Training a CAD Classifier with Different Polyp Morphologies for Virtual Colonoscopy

Murat Dündar, Matthias Wolf, Sarang Lakare, and Marcos Salganicoff

IKM CAD & Knowledge Solutions
Siemens Medical Solutions Inc., USA
51 Valley Stream Parkway, MS E51
Malvern, PA 19355
{murat.dundar,mwolf,
sarang.lakare,marcos.salganicoff}@siemens.com

Abstract. In this study we introduce a learning algorithm for dealing with different polyp morphologies in designing a Computer Aided Detection system for Virtual Colonoscopy. The proposed approach takes advantage of the subclass information available for polyp candidates in the training data and jointly optimizes multiple hyperplane classifiers each of which is designed to classify negative candidates from a subclass of polyp candidates. This yields a polyhedral decision surface with flat faces with the number of such faces equivalent to the number of different polyp morphologies. Flat faces provides robustness to the classifier whereas multiple faces contributes to the flexibility required to deal with different polyp types. We evaluate the performance of the proposed technique on a real-world Colon dataset and compare the proposed classifier against the state-of-the-art linear classifier and a multi-class classifier in terms of the area under the receiver operating characteristics (ROC) curves in the clinically admissible range of 0-4fp/vol average.

1 Introduction

In this study we propose a learning algorithm that deals with the multi-mode nature of the training data. Even though the technique is mainly motivated by the existence of different polyp morphologies in colorectal cancer and currently validated only with our Colon dataset, we believe it can be applied to any target detection problem with multi-mode data characteristics.

Typical examples of different polyp morphologies are given in Fig. 1. A polyp with a broad base is called “sessile”; if it has a separate stalk it is called “pedunculated”. A polyp is considered “flat” if the width to height ratio is greater than 2 or the vertical elevation above the colon wall is less than 3mm. Even though there are certain attributes shared across all three polyp morphologies such as all protruding from the colon wall, having certain thickness, and having similar intensity values, the dominant image and shape characteristics that allow us to easily distinguish one polyp morphology from other visually makes it very unrealistic to assume that candidates representing different polyp morphologies all belong to the same distribution in the feature space.

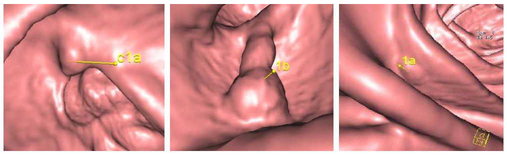

Fig. 1. Polyp morphologies (from left to right): Sessile, pedunculated, and flat polyp

1.1 Related Work

Traditionally the classifier is trained by pooling all candidates representing different subclasses, i.e. polyp morphologies into a single positive class and learning a classifier that will distinguish as many members of the positive class from a large cluster of negatives in the feature space while making sure certain regularization criteria are satisfied to avoid overfitting. The classifier trained this way does not address the multi-mode nature of the data and leaves room for potential improvements.

One way to tackle with this problem is finite mixture models (McLachlan & Peel, 2004). The positive class can be modeled by a mixture model, one mode for each subclass, and a maximum likelihood classifier can be designed to classify positive candidates from the negative ones. Given that the positive data is very scarce and the number of features are relatively large it is almost impractical to estimate mixture model parameters (covariance matrices and means of each mode in the case of a normal mixture model) in the high dimensional feature space without encountering numerical issues.

Discriminative techiques such as Support Vector Machines (Vapnik, 1995), Kernel Fisher Discriminant (Mika et al., 2000), Relevance Vector Machines (Tipping, 2000) to name few are also used in this domain. These techniques deal with the unbalanced nature of the data by assigning different cost factors to the negative and positive samples in the objective function. The kernel evaluation of these techniques yields nonlinear decision boundaries suitable for classifying multi-mode data. Even though kernel-based classifiers have the capacity to learn higly nonlinear decision boundaries allowing great flexibility, it is well-known that in real-world applications where feature noise and redundancy is a problem, too much capacity usually hurts the generalizability of a classifier. Previous studies in the CAD domain (undisclosed, 2004) shows that kernel-based nonlinear classifiers are more prone to overfitting training data and yielding poor generalization performance than linear classifiers.

