Regenerative Random Forest with Automatic Feature Selection to Detect Mitosis in Histopathological Breast Cancer Images

Angshuman Paul[1], Anisha Dey[2], Dipti Prasad Mukherjee[1],
Jayanthi Sivaswamy[2], and Vijaya Tourani[3]

[1] Indian Statistical Institute, Kolkata, India,
[2] Center for Visual Information Technology, International Institute of Information Technology, Hyderabad, India,
[3] CARE Hospital, Hyderabad, India

Abstract. We propose a fast and accurate method for counting the mitotic figures from histopathological slides using regenerative random forest. Our method performs automatic feature selection in an integrated manner with classification. The proposed random forest assigns a weight to each feature (dimension) of the feature vector in a novel manner based on the importance of the feature (dimension). The forest also assigns a misclassification-based penalty term to each tree in the forest. The trees are then regenerated to make a new population of trees (new forest) and only the more important features survive in the new forest. The feature vector is constructed from domain knowledge using the intensity features of nucleus, features of nuclear membrane and features of the possible stroma region surrounding the cell. The use of domain knowledge improves the classification performance. Experiments show at least 4% improvement in F-measure with an improvement in time complexity on the MITOS dataset from ICPR 2012 grand challenge.

Keywords: Breast cancer grading, random forest, weighted voting, tree penalty, forest regeneration, feature selection

1 Introduction

Detecting and counting of mitotic cells from histopathological images are key steps in breast cancer diagnosis as mitosis signals cell division. Interest in automating this process is driven from the arduous and error-prone nature of the manual detection. Research in this area has received much interest after the launch of mitosis detection as a challenge in ICPR 2012. The images that need to be analyzed are obtained with Hematoxylin & Eosin ($H\&E$) staining which render the cell nuclei dark blue against pinkish background (see Fig. 1).

Early solutions proposed for this problem include Gamma-Gaussian mixture modeling [7] and independent component analysis [6]. The difficulty in identifying appropriate features has been addressed by augmenting handcrafted features with those learnt by a convolutional neural network [9] or using only features

N. Navab et al. (Eds.): MICCAI 2015, Part II, LNCS 9350, pp. 94–102, 2015.
DOI: 10.1007/978-3-319-24571-3_12

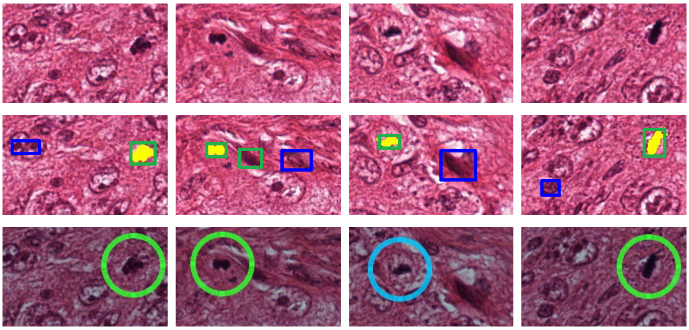


Fig. 1. Sample results: Row 1: Original sub-images , Row 2: Mitosis detection with proposed method (green/blue boxes indicate true/false positive) and Row 3: Mitosis detection with IDSIA [4] (green/cyan circles indicate true positive/false negative). Ground truth provided by [1] are shown as yellow regions in the middle row.

learnt with a deep neural network [4]. But this benefit of neural networks comes at a cost of tuning effort and training time for optimal performance. For instance, [4] reports a training time of 1 day on a GPU and processing time of 8 minutes per image. The latter poses a major deterrent for considering automated mitosis detection with high throughput processing of tissue microarrays. More extensive review can be found in [11]. Recently, random forests with population update [10] have been tried for mitosis detection where tree weights are changed based on classification performance. An in-depth biological study of the breast cancer tissues reveal that key factors in mitosis detection are: color of the nucleus, shape of the nuclear membrane and texture of the surrounding region. Presence of nucleus in a stromal region rules out the possibility of it being mitotic, while the absence or rupture of the nuclear membrane signals mitosis. Hence, we propose a fast mitosis detection method with the following contributions (i) new features based on domain knowledge mentioned above and (ii) regenerative random forest-based classification *with* automatic feature selection, unlike [10]. Automatic feature selection is achieved using a novel feature weighting scheme. Feature weights are based on the importance of a feature and we reject features with low weights. A new generation of forest (new population of trees) is created which operates on a reduced feature set. During the test phase, each tree of the trained forest votes with its corresponding weights to perform the classification.

