

Graph Embedding to Improve Supervised Classification and Novel Class Detection: Application to Prostate Cancer

Anant Madabhushi¹, Jianbo Shi², Mark Rosen²,
John E. Tomaszewski², and Michael D. Feldman²

Rutgers University, Piscataway, NJ 08854,
University of Pennsylvania, Philadelphia, PA 19104
`anantm@rci.rutgers.edu`

Abstract. Recently there has been a great deal of interest in algorithms for constructing low-dimensional feature-space embeddings of high dimensional data sets in order to visualize inter- and intra-class relationships. In this paper we present a novel application of graph embedding in improving the accuracy of supervised classification schemes, especially in cases where object class labels cannot be reliably ascertained. By refining the initial training set of class labels we seek to improve the prior class distributions and thus classification accuracy. We also present a novel way of visualizing the class embeddings which makes it easy to appreciate inter-class relationships and to infer the presence of new classes which were not part of the original classification. We demonstrate the utility of the method in detecting prostatic adenocarcinoma from high-resolution MRI.

1 Introduction

The aim of embedding algorithms is to construct low-dimensional feature-space embeddings of high-dimensional data sets [1–4]. The low-dimensional representation is easier to visualize and helps provide easily interpretable representations of intra-class relationships, so that objects that are closer to one another in the high dimensional ambient space are mapped to nearby points in the output embedding. Recently researchers have begun exploring the use of embedding for solving different problems. Dhillon [1] employed embedding for visually understanding the similarity of different classes, distance between class clusters, and to evaluate the coherence of each of the class clusters. Iwata *et al.* [2] described a parametric embedding method to provide insight into classifier behavior. Euclidean embedding of co-occurrence data has also been successfully applied to classifying text databases [3] and for detecting unusual activity [4]. In this paper we demonstrate a novel application of graph embedding in (i) improving the accuracy of supervised classification tasks and (ii) for identifying *novel classes*, i.e. classes not included in the original classification.

In [5] we presented a computer-aided detection (CAD) methodology for detecting prostatic adenocarcinoma from high resolution MRI, which in several

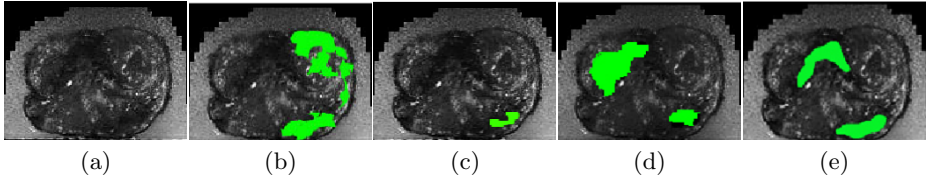


Fig. 1. (a) Original MR image of the prostate, (b) ground truth for tumor (in green) determined manually from the corresponding histology [5]. Three expert segmentations (Fig. 1(c)-(e)) based on visual inspection of Fig. 1(a) without the accompanying histologic information. Note the low levels of inter-expert agreement.

instances outperformed trained experts. It was found that the false positive errors due to CAD were on account of,

- Errors in tumor ground truth labels on MRI, since the tumor labels were established by manually registering the corresponding histologic and MR slices (both MR and histologic slices being of different slice thickness) and due to the difficulty in identifying cancer (Fig. 1).
- The presence of objects with characteristics between tumor and non-tumor (e.g. pre-cancerous lesions). Since the system is not trained to recognize these novel classes, the classifier *forces* these objects into one of the original classes, contributing to false positives.

In order to detect novel classes, we need to first eliminate true outliers due to human errors from the training set. The implications of outlier removal from the training set are two fold.

- (1) It can significantly improve the accuracy of the original classification, and
- (2) It ensures that objects that now lie in the overlap between the object classes after outlier removal, truly represent the novel classes.

We borrow a graph embedding technique used in the computer vision domain [4] for improving classification accuracy and for novel class detection. However, while in [4] both the object and its representative class are co-embedded into the low-dimensional space, in our case, the embedding algorithm takes as input the *a-posteriori* likelihoods of objects belonging to the tumor class. Note that graph embedding differs from data reduction techniques like PCA [7] in that the relationship between adjacent objects in the higher dimensional space is preserved in the co-embedded lower dimensional space. While we have focused on one specific CAD application in this paper [5], we emphasize that our methods are applicable to most supervised or semi-supervised classification tasks, especially those in which class labels cannot be reliably ascertained.