One acceptable techique in this domain would be to design a series of hyperplane classifiers one for each of different subclass, each trained independently to differentiate candidates of a given subclass from the negative class. The final classification decision is to assign a given candidate to the positive class, if the candidate is classified as positive by any one of the classifiers. The main problem with this approach is that since the classifiers are trained independently, each classifier only utilizes a portion of the positive samples. That is to say already scarce positive data is split among the classifiers and as a result individual classifiers potentially become more prone to overfitting.

Multiple linear classifiers brings us the flexibility to model multi-mode data without refering to nonlinear classifiers whereas the linearity of each individual classifier provides robustness, i.e. potentially makes the classifier more resistant to overfitting. In

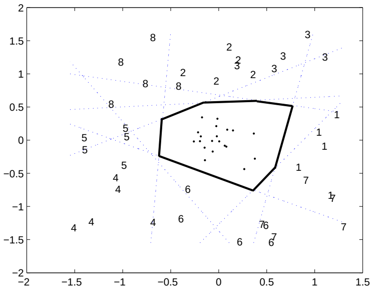

Fig. 2. A schematic example demonstrating the proposed algorithm. Dark circles depicting negative samples, numbers representing positive samples. The decision boundary is shown with the solid lines.

that regard this is a promising approach. However having to train each classifier independently and making use of only a small portion of the positive samples at each time makes this approach less desirable from a machine learning perspective.

2 Proposed Approach

In this study we propose a methodology to jointly optimize multiple hyperplane classifiers to learn a polyhedral decision surface. The number of such hyperplane classifiers is equivalent to the number of subclasses identified in the positive class. Each of these classifiers is designed to classify candidates belonging to the corresponding subclass from all of the negative candidates.

In Figure 2 the proposed algorithm is demonstrated with a toy example. Negative samples are depicted by the dark circles in the middle, whereas positive samples are depicted with the numbers with each number corresponding to a different subclass. All eight classifiers are optimized simultaneously and polygon shown with dark lines is obtained as a decision boundary that classifies positive samples from the negative ones.

We are now ready to formulate the proposed algorithm. We first start with a brief overview of hyperplane classifier with hinge-loss, which is also the foundation for Support Vector Machines.

2.1 Hyperplane Classifiers

We are given a training dataset $\{(x_i, y_i)\}_{i=1}^{\ell}$, where $x_i \in \Re^d$ are input variables and $y_i \in \{-1, 1\}$ are class labels. We consider a class of models of the form $f(x) = \alpha^T x$, with the sign of $f(x)$ predicting the label associated with the point x. An hyperplane classifier with hinge loss can be designed by minimizing the following cost function.

$$\mathcal{J}(\alpha) = \varPhi(\alpha) + \sum_{i=1}^{\ell} w_i \left(1 - \alpha^T y_i x_i\right)_+ \tag{1}$$

where the function $\Phi : \Re^{(d)} \Rightarrow \Re$ is a regularization function or regularizer on the hyperplane coefficients and $(k)_+ = max(0, k)$ represents the hinge loss, and $\{w_i : w_i \geq 0, \forall i\}$ is the weight preassigned to the loss associated with x_i. For balanced data usually $w_i = \nu$, but for unbalanced data it is a common practice to weight positive and negative classes differently, i.e. $\{w_i = \nu_1, \ \forall i \in C^+\}$ and $\{w_i = \nu_2, \ \forall i \in C^-\}$ where C^+ and C^- are the corresponding sets of indices for the positive and negative classes respectively.