The rest of the paper is organized as follows: section 2 describes our proposed method and proves that the process of forest regeneration converges with maximum classification accuracy with the training data. We present the experimental results in section 3 and the paper concludes in section 4.

2 Methods

We focus on several important biological cues for mitosis detection. First, after chromosomal condensation in interphase, the nucleolus disappears and the nuclear envelope breaks down during prophase. Consequently, most of the liquid contents of the nucleus is released into the cytoplasm making the nucleus denser with darker appearance compared to non-mitotic nucleus [2] after $H\&E$ staining. This intensity pattern of mitotic nuclei remains almost phase invariant during mitosis. Thus, the nuclear intensity pattern provides discriminative features between mitotic and non-mitotic cells. Second, it has been observed that the stroma or the connecting tissues contain non-cancerous fibroadenomas and a large number of non-mitotic cells. Thus, if stroma is found surrounding an unknown cell, the cell can be classified as non-mitotic. Hence, we evaluate the texture features of the regions surrounding the cell to find whether the cell belong to stroma or not. Further, after the nuclear membrane breaks down during mitosis, the nuclear contour of a mitotic cell becomes irregular compared to the contour of a non-mitotic cell with nuclear membrane. Thus, the pattern of the contour of nucleus also provides useful information in categorizing the mitotic and the non-mitotic cells. But the above features can be correctly evaluated and used for classification only if the nuclei are accurately segmented. So, the proposed method consists of three steps, namely: nuclei segmentation, feature extraction, classification. We segment the nuclei using [10]. These nuclei are used for feature extraction. In the next few paragraphs we discuss the feature extraction and classification steps in detail.

2.1 Feature Extraction

The blue channels of the $H\&E$ stained images lack useful information since both mitotic and non-mitotic nuclei are almost uniformly stained in blue [8]. Hence, we use only the red and green channel images for feature extraction. First, we extract the red and green channel histograms of the segmented nuclei. Next we look for the features of stroma by taking a region with 3 times the area enclosing the nucleus. In each such area, we evaluate the Haralick texture feature [5] of the region excluding pixels representing nuclei. We further detect the features associated with nuclear membrane. Here, we rely on the fact that a detected object (nucleus) with nuclear membrane has smooth and almost circular boundary whereas the mitotic nucleus with ruptured nuclear membrane generally has a rough boundary with non-circular shape. In order to characterize the detected object contour (whether the contour is circular or not) we evaluate the solidity, extent and lengths of major and minor axes of an ellipse equivalent to the detected object. This yields a 604-dimensional feature vector for each of the detected object (nucleus). Let the feature vector of k^{th} nucleus be denoted as $\mathcal{F}(k)$. Intensity histogram of nucleus, stromal texture and nuclear membrane contribute 512, 88 and 4 features respectively. We use these feature vectors in the next step for classification of the nuclei in mitotic and non-mitotic category.

2.2 Classification

Our training dataset relies on manual ground truth [1] which inherently includes observer bias. Further a wide variation of feature magnitudes among the nuclei of same category (mitotic or non-mitotic) dictates the use of an ensemble classifier rather than a single classifier. Further, the variability in class-specific signatures in this problem precludes an explicit feature initialization in the classifier. All these prompt us to use random forest classifier [3] which is an ensemble classifier without the requirement for explicit feature initialization. A random forest is constructed of T trees each of which is grown using a bootstrap sample [3] from the training dataset. While splitting a node in each such tree, a subset of f number of features from $\mathcal{F}(k)$ are chosen. The best out of these f features is used for splitting the node. But, all of the features (dimensions) of $\mathcal{F}(k)$ may not have class-discriminative information. So, we propose a novel regenerative random forest that keeps on improving its classification performance by eliminating less discriminative features (dimensions) as the new populations of trees (new forest) are produced and thus performs feature selection and classification in an integrated manner.