This paper is organized as follows. Section 2 provides a brief description of our methodology, and a detailed description of the individual modules is given in Section 3. In Section 4 we present the results of quantitative evaluation of our methodology on a CAD system for prostate cancer. Concluding remarks are presented in Section 5.

2 System Overview for Prostate Cancer Detection

Fig. 2 shows the main modules and pathways comprising the system. Fig. 3 shows the results of graph embedding on a high resolution MR study of the prostate (Fig. 3(a)). Fig. 3(c) is a map of the posterior likelihood of every voxel belonging to the tumor class; the posterior likelihood being derived from the prior distribution (dashed line in Fig. 3(f)), obtained with the initial set of tumor class labels (Fig. 3(b)) and Fig. 3(e) shows the corresponding *probability* image using the refined prior distribution after graph embedding (solid line in Fig. 3(f)). The plot of graph embedding (Fig. 3(d)) shows considerable overlap (ellipse 3) between the tumor (red circles) and non-tumor (black dots) classes. Using the refined probability map in Fig. 3(e), the resultant embedding (Fig. 3(f)) shows a clear separation between the two classes (ellipses 1, 2). The increased class separation is also reflected in the increased image contrast of Fig. 3(e) over Fig. 3(c). Fig. 3(g) shows a novel way of visualizing the graph embeddings in Fig. 3(f), with objects that are adjacent in the embedding space being assigned similar colors. Objects that lie in the overlap of the class clusters after outlier removal (ellipse 3 in Fig. 3(f)) correspond to the apparent false positive area (marked as FP) in Fig. 3(g). This region is actually inflammation induced by atrophy (confirmed via the histology slice in Fig. 3(h)).

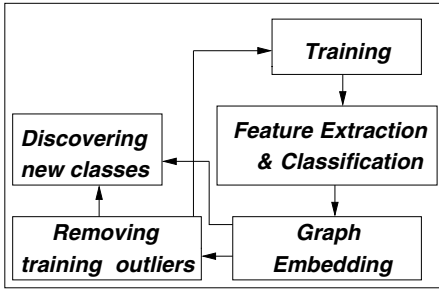


Fig. 2. Training distributions for individual features are generated using existing class labels, and each voxel assigned a posterior likelihood of being tumor. Graph embedding on the posterior likelihoods is used to remove training outliers and (i) improve the prior distributions and (ii) identify new classes.

3 Methodology

3.1 Notation

We represent a 3D image or *scene* by a pair $\mathcal{C} = (C, g)$, where C is a finite 3-dimensional rectangular array of voxels, and g is a function that assigns an integer intensity value $g(c)$ for each voxel $c \in C$. The feature scenes $\mathcal{F}^i = (C, f^i)$ are obtained by application of K different feature operators, for $1 \leq i \leq K$. The tumor class is denoted by ω_t and S_{ω_t} denotes the *true ground truth* set, such that for any voxel $d \in S_{\omega_t}$, $d \hookrightarrow \omega_t$ where \hookrightarrow denotes the “belongs to” relationship. \hat{S}_{ω_t} is the surrogate of ground truth S_{ω_t} obtained by experts by visually registering the MR and the histologic slices [5]. $\hat{S}_{\omega_t}^T \subset \hat{S}_{\omega_t}$ is the training set used for generating the prior distributions $\hat{p}(f^i | c \hookrightarrow \omega_t)$ for each feature f^i . Given $\hat{p}(f^i | c \hookrightarrow \omega_t)$, the *a-posteriori* probability that voxel $c \hookrightarrow$