The function $\left(1 - \alpha^T y_i x_i\right)_+$ is a convex function. The weighted sum of convex functions is also convex. Therefore for a convex function $\Phi(\alpha)$ (1) is also convex. The problem in (1) can be formulated as a mathematical programming problem as follows:

$$\begin{array}{rl} \min\limits_{(\alpha,\xi) \in R^{d+\ell}} & \Phi(\alpha) + \sum_{i=1}^{\ell} w_i \xi_i \\ \text{s.t.} & \xi_i \geq 1 - \alpha^T y_i x_i \\ & \xi_i \geq 0, \ \forall i \end{array} \tag{2}$$

For $\Phi(\alpha) = \|\alpha\|_2^2$, where $\|.\|_2$ is the 2-norm, (2) results in the conventional Quadratic-Programming-SVM, and for $\Phi(\alpha) = |\alpha|$, where $|.|$ is the 1-norm it yields the sparse Linear-Programming-SVM.

2.2 Polyhedral Decision Boundaries

Joint learning of multiple hyperplanes can be achieved by optimizing the following cost function

$$\mathcal{J}(\alpha_1, \ldots \alpha_K) = \sum_{k=1}^{K} \Phi_k(\alpha_k) + \nu_1 \sum_{k=1}^{K} \sum_{i \in C_k^+} (e_{ik})_+ + \nu_2 \sum_{i \in C^-} \max\left(0, e_{i1}, \ldots, e_{iK}\right) \tag{3}$$

where $e_{ik} = 1 - \alpha_k^T y_i x_{ik}$, and $(e_{ik})_+, \ \forall i \in C_k^+$, defines the hinge loss of the i-th training example $\{(x_{ik}, y_{ik})\}$ in subclass-k induced by classifier k. C_k^+ is the set of indices of the positive samples in subclass-k. Note that classifier k is designed to classify positive examples in the subclass-k from the negative examples. The first term in (3) is a summation of the regularizers for each of the classifiers and the second and third terms accounts for the losses induced by the positive and negative samples respectively. Unlike (1) the loss function here is different for the negative samples. The loss induced by a negative sample i, $i \in C^-$ is zero only if $\forall k : 1 + \alpha_k^T x_i \leq 0$, which corresponds to the “AND” operation. The problem (3) can be formulated as follows

$$\begin{array}{rl} \min\limits_{(\alpha,\xi) \in R^{Kd+\ell}} & \sum_{k=1}^{K} \Phi_k(\alpha_k) + \nu_1 \sum_{k=1}^{K} \sum_{i \in C_k^+} \xi_{ik} \\ & \qquad + \nu_2 \sum_{i \in C^-} \xi_i \\ \text{s.t.} & \xi_{ik} \geq 1 - \alpha_k^T x_{ik} \\ & \xi_{ik} \geq 0 \\ & \xi_i \geq 1 + \alpha_k^T x_i \\ & \xi_i \geq 0 \end{array} \tag{4}$$

where the first two constraints are imposed for $\forall i \in C_k^+$, $k = 1, ..., K$ and the last two constraints are imposed for $\forall i \in C^-$, $k = 1, ..., K$. Note that for a convex function $\Phi(\alpha)$ the problem in (4) is convex. In a nutshell we designed K classifiers, one for each of the binary classification problems, i.e. subclass-k of the positive class vs negative class. Then we construct a learning algorithm to jointly optimize these classifiers such that the cost induced by a negative sample is zero if and only if all of the K classifiers classifies this sample correctly, i.e. $\forall k : 1 + \alpha_k^T x_i \leq 0$. Since each positive sample is only used once for training the classifier k, the cost induced for a positive sample is zero as long as it is classified correctly by the corresponding classifier k, i.e. $1 - \alpha_k^T x_{ik} \leq 0$.

3 Experimental Results

We validate the proposed polyhedral classifier denoted as *polyhedral* with respect to its generalization performance. We compared its performance against two other techniques. The first one is a hyperplane classifier with hinge loss denoted as *sparse svm* and the second one is a classifier obtained by multiple hyperplane classifiers each trained independently, denoted as *multi-class*. Throughout the experiments we set the $\Phi_k(\alpha_k) = |\alpha_k|$ for all three techniques.

