Multiple-instance learning improves CAD detection of masses in digital mammography

Balaji Krishnapuram[1], Jonathan Stoeckel[2], Vikas Raykar[1], Bharat Rao[1], Philippe Bamberger[2], Eli Ratner[2], Nicolas Merlet[2], Inna Stainvas[2], Menahem Abramov[2], and Alexandra Manevitch[2]

[1] CAD and Knowledge Solutions (IKM CKS), Siemens Medical Solutions Inc.,
Malvern PA 19355, USA,
[2] Siemens Computer Aided Diagnosis Ltd., Jerusalem, Israel

Abstract. We propose a novel *multiple-instance learning* (MIL) algorithm for designing classifiers for use in *computer aided detection* (CAD). The proposed algorithm has 3 advantages over classical methods. First, unlike traditional learning algorithms that minimize the candidate level misclassification error, the proposed algorithm directly optimizes the patient-wise sensitivity. Second, this algorithm automatically selects a small subset of statistically useful features. Third, this algorithm is very fast, utilizes all of the available training data (without the need for cross-validation etc.), and requires no human hand tuning or intervention. Experimentally the algorithm is more accurate than state of the art support vector machine (SVM) classifier, and substantially reduces the number of features that have to be computed.

1 Background

Traditionally, in an almost universal architecture, CAD algorithms operate in a sequence of three stages. In the first stage, a candidate generation (CG) algorithm identifies suspicious regions. In the second stage each suspicious region is characterized by a set of features. In the third, classification stage, each region is evaluated in light of the features and a decision is made whether the region is sufficiently suspicious that it should be highlighted to a radiologist. This paper focuses largely on the design of the classifier for the third stage of this architecture.

Many off-the-shelf classifier learning algorithms have been used during the design of CAD algorithms, e.g. support vector machines (SVM) [1], neural networks (NN) [3], etc. However, the derivations behind most of these algorithms make unwarranted assumptions that are violated in CAD data sets. For example, most classifier-learning algorithms assume that the training samples are independent and identically distributed (i.i.d.). However, there are high levels of correlations among the suspicious locations from the same region of a breast (both within a breast image, and across multiple images of the same breast), so the training samples are clearly not independent.

Further, these standard algorithms try to maximize classification accuracy over all candidates. However, this particular accuracy measure is not very relevant for CAD. For example, often several candidates generated from the CG point to the same underlying lesion in a breast. Even if one of these is highlighted to the radiologist and other adjacent or overlapping candidates are missed, the underlying lesion (hence the patient or image) would have been detected. Hence, CAD system accuracy is measured in terms of Free-Response Operator Characteristic (FROC) curves plotting per-image (or per-patient sensitivity when multiple images are available for a patient), versus false-alarms (FA) per-case. This is only loosely related to the accuracy-measure optimized by off-the-shelf methods.

Previous work presented to the machine-learning community has shown that modeling CAD classifier learning as a multiple-instance learning (MIL) problem largely alleviates the above concerns [2]. However, that paper was not targeted to the applied radiology community, and the algorithm was not applied to mammography. Compared to previous (non-probabilistic) MIL algorithms, the method proposed in this paper is much faster in terms of run-time, needs no specialized optimization packages, needs no parameter tuning on validation sets (or cross validation), and automatically selects a small set of diagnostically useful features.

2 Method

2.1 Notation

Consider a parametric family of classification functions $p(y|\mathbf{x}, \mathbf{w})$ that take a d-dimensional feature vector $\mathbf{x} \in \mathbf{R}^d$ as input, and produces the probability that the sample $\mathbf{x}$ belongs to one of the two classes $y \in \{0, 1\}$. This family of functions is parameterized by the weight vector $\mathbf{w} \in \mathbf{R}^d$. To learn this classification function we are given N training samples $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$.

2.2 Logistic Regression

An example of a parametric family is *logistic regression.* The posterior probability for the positive class is modeled as

$$p(y = 1|\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^\top \mathbf{x}), \tag{1}$$