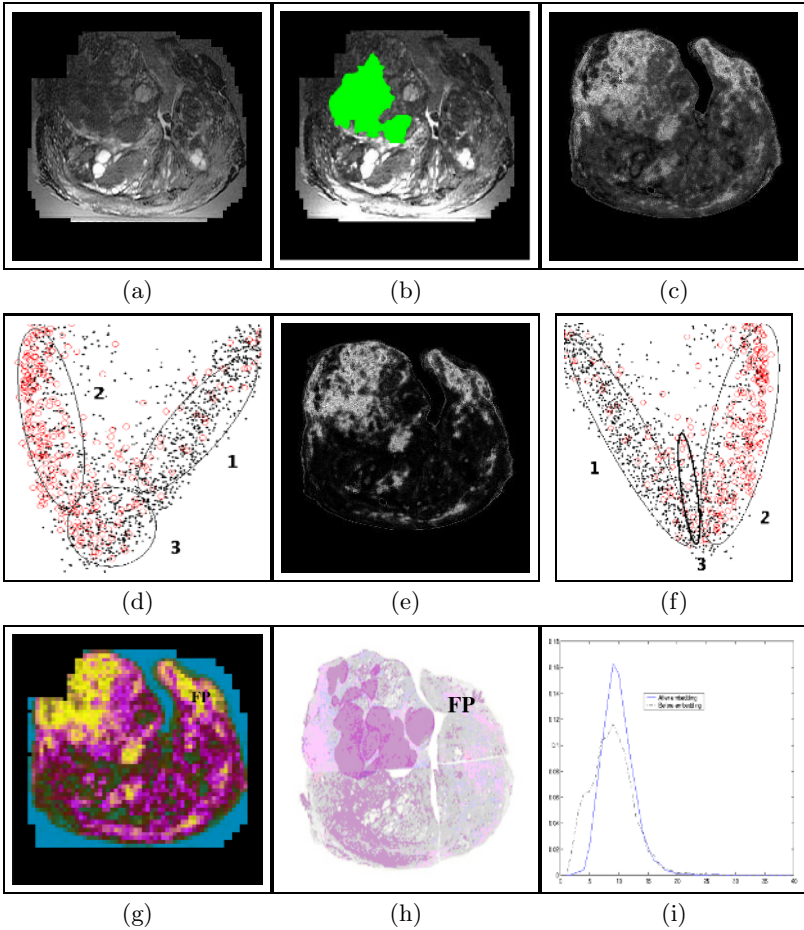


Fig. 3. (a) Original MR scene \mathcal{C} , (b) surrogate of ground truth (in green) for cancer (\hat{S}_{ω_t}) superposed on (a), (c) combined likelihood scene showing tumor class probability before outlier refinement via embedding, (d) graph embedding of tumor/non-tumor class likelihoods in (c), (e) combined likelihood scene showing tumor class probabilities after outlier removal, (f) graph embedding of tumor/non-tumor class likelihoods in (e), (g) RGB representation of graph embeddings in (f), and (h) the histology slice corresponding to the MR slice in (a). Note the greater contrast between intensities in (e) compared to (c), reflecting the increased separation between the tumor and non-tumor clusters after outlier removal. This is also reflected in the overlap of the tumor (red circles) and non-tumor (black dots) clusters in the embedding plot before outlier removal (ellipse 3 in (d)) and the more distinct separation of the two clusters after outlier removal (ellipse 3 in (f)), constitute the intermediate class (between tumor and non-tumor). Also note the tighter envelope of the prior distribution of feature f^i (3(i)) after embedding (solid line) compared to before (dashed line). The embedding scene in 3(g) also reveals that an apparent false positive area (FP on 3(g)) actually corresponds to a new object class not included in the original classification (inflammation induced by atrophy, confirmed via the histology slice (h)).

ω_t for f^i is given as $\hat{P}(c \leftrightarrow \omega_t | f^i)$. $\hat{P}(c \leftrightarrow \omega_t | \mathbf{f})$, for $\mathbf{f} = [f^i | i \in \{1, \dots, K\}]$, is the combined posterior likelihood obtained by combining $\hat{P}(c \leftrightarrow \omega_t | f^i)$, for $1 \leq i \leq K$. $\tilde{p}(f^i | c \leftrightarrow \omega_t)$, $\tilde{P}(c \leftrightarrow \omega_t | f^i)$, and $\tilde{P}(c \leftrightarrow \omega_t | \mathbf{f})$ denote the corresponding prior, posterior, and combined posterior likelihoods obtained after refinement by embedding. $\hat{\mathcal{L}} = (C, \hat{l})$ denotes the combined likelihood scene (Fig. 3(d)), such that for $c \in C$, $\hat{l}(c) = \hat{P}(c \leftrightarrow \omega_t | \mathbf{f})$. $\tilde{\mathcal{L}} = (C, \tilde{l})$, where for $c \in C$, $\tilde{l}(c) = \tilde{P}(c \leftrightarrow \omega_t | \mathbf{f})$, similarly denotes the corresponding likelihood scene (Fig. 3(e)) after refinement by graph embedding.

3.2 Feature Extraction and Classification

A total of 35 3D texture feature scenes $\mathcal{F}^i = (C, f^i)$, for $1 \leq i \leq 35$, are obtained from the MR scene \mathcal{C} . The extracted features include 7 first order statistical features at two scales, 8 Haralick features at two scales, 2 gradient features, and 18 Gabor features corresponding to 6 different scales and 3 different orientations. A more detailed description of the feature extraction methods has been previously presented in [5]. The *a-posteriori* likelihoods $\hat{P}(c \leftrightarrow \omega_j | f^i)$ for each feature f^i can be computed using Bayes Theorem [6] as, $\hat{P}(c \leftrightarrow \omega_j | f^i) = \hat{P}(c \leftrightarrow \omega_j) \frac{\hat{p}(f^i | c \leftrightarrow \omega_j)}{\hat{p}(f^i)}$, where $\hat{P}(c \leftrightarrow \omega_j)$ is the *a-priori* probability of observing the class ω_j , $\hat{p}(f^i) = \sum_{j=1}^{\mathcal{B}} \hat{p}(f^i | c \leftrightarrow \omega_j) \hat{P}(c \leftrightarrow \omega_j)$, where \mathcal{B} refers to the number of classes. The combined posterior likelihood $\hat{P}(c \leftrightarrow \omega_j | \mathbf{f})$, for $\mathbf{f} = [f^i | i \in \{1, \dots, K\}]$, can be obtained from $\hat{P}(c \leftrightarrow \omega_j | f^i)$, by using any of the various feature ensemble methods, e.g. ensemble averaging, GEM [5], majority voting.

3.3 Graph Embedding for Analyzing Class Relationships

Our aim is to find a placement (embedding) vector $\hat{X}(c)$ for each voxel $c \in C$ and the tumor class ω_t such that the distance between c and class ω_t is monotonically related to the *a-posteriori* probability $\hat{P}(c \leftrightarrow \omega | \mathbf{f})$ in a low-dimensional space [2]. Hence if voxels $c, d \in C$ both belong to class ω_t , then $[\hat{X}(c) - \hat{X}(d)]^2$ should be small. To compute the optimal embedding, we first define a confusion matrix W representing the similarity between any two objects $c, d \in C$ in a high dimensional feature space.

$$W(c, d) = e^{-\|\hat{P}(c \leftrightarrow \omega_t | \mathbf{f}) - \hat{P}(d \leftrightarrow \omega_t | \mathbf{f})\|} \in \mathcal{R}^{|C| \times |C|} \tag{1}$$

Computing the embedding is equivalent to optimization of the following function,

$$E_W(\hat{X}) = \frac{\sum_{(c,d) \in C} W(c, d) (\hat{X}(c) - \hat{X}(d))^2}{\sigma_{\hat{X}}^2} \tag{2}$$

Expanding the numerator of (2) we get $2\hat{X}^T(D - W)\hat{X}$, where $D(c, d) \in \mathcal{R}^{|C| \times |C|}$ is a diagonal matrix with $D(c, c) = \sum_d W(c, d)$. Using the fact that

$$\sigma_{\hat{X}}^2 = \sum_{c \in C} \hat{X}^2(c) \hat{P}(c \leftrightarrow \omega_t | \mathbf{f}) - \left(\sum_{c \in C} \hat{X}(c) \hat{P}(c \leftrightarrow \omega_t | \mathbf{f}) \right)^2, \tag{3}$$

it can be shown that $\hat{P}(c \hookrightarrow \omega_t | \mathbf{f}) \approx \frac{1}{\gamma} D(c, c)$, where $\gamma = |C| - 1$, and $|C|$ represents the cardinality of set C . Centering the embedding around zero (i.e. $\hat{X}^T \gamma = 0$), we get $\sigma_{\hat{X}}^2 = \frac{1}{\gamma} \hat{X}^T D \hat{X}$. Putting all these together we can rewrite (2) as,

$$E_W(\hat{X}) = 2\gamma \frac{\hat{X}^T (D - W) \hat{X}}{\hat{X}^T D \hat{X}}. \tag{4}$$