3.1 Data and Experimental Settings

The database of high-resolution CT images used in this study were obtained from two different sites across US. The 370 patients were randomly partitioned into two groups: training (n=167) and test (n=199). The test group was sequestered and only used to evaluate the performance of the final system.

Training Data Patient and Polyp Info: There were 170 patients with 340 volumes. A total of 185 polyps were identified as ground truth at the volume level in the 6-25mm range. The candidate generation algorithm generates a total of 67436 false positives or on the average 198 false positives per volume (fp/vol).

Testing Data Patient and Polyp Info: There were 201 patients with 395 volumes. A total of 223 polyps were identified as ground truth at the volume level in the 6-25mm range. The candidate generation algorithm generates a total of 79057 false positives or 200 false positives per volume (fp/vol).

A total of 101 features are extracted to capture shape and intensity characteristics of each candidate. Three different polyp types are identified, namely flat, pedunculated and sessile. Sparse SVM is trained using all three polyp categories as one class. Polyhedral and Multi-class techniques each learn a hyperplane classifier for each of the three polyp types with the former one learning these classifiers jointly and the latter one independently. The classifier paramaters ν_1 and ν_2 are jointly estimated through 10-fold patient-wise cross validation technique over the training data such that the area under the receiver operating characteristics (ROC) curve defined by the average false positive per volume values of 0 to 4 is optimized. For both ν_1 and ν_2 a discrete set of five values are considered.

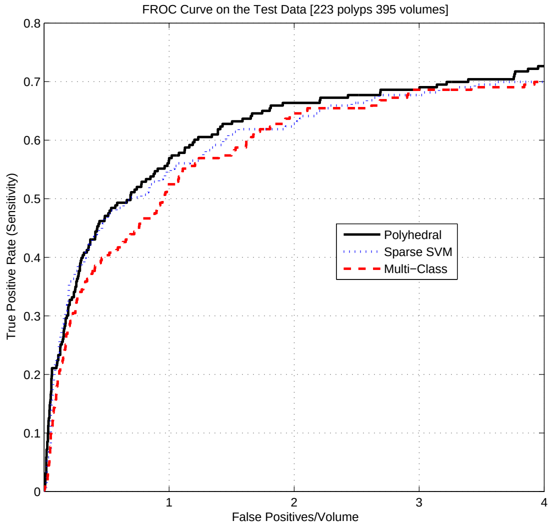

Fig. 3. FROC curves obtained by the three classifiers on the test data.

3.2 Performance Evaluation

As shown in Figure 3 the Free-response ROC (FROC) curve corresponding to the proposed *polyhedral* classifier is consistently dominating the other two curves across almost the entire region of interest indicating superior prediction performance. At the same fp/vol we see that the polyhedral classifier detects as many as 15 polyps more than the multi-class classifier, and 10 polyps more than the sparse SVM.

4 Conclusions

In this study we have presented a methodology to take advantage of the subclass information available in the positive class while training a classifier. The subclass information which is neglected in conventional binary classifiers provides a better insight of the dataset and when incorporated into the learning mechanism acts as an implicit regularizer on the classifier coefficients. We believe this is an important contribution for applications where the number of positive candidates is limited and feature noise is prevalent.

Bibliography

McLachlan, G., & Peel, D. (2004). *Finite mixture models*. Wiley-Interscience.

Mika, S., Rätsch, G., & Müller, K.-R. (2000). A mathematical programming approach to the kernel fisher algorithm. *NIPS* (pp. 591–597).

Tipping, M. E. (2000). The relevance vector machine. In S. Solla, T. Leen and K.-R. Muller (Eds.), *Advances in neural information processing systems 12*, 652–658. Cambridge, MA: MIT Press.

undisclosed (2004). .

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer.