Addressing image variability while learning classifiers for detecting clusters of micro-calcifications

Glenn Fung, Balaji Krishnapuram, Sriram Krishnan, and Bharat Rao

Siemens Medical Solutions, 51 Valley Stream Pkwy, Malvern, PA-19355, USA

Abstract. In the context of computer aided mammography, many standard algorithms (e.g. SVM and neural networks) have been used for detecting lesions. However, these general purpose learning methods make implicit assumptions like sample independence that are commonly violated. A new ensemble algorithm is proposed to explicitly account for the small fraction of outlier images which tend to produce a large number of false positives. A bootstrapping procedure is used to ensure that the candidates from these outlier images do not skew the statistical properties of the training samples. We compared a standard state-of-the-art method (SVM) for detecting clusters of micro-calcifications, with our ensemble algorithm. This algorithm significantly improved the test set results, especially in the operating region of interest (around 0.2 FP per image).

1 Introduction

In *computer aided diagnosis* (CAD) applications the goal is to detect structures of interest to physicians in medical images: *e.g.* to identify potentially malignant lesions in mammography. In an almost universal paradigm, this problem is addressed by a 3 stage system: identification of potentially unhealthy candidate *regions of interest* (ROI) from a medical image, computation of descriptive features for each candidate, and classification of each candidate (*e.g.* normal or diseased) based on its features.

This paper focusses on automatic algorithms for designing (*i.e.* learning) pattern classifiers for the third stage. Automatic learning algorithms are an important part of the modern methodology for efficiently designing computer aided diagnostic products. Besides improving the diagnostic accuracy, these technologies greatly reduce the time required to develop algorithms that act as "second readers".

In the context of computer aided mammography, many standard algorithms (*e.g.* SVM, Back-propagation for Neural Nets, Kernel Fisher Discriminants) have been used to learning classifiers for detecting malignant lesions in computer aided mammography [1–3]. However, these general purpose learning methods make implicit assumptions that are commonly violated in CAD applications, often resulting in sub-optimal prediction accuracy for the classifiers that they learn. For example, these methods almost universally assume that the training samples are *independently* drawn from an *identical*—albeit unobservable—underlying distribution (i.i.d. assumption).

We propose a new ensemble algorithm that is designed to improve the classification accuracy. This algorithm explicitly accounts for the fact that a small fraction of outlier images tend to produce a large number of false (true) positives in the training set used

to learn classifiers, whereas a large number of other images only contribute very few negative (positive) training samples each. A bootstrapping procedure is used to ensure that the candidates from these outlier images do not skew the statistical properties of the training samples.

When we learnt a classifier using standard state-of-the-art methods (SVM) for detecting clusters of micro-calcs, the resulting system performed (generalized) poorly on a hold out set of test samples, in terms of per-image sensitivity & per-patient sensitivity. By contrast, the proposed methods significantly improved the ROC curves, especially in the operating region of interest (around 0.2 FP per image).

The rest of the paper is organized as follows. Section 2 highlights some of the assumptions that underly almost all algorithms for learning pattern classifiers, and indicates why some of them may be inappropriate for CAD. Based on this analysis, Section 3 develops a novel method for learning classifiers that detect clusters of microcalcifications. Experimental results are provided in Section 4. We conclude with a discussion of the broader applicability of the proposed algorithm and some ideas for future extensions in Section 5.

2 Common assumptions while learning pattern classifiers

2.1 Creation of the training data

During the design of a CAD system, considerable human intervention and domain knowledge engineering is employed in the first two stages of a CAD system for (a) candidate generation (CG): identifying all potentially suspicious regions in a candidate generation stage with very high sensitivity, and (b) feature-extraction: description of each such region quantitatively using a set of medically relevant features. For example quantitative measurements based on texture, shape, intensity and contrast and other such characteristics may be used to characterize any region of interest (ROI). Subsequently, for learning the classifier to be used in the third stage, a training dataset is created by obtaining features to describe each candidate ROI in the training images, and class labels are assigned to them based upon the overlap and/or distance from any radiologist-marked (diseased) region.