where $\sigma(z) = 1/(1 + e^{-z})$ is the *logistic sigmoid* link function. The notation is overloaded so that $\sigma(\mathbf{z})$ of a vector $\mathbf{z} = [z_1, z_2, \ldots, z_N]^\top$ denotes the element wise application of the sigmoid function to each element of the vector, *i.e.*, $\sigma(\mathbf{z}) = [\sigma(z_1), \sigma(z_2), \ldots, \sigma(z_N)]^\top$. More generally, for example in a two layer neural network, the weight $\mathbf{w}$ could consist of a weight matrix $\mathbf{W}_1$ and a weight vector $\mathbf{w}_2$, so that $p(y = 1|\mathbf{x}, \mathbf{w} = \{\mathbf{W}_1, \mathbf{w}_2\}) = \sigma(\mathbf{w}_2^\top \sigma(\mathbf{W}_1^\top \mathbf{x}))$. For binary classification, $p(y = 0|\mathbf{x}, \mathbf{w}) = 1 - p(y = 1|\mathbf{x}, \mathbf{w})$. Learning a classifier implies choosing the weight vector $\mathbf{w}$ given the training data $\mathcal{D}$.

2.3 Traditional statistical learning of classifiers

The following theory is very general and extends to many non-linear classifier models $p(y|\mathbf{x}, \mathbf{w})$ such as Neural Networks, Gaussian Processes, etc. Though we have implemented our Multiple-Instance Learning method to extend several baseline classifiers, for succinct presentation we shall focus on the binary logistic regression model $p(y = 1|\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^\top \mathbf{x})$, in the rest of this paper.

Maximum-Likelihood estimator Traditional learning methods assume that the samples are drawn i.i.d., so the overall log-likelihood for the entire training data set factorizes as

$$\begin{aligned} l(\mathbf{w}) &= \log p(y_1, y_2 \ldots, y_N|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N, \mathbf{w}) = \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i, \mathbf{w}) \\ &= \sum_{\forall i|y_i=1} \log \sigma(\mathbf{w}^\top \mathbf{x}_i) + \sum_{\forall i|y_i=0} \log(1 - \sigma(\mathbf{w}^\top \mathbf{x}_i)). \end{aligned} \tag{2}$$

Most training algorithms maximize this log-likelihood. However the maximum-likelihood solution in practice can exhibit severe over-fitting especially for high-dimensional data. This is addressed by using a prior on w and then finding the *maximum a-posteriori* (MAP) solution.

Maximum a-posteriori estimator Using Bayes's rule the log-posterior L can be written as

$$L(\mathbf{w}) = l(\mathbf{w}) + \log p(w) - \log p(y_1, y_2 \ldots, y_N|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N). \tag{3}$$

Since the last term is independent of $\mathbf{w}$ the MAP estimate is given by

$$\widehat{\mathbf{w}}_{\text{MAP}} = \arg\max_{\mathbf{w}} \left[l(\mathbf{w}) + \log p(\mathbf{w}) \right]. \tag{4}$$

Often one uses a zero mean Gaussian prior $(\mathcal{N}(\mathbf{w}|0, \mathbf{A}^{-1}))$ on the weights w with inverse-covariance (precision) matrix $\mathbf{A} = diag(\alpha_1 \ldots \alpha_d)$ (also referred to as the *hyper-parameters*).

$$p(\mathbf{w}) = (2\pi)^{-d/2} |\mathbf{A}^{-1}|^{-1/2} \exp\left(-\frac{\mathbf{w}^\top \mathbf{A} \mathbf{w}}{2} \right). \tag{5}$$

As $\alpha_i \rightarrow 0$, the prior becomes diffuse-*i.e.* more or less uniform-resulting in maximum-likelihood estimate. However, as $\alpha_i \rightarrow \infty$, the prior is sharply concentrated around 0, preventing the magnitude of $||\mathbf{w}||^2$ from growing large regardless of the training data, reducing the risk of over-fitting to the training set.

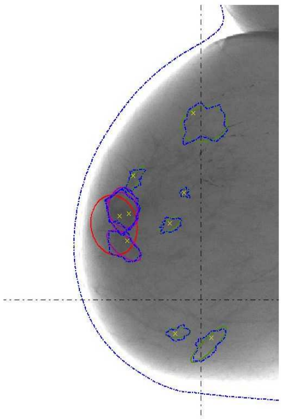

Fig. 1. An MLO view of the right breast illustrating the concept of multiple candidates pointing to the same ground truth. The red ellipse is the lesion as marked by the radiologist (ground truth). The blue contours are the candidates generated by our algorithm. All the blue contours which overlap with the red ellipse share the same ground truth and constitute our notion of a positive bag.

2.4 Multiple Instance Learning for classifiers

During the development of a training set, a CG runs on a set of cases, and then features are extracted for the candidates identified in this stage. Next, we identify which of these candidates overlaps with a radiologist marked ground-truth (lesion), and label these as positive candidates, and the rest are labeled as negative. During this process, we obtain information about which candidates point to the same underlying ground-truth lesion (See Fig. 1). While this information is typically discarded during the development of traditional classifiers, we propose to utilize this information to extract more statistical power from the data.