Regenerative Random Forest: Our present aim is to classify the detected nuclei into two classes: mitotic and non-mitotic. For this, let $\mathcal{S}$ be the training dataset with s number of training data. A feature vector $\mathcal{F}(k)$ of dimension 604 is associated with k^{th} such data. During first step of training, we build T number of binary trees. For each tree τ, we construct a subset of training feature vectors Ψ^{τ} (with $\#\Psi^{\tau} = s$) by selecting s number of training feature vectors randomly with replacement from $\mathcal{S}$. At each node of tree τ, we randomly choose a set (Λ) of f number of features (dimensions) from each training feature vector $F(k) \in \Psi^{\tau}$ without replacement for splitting the node into its two child nodes. Now, for selecting the split point, we consider the fact that an ideal split should result in child nodes each of which contains training data of exactly one class. Let a node u be composed of ξ_f^u number of training feature vectors and the class label of j^{th} feature vector be denoted by c_j. We indicate the probability of class c_j in node u by $p(u, c_j)$. Then the total entropy of node u is defined as: $H(u) = \sum_{c_j} p(u, c_j) \log \frac{1}{p(u,c_j)}; \forall j \in \ \xi_f^u$. So, the total entropy of two child nodes after a split from parent node v using feature (dimension) f_l is:

$$E(v, f_l) = \sum_{u=1}^{2} H(u) = \sum_{u=1}^{2} \sum_{c_j} p(u, c_j) \log \frac{1}{p(u, c_j)}; \forall j \in \ \xi_f^u. \tag{1}$$

In case of an ideal split from parent node v, $p(u, c_j) = 1$ making $H(u) = 0; \forall u$, where u is a child node of v. Thus, from (1), we get $E(v) = \sum_{u=1}^{2} H(u) = 0$ which indicates that the best split should result in minimum total entropy of the child nodes. Hence we look for the split point in node v based on a feature (dimension) $f_\chi \in f$ that minimizes the total entropy of its child nodes. So,

$$f_\chi = \underbrace{argmin}_{f_l} \{E(v, f_l)\} = \underbrace{argmin}_{f_l} \left\{ \sum_{u=1}^{2} \sum_{\forall j \in \ \xi_f^u} p(u, c_j) \log \frac{1}{p(u, c_j)} \right\}. \tag{2}$$