2.2 Characteristic properties of the data

A few important characteristics of the data are relevant for designing classifiers that generalize well. First, there is a form of stochastic dependence between the labeling errors of a group of candidates, all of which are spatially proximate to the same radiologist mark. Further, the features used to describe spatially adjacent or overlapping samples are also highly correlated. As a result, both the labels and the features for the training samples from an image tend to be highly correlated: the inter sample correlation is particularly high for spatially adjacent candidates.

Second, some types of biological or image structures tend to be identified much more often by CG algorithms in the form of many spatially adjacent candidates. This introduces a sampling bias in the training dataset as compared to the frequency of occurrence of these structures in screening populations. Also, some training images tend

to contain far more false positive candidates as compared to the rest of the training database, due to noise or various imaging artifacts present in them.

2.3 Shortcomings in standard classification algorithms

In the CAD literature, many machine learning algorithms—such as *neural networks*, *support vector machines* (SVM), and *Fisher's linear discriminant*—have been employed to train classifiers. However, almost all the standard methods for classifier design explicitly make certain assumptions that are violated by the somewhat special characteristics of the data as discussed above.

In particular, most of the algorithms assume that the training samples or instances are drawn *identically* and *independently* from an underlying (unknown) distribution. However, as mentioned above, due to spatial adjacency of the regions identified by a candidate generator, both the features and the class labels of several adjacent training candidates are highly correlated.

Further, the standard methods for classifier design implicitly assume that the appropriate measure for evaluating the classifier is based only on the accuracy of the system on a per-lesion basis. In other words, these algorithms try to most correctly classify each candidate from the CG algorithm; they do not account for the sampling bias introduced by the common tendency of CG algorithms to produce candidates corresponding to certain types of structures and fewer candidates corresponding to others.

The appropriate measure of accuracy for evaluating a CAD system is different from the standard measures that are optimized by conventional classifiers. In particular, even if one of the candidates that refers to the underlying malignant structure is correctly highlighted to the radiologist, the *lesion* is detected. Thus, correct classification of every candidate instance is not as important as the ability to detect *at least one* candidate that points to a malignant lesion. At another level, in many CAD problems it is even more relevant to measure the accuracy in terms of FROC curves plotting the per-patient sensitivity—the fraction of diseased patients correctly identified by the system—versus the rate of false positives per patient.

These consideration motivated the development of a novel algorithm for learning ensemble classifiers in an effort to adjust for the sampling bias of the CG algorithm and the correlations between subsets of samples for the same image or patient.

3 Learning ensemble classifiers for CAD using boot-strapping

Instead of learning a single classifier, we learn a set or ensemble of k classifiers. The final prediction of the ensemble is obtained by weighted voting, this is, the final prediction consists in a weighted sum (average) of the predictions of the members of the ensemble. Furthermore, in order to achieve a diverse ensemble, we use the technique known as bagging [], where each classifier is trained on a random redistribution of the training set. In our case, each classifier's training set is generated by randomly drawing, without replacement, N^+ positive examples and N^- negative examples from the original training set.

It is a well-known fact in the machine learning community that very unbalanced training sets (number of negatives points or false positive candidates is much larger that the number of positive points or number of real microcalcifications) tend to make most machine learning algorithms to be biased toward the majority class, producing poor generalization over the minority class. In order to address this issue and to reduce computational complexity, for each of the classifiers in the ensemble we chose N^- to be a relatively smaller number ($N^+ = 1000$ was chosen by tuning in our experiments). The number of positives N^+ was chosen as a function of the number of positive images in the training set. For each positive image in the training set only i positive datapoints were randomly chosen from all the positive candidates in the image.

Each one of the linear classifiers in the ensemble is obtained by solving the Relevance Vector Machine formulation (RVM) [].RVM produces a linear classifier that makes its predictions using only a small number of relevant features which are automatically selected from the original large pool of features. Enforcing each member of the ensemble to depend on an small number of features also promotes diversity of the ensemble since each classifier tends to make predictions based on different features.