The global energy minimum of this function is achieved by the eigenvector corresponding to the second smallest eigenvalue of,

$$(D - W) \hat{X} = \lambda D \hat{X}. \tag{5}$$

For voxel $c \in C$, the embedding $\hat{X}(c)$ contains the coordinates of c in the embedding space and is given as, $\hat{X}(c) = [\hat{e}_a(c) | a \in \{1, 2, \dots, \beta\}]$, where $\hat{e}_a(c)$, are the eigen values associated with c .

3.4 Improving Training Distributions by Refining Ground Truth

In several classification tasks (especially in medical imaging), S_{ω_t} , the set of *true ground truth* class labels is not available. For the CAD problem tackled in this work, only an approximation of the ground truth (\hat{S}_{ω_t}) is available, so that there exist objects $d \in \hat{S}_{\omega_t}$ which do not belong to class ω_t . Consequently the prior distributions $\hat{p}(f^i | c \hookrightarrow \omega_t)$, for $1 \leq i \leq K$, and the posterior probabilities $\hat{P}(c \hookrightarrow \omega_t | f^i)$ reflect the errors in \hat{S}_{ω_t} , since $\hat{p}(f^i | c \hookrightarrow \omega_t)$ is generated from a training set $\hat{S}_{\omega_t}^T \subset \hat{S}_{\omega_t}$. Clearly a more accurate estimate (\tilde{S}_{ω_t}) of S_{ω_t} would result in more accurate prior distributions $\tilde{p}(f^i | c \hookrightarrow \omega_t)$, for $1 \leq i \leq K$, and consequently a more accurate posterior likelihoods $\tilde{P}(c \hookrightarrow \omega_t | f^i)$. To obtain \tilde{S}_{ω_t} we proceed as follows,

- (1) The embedding of all voxels $c \in C$, $\hat{X}(C)$ is determined.
- (2) The K -means algorithm is applied on the embedding coordinates $\hat{X}(C)$ to cluster objects $c \in C$ into Z disjoint partitions $\{\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_Z\}$.
- (3) We obtain the union of those disjoint partitions \mathcal{P}_z , for $1 \leq z \leq Z$, sizes of which are above a pre-determined threshold θ . The rationale behind this is that outliers will be partitioned into small sets. \tilde{S}_{ω_t} is then obtained as,

$$\tilde{S}_{\omega_t} = \hat{S}_{\omega_t} \bigcap \left[\bigcup_z \mathcal{P}_z \right], \text{ where } |\mathcal{P}_z| \geq \theta, \text{ for } z \in \{1, 2, \dots, Z\}. \tag{6}$$

The intuition behind Equation 6 is that we only consider objects in \hat{S}_{ω_t} for inclusion into \tilde{S}_{ω_t} . This avoids inclusion of potentially new outliers. Note that, since this procedure is only for the training step, we are not concerned with including every object in class ω_t into \tilde{S}_{ω_t} . Instead, our aim is to ensure as far as possible that for every object $c \in \hat{S}_{\omega_t}$, $c \hookrightarrow \omega_t$.

- (4) New *a-priori* distributions $\tilde{p}(f^i | c \hookrightarrow \omega_t)$, for $1 \leq i \leq K$, are then generated from training set $\tilde{S}_{\omega_t}^T \subset \tilde{S}_{\omega_t}$ and the new posterior likelihoods $\tilde{P}(c \hookrightarrow \omega_t | f^i)$ and combined likelihood $\tilde{P}(c \hookrightarrow \omega_t | \mathbf{f})$, for $\mathbf{f} = [f^i | i \in \{1, \dots, K\}]$, are computed.