In particular we rely on the notion that all the positive candidates (with label $y_i = 1$) that point to the same radiologist marked ground-truth belong to the same *positive bag* (See Fig. 1); the training set consists of a large number of such bags corresponding to the number of lesions in the ground truth. In our notation, $\mathbf{x}_i^j$ refers to the i^{th} candidate in the j^{th} bag. All other candidates in the training data set are negative (*i.e.* they have a class label $y_i = 0$). Even if one candidate in a positive bag is declared positive by a classifier and displayed to a radiologist, the underlying lesion would be detected.

Thus, we want the classifier to reject as many negative candidates as possible, but instead of insisting that every positive candidate be labeled as positive by the classifier, we only want the classifier to label *at least one sample in a bag* as a positive (leading to the detection of the lesion). This mirrors the CAD objective.

This can be accomplished by using a *noisy-or* model for assigning a probability that a positive bag is correctly identified by the classifier. In this model, the probability that a positive bag is incorrectly classified is the probability that every sample individual in that bag is incorrectly labeled as a negative. Hence, the j^{th} bag is correctly classified with probability

$$p(y^{\text{bag } j} = 1|\{\mathbf{x}_i^j\}_i, \mathbf{w}) = 1 - \prod_i \left(1 - \sigma(\mathbf{w}^\top \mathbf{x}_i^j)\right). \tag{6}$$

Hence, the log-likelihood function is changed to

$$l(\mathbf{w}) = \sum_{\text{bags } j} \log \left[1 - \prod_i \left(1 - \sigma(\mathbf{w}^\top \mathbf{x}_i^j)\right)\right] + \sum_{\forall i | y_i = 0} \log(1 - \sigma(\mathbf{w}^\top \mathbf{x}_i)). \tag{7}$$

The log-posterior $L(\mathbf{w})$ resulting from the above $l(\mathbf{w})$ may also be optimized for any fixed set of hyper-parameters $\mathbf{A}$.

2.5 Feature selection for optimizing MIL-classifiers

We imposed a prior of the form $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|0, \mathbf{A}^{-1})$, parameterized by d hyper-parameters $\mathbf{A} = diag(\alpha_1 \dots \alpha_d)$. Clearly, as the precision $\alpha_k \to \infty$, *i.e*, the variance for $\mathbf{w}_k$ tends to zero (thus concentrating the prior sharply at zero). Hence, regardless of the evidence of the training data, the posterior for $\mathbf{w}_k$ will also be sharply concentrated on zero, thus that feature will not affect the classification result-hence, it is effectively removed out via feature selection. Therefore, the discrete optimization problem corresponding to feature selection (should each feature be included or not?), can be more easily solved via an easier continuous optimization over hyper-parameters. If one could maximize the marginal likelihood $p(\mathcal{D}|\mathbf{A}) = p(y_1, y_2 \dots, y_N | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N, \mathbf{A})$ this would perform optimal feature selection. We choose the hyper-parameters to maximize the marginal likelihood.

$$\widehat{\mathbf{A}} = \arg\max_{\mathbf{A}} p(\mathcal{D}|\mathbf{A}) = \arg\max_{\mathbf{A}} \int p(\mathcal{D}|\mathbf{w}) p(\mathbf{w}|\mathbf{A}) d\mathbf{w}. \tag{8}$$

Since this integral is not easy to compute for our MIL model we use an approximation to the marginal likelihood via the Taylor series expansion. The marginal likelihood $p(\mathcal{D}|\mathbf{A})$ can be written as $p(\mathcal{D}|\mathbf{A}) = \int e^{\Psi(\mathbf{w})} d\mathbf{w}$, where $\Psi(\mathbf{w}) = l(\mathbf{w}) + \log p(\mathbf{w}|\mathbf{A})$. Approximating Ψ using a second order Taylor series around $\widehat{w}_{\text{MAP}}$, $\Psi(\mathbf{w}) \approx \Psi(\widehat{\mathbf{w}}_{\text{MAP}}) + \frac{1}{2}(\mathbf{w} - \widehat{\mathbf{w}}_{\text{MAP}}) \mathbf{H}(\widehat{\mathbf{w}}_{\text{MAP}}, \mathbf{A})(\mathbf{w} - \widehat{\mathbf{w}}_{\text{MAP}})^\top$, we have the following approximation to the marginal likelihood

$$p(\mathcal{D}|\mathbf{A}) \approx p(\mathcal{D}|\widehat{\mathbf{w}}_{\text{MAP}}) p(\widehat{\mathbf{w}}_{\text{MAP}}|\mathbf{A}) (2\pi)^{d/2} | - \mathbf{H}^{-1}(\widehat{\mathbf{w}}_{\text{MAP}}, \mathbf{A})|^{1/2}. \tag{9}$$