Next, we present our proposed algorithm to learn an ensemble classifier for detecting clusters of micro-calcifications from digital mammograms:

Algorithm 1 ***BuildEnsemble*** **return:** $W = \left[w^1, \ldots, w^{nc}\right]$**:**

0. Given
   - *the number nc that define the number of classifiers in the ensemble.*
   - *The* training set *comprised of a matrix $A \in R^{m \times n}$ (m is the number of points and n is the number of input features and the vector $l \in \{1, -1\}^m$ containing the labels.*
   - *The number of positive points N^+ and negative points N^- to be randomly selected to train each one of the nc classifiers members of the ensemble.*

1. initialize $k = 0$
2. If $k = nc$, stop, return the matrix $W \in R^{n \times nc}$ hyperplane coefficients $W = \left[w^1, \ldots, w^{nc}\right]$
3. otherwise, generate training set for classifier k by randomly drawing, without replacement, N^+ positive examples and N^- negative examples from the original training set.
4. Obtain the coefficients w^k for classifier k by solving the RVM formulation []
5. do $k = k + 1$

Given an unseen datapoint (column vector) $x \in R^n$, the final ensemble classifier prediction is given by:

$$pred(x) = \frac{\sum_{k=1}^{nc} \exp(x^T w^k)}{nc}$$

4 Experiments

Our numerical experiments were performed in a dataset consisting of 37098 microcalcification clusters candidates extracted from 1891 digitized film-screen mammography (FSM) images belonging to 621 cases (242 Malignant and 379 normals). Each candidate consists in a vector of 1051 descriptors or features that were extracted from the microcalcification clusters candidates based on shape, texture, density , etc. The images of all the cases were digitized at high resolution (600 dpi, 12 bit) by a prototype CAD device developed by Siemens CAD, Israel. In order to validate the generalization performance of the proposed system, the available 621 cases were randomly divided into two subsets:

- A *training set* comprised of 945 images from 311 cases (190 normals and 121 malignants). 744 of the The 945 images belong to the normal cases (normal images) and the remaining 201 images belong to the malign cases. The total number of candidates in the training set is 18459, only 443 of these candidates are real microcalcification clusters, the remaining 18016 are false positives, i.e. candidates pointing to structures in the breast that are not microcalcification clusters.
- A *testing* or *validation set* comprised of 946 images from 310 cases (189 normals and 121 malignants). 754 of the The 946 images belong to the normal cases (normal images) and the remaining 192 images belong to the malign cases. The total number of candidates in the training set is 18639, 462 of these candidates are real microcalcification clusters, the remaining 18177 are false positives.

The number of positive datapoints i indicates the number of positive candidates to be randomly chosen from each positive image. The final number of positive candidates then, depends explicity on the choice of i and the number of positive images. Since our training set contains 201 malignant images, when $i = 2$, this results in randomly chosing up to 402 positive candidates to be included in the training set, this is, up to 2 for each malignant image (some images may not have 2 positive clusters, in that case only one was picked). We tried different values of i and $i = 1$ gave the best results in our problem. The idea behind this positive datapoints sampling scheme is to drive the ensemble classifier performance to be optimized per image instead of per cluster. In other words, by sampling positive clusters uniformly across all the positive images the classifier gets to learn a more heterogeneous concept of positive or malignancy cluster. By using all the positives candidates , the classifier may get biased by some of the rare images with an unusual number of positive candidates (see Figure 1) and that are not representative of the general population of positive images.

The number of classifiers in the ensemble k was fixed to 101 based on empirical experience.

4.1 Comparison to an standard SVM formulation

In order to show the effectiveness of our approach we our numerical experiments included comparisons to the smooth support vector machine (SSVM) [4]. SSVM is an efficient SVM formulation that consists in using an smoothing version of the plus function to reformulate the SVM problem as an unconstrained optimization problem that can be solved very fast and that can handle large datasets.

