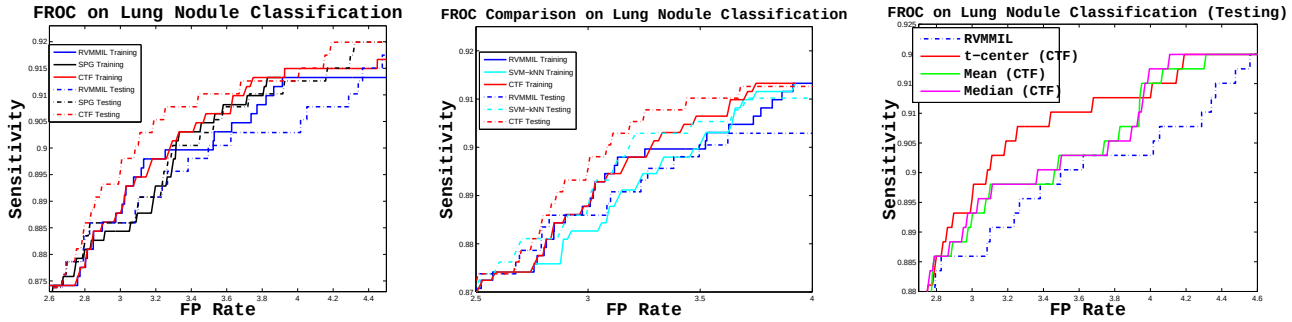

Figure 2: Left: FROC analysis using our proposed CTF method, RVMMIL classifier and SPG in training and testing of the lung nodule dataset. Middle: FROC analysis using our proposed CTF method, RVMMIL classifier and SVM-kNN classification scheme, in training and testing [18]. Right: FROC analysis using t-center, mean or median as estimated templates in CTF, compared with RVMMIL classifier in testing.

Nodules in Computed Tomography Scans: The ANODE09 Study. *Medical Image Analysis*, pages 707–22, 2010.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

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

[17] 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.

[18] 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.