Fig. 3(c), (e) correspond to the *likelihood scenes* ($\hat{\mathcal{L}}, \tilde{\mathcal{L}}$) obtained from distributions $\hat{p}(f^i|c \rightarrow \omega_t)$ and $\tilde{p}(f^i|c \rightarrow \omega_t)$ respectively. The intensity at every voxel $c \in C$ in Fig. 3(c), (e) is given by the *a-posteriori* likelihoods $\hat{P}(c \rightarrow \omega_t|\mathbf{f})$ and $\tilde{P}(c \rightarrow \omega_t|\mathbf{f})$, for $\mathbf{f} = [f^i|i \in \{1, \dots, K\}]$, respectively. While Fig. 3(e) is almost a bi-level image, suggesting distinct separation between the tumor and non-tumor classes, Fig. 3(c) is more fuzzy, indicating considerable overlap between the two classes. This is reflected in the plot of class embeddings $\hat{X}(C)$ obtained from $\hat{P}(c \rightarrow \omega_t|\mathbf{f})$ in which considerable overlap (ellipse 3) exists between the two classes (Fig. 3(d)), while in the plot of $\tilde{X}(C)$, the graph embedding obtained from $\tilde{P}(c \rightarrow \omega_t|\mathbf{f})$ (Fig. 3(f)), there is a more distinct separation of class clusters.

3.5 Discovering Novel Classes

Even after removing outliers from the ground truth, there exist objects that occupy the transition between tumor and non-tumor clusters (observe ellipse 3 in Fig. 3(f)), suggesting that the characteristics of these objects are between that of the tumor and benign classes. In Fig. 3(g) is shown a novel way of visualizing and identifying objects from these intermediate classes. Since $\tilde{X}(c)$ contains the embedding coordinates of voxel c , we can represent $\tilde{X}(C)$, the embedding over scene C , as a RGB image in which the value at voxel c is given by the three principal eigen values associated with c . Objects that are adjacent to each other in the embedding space have a similar color (Fig. 3(g)). The apparent false positive area (labeled as FP on Fig. 3(g)), on inspecting the corresponding histology slice (Fig. 3(h)) was found to be inflammation induced by atrophy on account of a prior needle insertion. This new class had not been considered in our original two class classification paradigm.

3.6 Algorithm

- For each scene we compute the corresponding feature scenes for each feature f^i .
- Prior distributions $\hat{p}(f^i|c \rightarrow \omega_t)$ for each feature f^i for class ω_t are obtained using training set $\hat{S}_{\omega_t}^T \subset \hat{S}_{\omega_t}$.
- Bayes Theorem [6] is used to compute posterior likelihoods $\hat{P}(c \rightarrow \omega_t|f^i)$, for $1 \leq i \leq K$. Combined likelihood $\hat{P}(c \rightarrow \omega_t|\mathbf{f})$, for $\mathbf{f} = [f^i|i \in \{1, \dots, K\}]$ is then computed from $\hat{P}(c \rightarrow \omega_t|f^i)$ using any standard ensemble method.
- Confusion matrix W is computed for $c, d \in C$ as $W(c, d) = e^{-|\hat{P}(c \rightarrow \omega_t|\mathbf{f}) - \hat{P}(d \rightarrow \omega_t|\mathbf{f})|} \in \mathcal{R}^{|C| \times |C|}$. Solve for the smallest eigen vectors of $(D - W)\hat{X} = \lambda D\hat{X}$ where the rows of the eigen vectors are the coordinates for the object c in the embedding space $\hat{X}(C)$.
- Partition $\hat{X}(C)$ into disjoint regions \mathcal{P}_z , for $1 \leq z \leq Z$, and compute new set of tumor class objects $\tilde{S}_{\omega_t} = \hat{S}_{\omega_t} \cap [\bigcup_z \mathcal{P}_z]$, where $|\mathcal{P}_z| \geq \theta$.
- Generate new prior distributions $\tilde{p}(f^i|c \rightarrow \omega_t)$, for $1 \leq i \leq K$, from new training set $\tilde{S}_{\omega_t}^T \subset \tilde{S}_{\omega_t}$ and compute new posterior likelihoods $\tilde{P}(c \rightarrow \omega_t|f^i)$ and combined posterior likelihood $\tilde{P}(c \rightarrow \omega_t|\mathbf{f})$, for $\mathbf{f} = [f^i|i \in \{1, \dots, K\}]$.