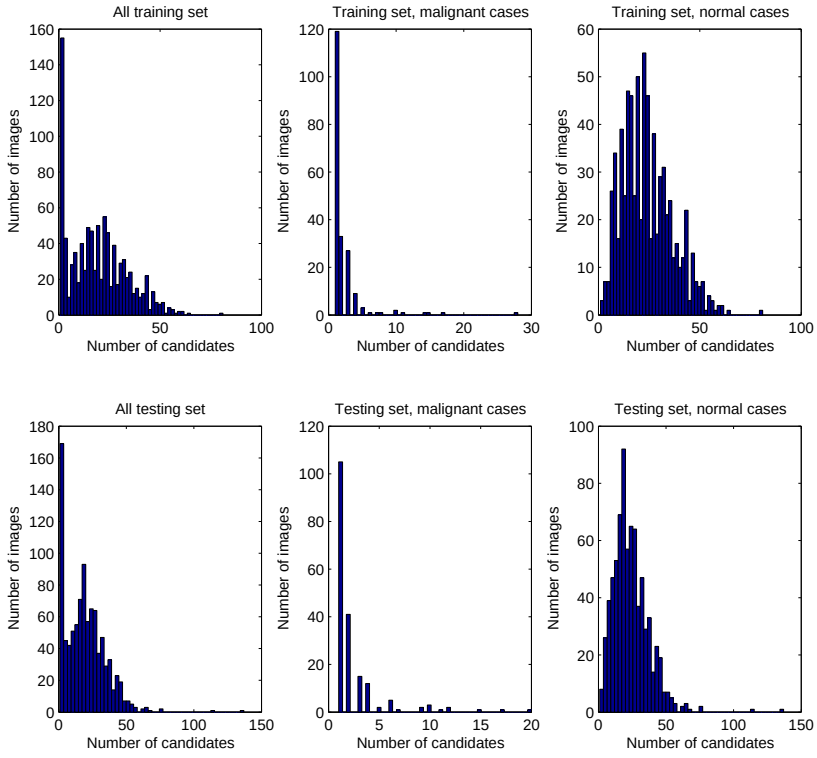

Fig. 1. Histogram on the training and testing sets showing number of images (y axis) with k candidates (x-axis). Histograms for all the candidates, malignant candidates only and for normal candidates only are shown. Note that some of the outlier images in both the training and the testing set have an usnussual large number of positive and negative candidates.

In order to be as fair as possible the parameter ν for the SSVM algorithm was determined by cross-validation from the set $\left\{2^{-}10, 2^{-}9, \ldots, 2^{1}0\right\}$.

Figures 2,3 and 4 show that our proposed method is considerable more robust and generalize better on the unseen cases (testing set) at all levels (per cluster, per image and per patient respectively). As can be seen in igures 2,3 and 4, at the 0.15 FP/image level our ensemble method obtained:

- 66.5% testing set sensitivity at the cluster level compared to 62.3% testing set sensitivity obtained by the SSVM algorithm.
- 88.8% testing set sensitivity at the image level compared to 79.5% testing set sensitivity obtained by the SSVM algorithm.
- 100.0% testing set sensitivity at the patient level compared to 95.0% testing set sensitivity obtained by the SSVM algorithm.

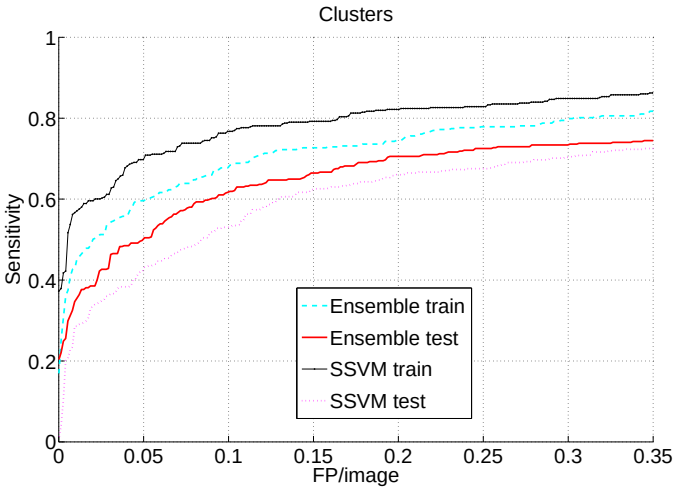

Fig. 2. Comparisons of SSVM and the proposed ensemble. The ROC curves illustrate the performance on the training and testing sets at the cluster level