We continue splitting a node u until u has a very small value (δ) of entropy. So, we split a node u only if $H(u) > \delta$. Now, suppose the ground truth class label of j^{th} training feature vector is given by c_j and tree τ predicts the proper class label with probability $q^0_\tau(j, c_j)$ [3]. Then, probability of misclassification on j^{th} training feature vector by tree τ is $1 - q^0_\tau(j, c_j)$. Once all the trees are grown, we calculate weight (importance) of each tree in a novel manner based on tree's classification performance on the labeled training data. Initially, each tree τ is assigned a weight $w^0_\tau = 1$. Now, we propose a penalty for each tree on the basis of above misclassification probabilities. The normalized penalty of tree τ is given by: $\Upsilon^0_\tau = \frac{1}{s} \sum_{\forall j \in \Psi^\tau} (1 - q^0_\tau(j, c_j))$. Consequently, the weight of a tree after the first population (first phase of training) is given by: $w_\tau = w^0_\tau - \kappa \Upsilon^0_\tau = w^0_\tau - \frac{\kappa}{s} \sum_{\forall j \in \Psi^\tau} (1 - q^0_\tau(j, c_j))$, where κ is a constant in $(0, 1]$. Subsequently, we assign weight to each feature of the feature vector $\mathcal{F}$. Consider a tree τ where feature x has been selected $\alpha_\tau(x)$ times in Λ, out of which it has been used $\beta_\tau(x)$ times for node splitting. Then we define the importance of feature x in tree τ as $\gamma_\tau(x) = \frac{\beta_\tau(x)}{\alpha_\tau(x)}$ and the global importance (weight) of feature x is defined as: $\eta(x) = \frac{1}{T} \sum_{\tau=1}^{T} \hat{w}_\tau \gamma_\tau(x)$, where $\hat{w}_\tau$ is the value of w_τ, normalized w.r.t. $max(w_\tau), \forall \tau \in T$. Note that $\alpha_\tau(x) = 0 \Rightarrow \beta_\tau(x) = 0$. So, if $\alpha_\tau(x) = 0$, we take $\gamma_\tau(x) = 1$. Clearly, the feature (dimension) that provides more class-discriminative information will be used for splitting more number of times as evident from (1) and (2) yielding high value of $\gamma_\tau(x)$ in each tree. Thus a more class-discriminative feature (dimension) x will have higher value of weight $\eta(x)$. Let μ and σ be the mean and standard deviation of feature weights $\eta(x), \forall x$. Also, let $\mathcal{L}$ be the set of features that has weight $\eta(x) < (\mu - \sigma)$. Clearly, the features in $\mathcal{L}$ are the redundant features with low weights. So, we remove the features in $\mathcal{L}$ and make $\mathcal{F} = \mathcal{F} - \mathcal{L}$. Then we create a new population of trees (new forest) which are trained with a subset of training data Ψ^τ. Ψ^τ is composed of s number of training feature vectors randomly selected with replacement from $\mathcal{S}$. Thus, the weight of tree τ in $(n+1)^{th}$ population is given by: $w^{n+1}_\tau = w^n_\tau - \kappa \Upsilon^n_\tau = w^n_\tau - \frac{\kappa}{s} \sum_{\forall j \in \Psi^\tau} (1 - q^n_\tau(j, c_j))$. Let $\mathcal{F}^n$ be the set of features in n^{th} population and $\mathcal{L}^n$ be the set of features to be removed from $\mathcal{F}^n$. Then the reduced feature set in $(n+1)^{th}$ population is given by: $\mathcal{F}^{n+1} = \mathcal{F}^n - \mathcal{L}^n$. Thus, our approach performs automatic selection of features that provide more class-discriminative information. Next, we show that the proposed approach results in maximization of classification performance of the forest on training data.

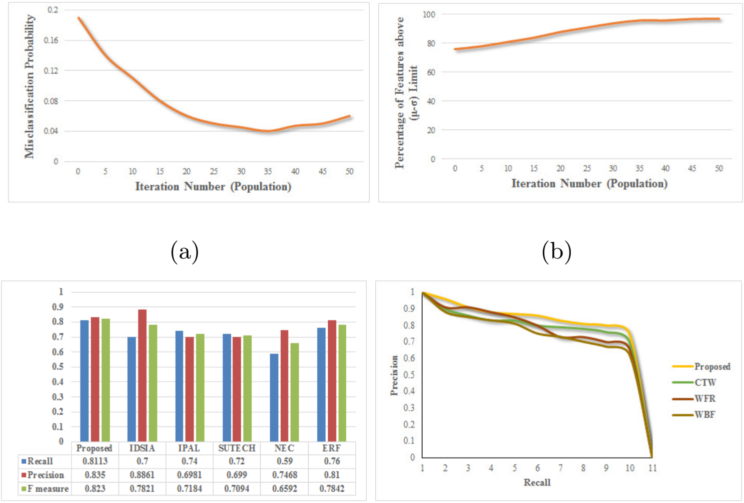
Maximization of Forest Classification Performance: Let the relative weight of r^{th} feature (dimension) be $\hat{\eta}(r) = \eta(r)/max(\eta(r))$. We consider an empirical dependence that the accurate classification probability by the forest Φ in the n^{th} population (P^n_Φ) depends on the relative weight of true discriminative features (dimensions). The relative weight $\hat{\eta}(r)$ indicates the probability of r^{th} feature (dimension) to be selected as the feature for splitting a node, given that the feature is there in Λ of the corresponding node. Consequently, P^n_Φ also depends on the probability that the above mentioned true discriminative feature is chosen in Λ. Let there be d number of true discriminative features in $\mathcal{F}$. Also, let the subset of features to be considered for node splitting in tree τ in the n^{th} population be Λ_n^τ. Given this, we obtain the probability that r^{th} true discriminative feature is selected in Λ_n^τ ($\#\Lambda_n^\tau = f$) is given by $\rho_n^\tau(r) = \binom{\#\mathcal{F}-1}{f-1}/\binom{\#\mathcal{F}}{f}$. Now, if the weight of the r^{th} discriminative feature is $\eta(r)$, then, we consider the empirical function Π that relates P_Φ^n and $\rho_n^\tau(r), \eta(r)$ as: $P_\Phi^n = \Pi\left(\sum_{r=1}^{d} \rho_n^\tau(r)\hat{\eta}(r)\right)$. So, if we go on removing redundant features using the equation $\mathcal{F}^{n+1} = \mathcal{F}^n - \mathcal{L}^n$, $\#\mathcal{F}$ will decrease. Thus, from straightforward observation, for every infinitesimal change in iteration (population number) $\frac{\Delta\rho_n^\tau(r)}{\Delta n} > 0$. Now, if we take sufficient training data, the relative weight of the features can be considered to be statistically independent of the subset of data chosen for training each tree. So, using the expression of P_Φ, we can write $\frac{\Delta P_\Phi^n}{\Delta n} > 0$ until we remove any important feature.