Using the prior $p(\mathbf{w}|\mathbf{A}) = \mathcal{N}(\mathbf{w}|0, \mathbf{A}^{-1})$, the log marginal likelihood can be written as

$$\log p(\mathcal{D}|\mathbf{A}) \approx l(\widehat{\mathbf{w}}_{\text{MAP}}) - \frac{1}{2} \widehat{\mathbf{w}}_{\text{MAP}}^\top \mathbf{A} \widehat{\mathbf{w}}_{\text{MAP}} + \frac{1}{2} \log |\mathbf{A}| - \frac{1}{2} \log | - \mathbf{H}(\widehat{\mathbf{w}}_{\text{MAP}}, \mathbf{A})|. \tag{10}$$

The hyper-parameters $\mathbf{A}$ are found by maximizing this approximation to the log marginal likelihood. There is no closed-form solution for this. Hence we use a iterative re-estimation method by setting the first derivative to zero. The derivative can be written as

$$\frac{\partial \log p(\mathcal{D}|\mathbf{A})}{\partial \alpha_i} = -\frac{1}{2}\widehat{\mathbf{w}}_i^2 + \frac{1}{2\alpha_i} - \frac{1}{2}\Sigma_{ii}, \tag{11}$$

where Σ_{ii} is the i^{th} diagonal element of $\mathbf{H}^{-1}(\widehat{\mathbf{w}}_{\mathrm{MAP}}, \mathbf{A})$. Assuming Σ_{ii} does not depend on $\mathbf{A}$ a simple update rule for the hyper-parameters can be written by equating the first derivative to zero.

$$\alpha_i^{\mathrm{new}} = \frac{1}{\mathbf{w}_i^2 + \Sigma_{ii}}. \tag{12}$$

The final algorithm has two levels of iterations: in an outer loop we update the hyper-parameters α_i and in an inner loop we find the MAP estimator $\widehat{\mathbf{w}}_{\mathrm{MAP}}$ given the hyper-parameters. After a few iterations we find that the hyper-parameters for several features tend to infinity causing numerical problems in implementation. This means that we can simply remove those irrelevant features from further consideration in future iterations.

3 Results

3.1 Run-time efficiency

We converge upon the optimal feature subset within about 10 iterations of the outer loop. Using a simple Newton-Raphson optimizer, for a fixed the inner loop finds the MAP estimator in about 5-8 iterations. On a 1 GHz laptop, the entire algorithm including automatic feature selection converges in under a minute even on training data sets with over 10,000 patients. The system needs absolutely no human intervention or tuning even to decide on the number of features to be used in the eventual classifier.

3.2 Accuracy

We compared the proposed algorithm against a state-of-the-art linear SVM classifier and against the proposed feature selection approach to non-MIL baseline logistic-regression. Each system was trained using a small proprietary digital mammography (FFDM) data set with 144 biopsy proven malignant-mass cases and 2005 normal cases from BI-RADS® 1 and 2 categories. The CG and feature extraction algorithms produced 127,509 candidates in total, each described by a set of 81 numerical features. The systems were evaluated on a held out set of 108 biopsy proven malignant cases and 1513 BI-RADS® 1 and 2 cases. The FROC curves on the held out set are produced below in Fig. 2. The proposed MIL algorithm automatically performed feature selection and only selected 40 features out of the original set of 81 features, while the non-MIL variant selected

56 features. No human intervention or cross-validation tuning was necessary for our algorithm. The regularization parameter of the SVM was tuned using 10-fold cross validation on the training set, and the optimal parameter was used to train a single SVM on the entire data.

4 Discussion

As seen in Fig. 2, the proposed algorithm was indeed more accurate when measured in terms of per-patient (and although not shown, per-image) FROCs. As expected, this statistically significant improvement on per-patient statistics comes at the cost of deteriorating the per-candidate statistics. This underscores the point of the paper: in CAD we care about a different set of performance metrics that are not optimized in conventional methods, and our algorithm optimizes them. Note that the code can optimize either per-lesion or per-patient or per-image statistics: one simply describes the unique bag-ID of a candidate in terms of the ID of the lesion, patient or image using the same code to accomplish this. The proposed method is fast, easy to implement, very general and broad in terms of CAD applicability, and it can support many different baseline learning algorithms such as (potentially non-linear and/or multi-class) Neural Nets, kernel classifiers, etc. We have already implemented such variations, but did not report them here due to space considerations.