5 Related Work & Discussions

Most classification approaches used till today in the domain of computer aided diagnosis have assumed the data to fulfill some general assumptions like sample independence, and an identical distribution for all patients. In many cases those algorithms have been used as so called black boxes. This despite the fact that these assumptions are violated due to many different factors e.g. samples of the same patient are correlated and might even come from a different distribution e.g. in the case of a very dense breast However, the results of this article show that taking explicitly the distribution of the data into account can improve the classification results. This improvement results in a real improvement in sensitivity and a decrease in false positives per case, leading to a real clinical benefit. Note especially that this improvement is significantly observed for the independent test set.

Since more explicitly modeling the data distribution seems to lead to improved results one might therefore want to consider an even more detailed approach to modeling the data in future work. The current model does not take spatial correlation of samples which are spatially close to each other in the image into account. Also samples in different images of the same patient might of course be correlated. These correlations could be modeled using random effects models AND WHAT OTHER METHODS.

References

1. Fu, J., Lee, S., Wong, S., Yeh, J., Wang, A., Wu, H.: Image segmentation, feature selection and pattern classification for mammographic microcalcifications. Comput Med Imaging Graph **9** (2005) 419–429

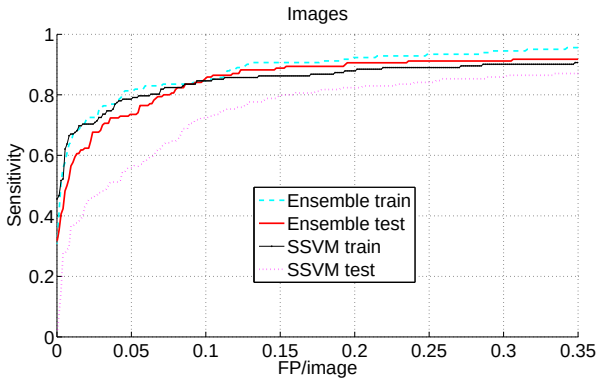

Fig. 3. Comparisons of SSVM and the proposed ensemble. The ROC curves illustrate the performance on the training and testing sets at the image level

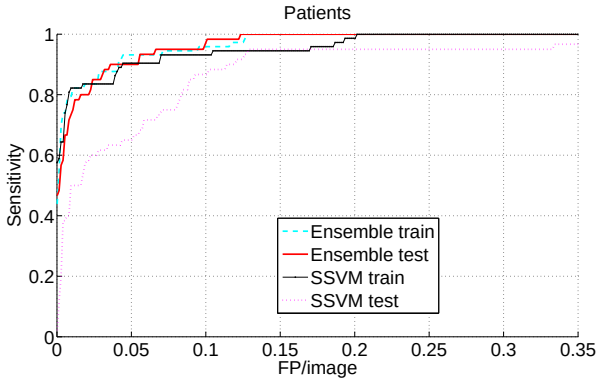

Fig. 4. Comparisons of SSVM and the proposed ensemble. The ROC curves illustrate the performance on the training and testing sets at the patient level

2. Pappadopolous, A., Fotiadis, D., Likas, A.: Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines. Artificial Intelligence in Medicine **34** (2005) 141–150
3. Wei, L., Yang, Y., Nishikawa, R., Jiang, Y.: A study of several machine learning methods for classification of malignant and benign microcalcifications. IEEE Transactions on Medical Imaging **24** (2005) 371–380
4. Lee, Y.J., Mangasarian, O.L.: SSVM: A smooth support vector machine. Computational Optimization and Applications **20** (2001) 5–22 Data Mining Institute, University of Wisconsin, Technical Report 99-03. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/99-03.ps.