4 Evaluating CAD Accuracy for Prostate Cancer on MRI

The likelihood scene $\hat{\mathcal{L}}$ is thresholded to obtain binary scene $\hat{\mathcal{L}}_B = (C, \hat{l}_B)$ so that for $c \in C$, $\hat{l}_B(c)=1$ iff $\hat{l}(c) \geq \delta$, where δ is a pre-determined threshold. $\tilde{\mathcal{L}}_B$ is similarly obtained. $\hat{\mathcal{L}}_B$ and $\tilde{\mathcal{L}}_B$ are then compared with \hat{S}_{ω_t} and \tilde{S}_{ω_t} respectively to determine Sensitivity and Specificity values for different values of δ . Receiver operating characteristic (ROC) curves (plot of Sensitivity versus 100-Specificity) provide a means of comparing the performance of detection tasks. A larger area under the ROC curve implies higher accuracy. A total of 33 MR images of the prostate were used for quantitatively comparing $\tilde{\mathcal{L}}$ and $\hat{\mathcal{L}}$ for different values of δ . Fig. 4(a) and (b) show the ROC curves for $\hat{\mathcal{L}}$ (dashed line) and $\tilde{\mathcal{L}}$ (solid line) for two different feature combination methods (ensemble averaging and majority voting) using 5 and 10 training samples respectively. The accuracy of $\tilde{\mathcal{L}}$ was found to be significantly higher compared to $\hat{\mathcal{L}}$ for both classification methods and different sets of training samples, as borne out by the larger area under the ROC curves in Fig. 4(a) and (b). All differences were found to be statistically significant.

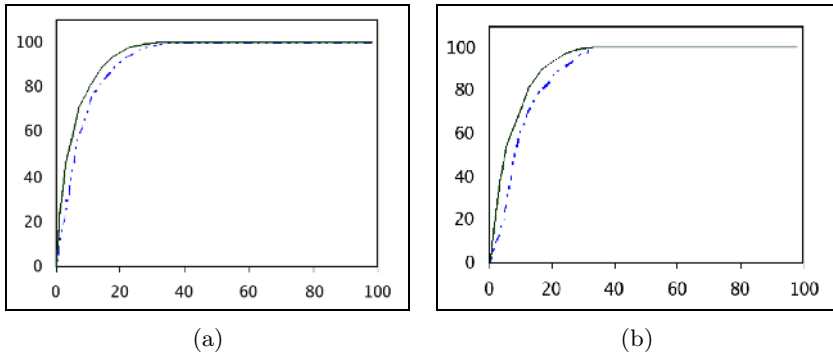


Fig. 4. ROC analysis of $\hat{\mathcal{L}}$ (dashed line) and $\tilde{\mathcal{L}}$ (solid line) using (a) ensemble averaging and 5 training samples, and (b) majority voting and 10 training samples

5 Concluding Remarks

In this paper we have presented a novel application of graph embedding in (i) improving the accuracy of supervised classification schemes, especially in cases where object class labels cannot be reliably ascertained, and (ii) for identifying novel classes of objects not present in the original classification. We have successfully employed this method to improve the accuracy of a CAD system for detecting prostate cancer from high resolution MR images. We were also able to identify a new class (inflammation due to atrophy). The method could be similarly used to detect pre-cancerous lesions, the presence of which has significant clinical implications.

References

1. I. Dhillon, D. Modha, W. Spangler, “Class Visualization of high-dimensional data with applications”, *Computational Statistics & Data Analysis*, 2002, vol. 41, pp. 59-90.
2. T. Iwata, K. Saito, *et al.*, “Parametric Embedding for Class Visualization”, *NIPS*, 2004.
3. A. Globerson, G. Chechik, *et al.*, “Euclidean Embedding of Co-occurrence Data”, *NIPS*, 2004.
4. H. Zhong, J. Shi and M. Visontai, Detecting Unusual Activity in Video, *CVPR*, 2004.
5. A. Madabhushi, M. Feldman, D. Metaxas, J. Tomaszeweski, D. Chute, Automated Detection of Prostatic Adenocarcinoma from High Resolution *in vitro* prostate MR studies, *IEEE Trans. Med. Imag.*, Accepted.
6. R. Duda, P. Hart, *Pattern Classification and Scene Analysis*, New York Wiley, 1973.
7. T. Joliffe, *Principal Component Analysis*, *Springer-Verlag*, 1986.