References

1. R. Campanini, A. Bazzani, A. Bevilacqua, D. Bollini, D. Dongiovanni, E. Iampieri, N. Lanconelli, A. Riccardi, M. Roffilli, and R. Tazzoli. A novel approach to mass detection in digital mammography based on SupportVector Machines (SVM). In *Proceedings of the 6th International workshop in digital Mammography (IWDM)*, pages 399–401, Bremen, Germany, 2002. Springer Verlag.
2. G. Fung, M. Dundar, B. Krishnapuram, and R. B. Rao. Multiple instance learning for computer aided diagnosis. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 425–432. MIT Press, Cambridge, MA, 2007.
3. B. Sahiner, H.P. Chan, N. Petrick, D. Wei, MA Helvie, DD Adler, and MM Goodsitt. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. *IEEE Transactions on Medical Imaging*, 15(5):598–610, 1996.

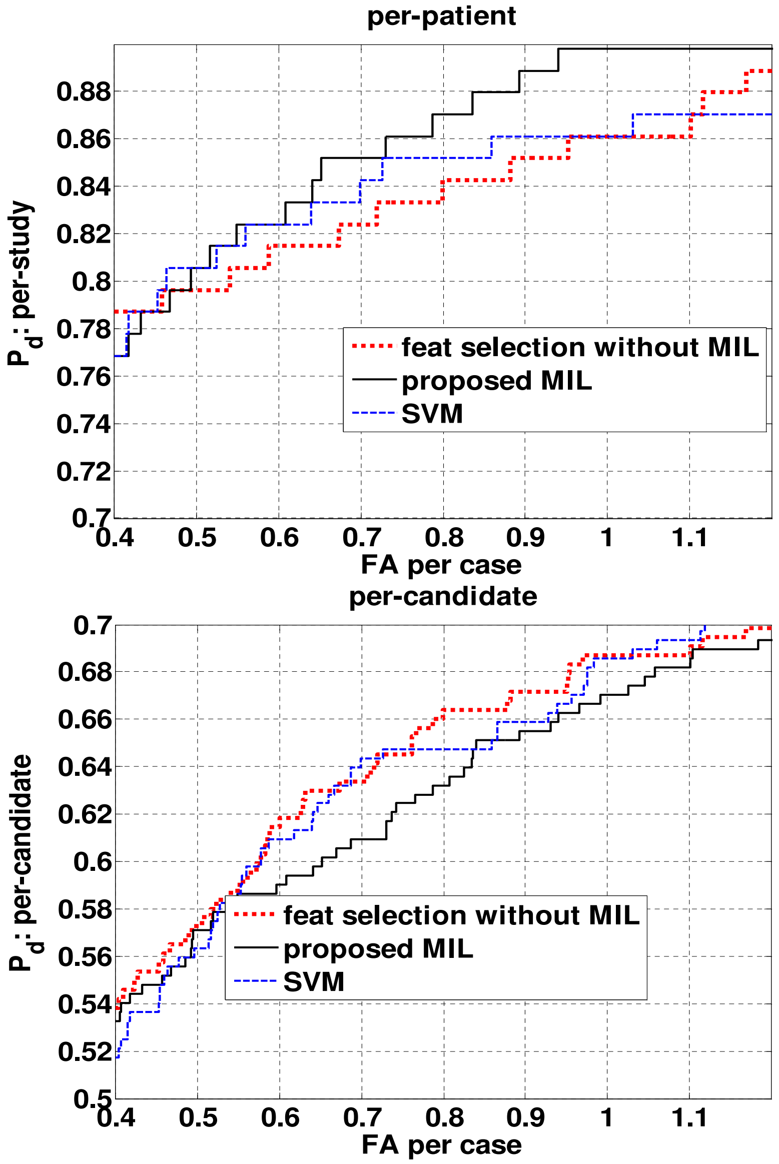

Fig. 2. The proposed MIL classifier (solid black) with automatic feature selection improves per-patient statistics at the cost of deteriorating per-candidate statistics. Competing methods: linear SVM (blue dashed) and the proposed feature selection approach without MIL (dotted red).