Forest Convergence: Increase in ΔP_Φ^n corresponds to decrease in misclassification probability of the forest on training data [3]. So, we conclude that our proposed approach converges with minimum misclassification probability. We terminate the population update when misclassification probability starts increasing after attaining a minima. This situation corresponds to removal of discriminative feature from the feature vector. Finally we discard the current population and select the trees of previous population (that attained misclassification minima). We compute the weights of the selected trees and use a weighted voting from these tress to classify test data. At this point, we discuss two situations that may arise during training. First, at some iteration number (population) n^*, the feature vector may have all the features with weight $\eta(x) \geq (\mu - \sigma)$, indicating that all the features now present in feature vector are important (class-discriminative). In this condition we terminate the regeneration process and go for testing. Fig. 2(a)-(b) shows a typical situation where convergence is achieved based on misclassification probability and as many as 96% of the features are above $(\mu - \sigma)$ limit at the convergence point. Second, as we go on reducing the features, at some population, although very rare, it may happen that $\#\mathcal{F} < f$. To avoid this, we terminate the training procedure pre-matured when $\#\mathcal{F} = f$. In the next section, we present and analyze the experimental results.

3 Experimental Results

3.1 Dataset and Parameters

We use the MITOS dataset [1] for our experiments. The training data contains 35 images with 233 mitotic cells and the test data contains 15 images with 103 mitotic cells. During classification, our forest is composed of $T = 1000$ trees and the initial feature vector contains 604 features. We take $f = 15$ for each node. When the proposed forest tends to misclassification minima, the tree weights have a high mean value and a low value of standard deviation as evident from previous discussions. In this near minima situation, the rate of change of tree



(a) (b)

(c) (d)

Fig. 2. Performance of proposed method: (a) Probability of misclassification, (b) percentage of features above $(\mu - \sigma)$ limit, (c) classification performances of different methods, (d) P-R curves for different methods (proposed, CTW, WFR and WBF).

weights (and consequently κ) should be low so that the forest does not cross the minima. Hence, we take $\kappa = \sigma_\tau^n(1 - \mu_\tau^n)$ where μ_τ^n and σ_τ^n are mean and standard deviation of tree weights in the n^{th} population.

3.2 Performance Measures and Comparisons

Following [1] we evaluate recall, precision and F-measures on the test data. We obtain recall and precision values of 0.8113 and 0.8350, respectively, yielding an F-measure of 0.823. We compare our classification performance with the four best results reported for MITOS dataset [1] and the result using the method proposed in [10] (abbreviated as ERF). The four best results are from groups IDSIA, IPAL, SUTECH and NEC respectively [1]. We show the comparative results in Fig. 2(c) where the proposed method is found to outperform other competing methods in terms of F-measure. Note that the ranking of the groups in [1] are performed in terms of F-measure. However, it can be observed that the precision value in the proposed method is less compared to IDSIA. The reason is imbalance in the number of dense non-mitotic nuclei in the training dataset due to which the forest did not properly learn the corresponding signature. The precision-recall (P-R) curves obtained by varying the number of features selected for each node (f) are shown in Fig. 2(d). We also examined the performance of the method a) while keeping constant tree weight (CTW) of value 1 across all generations (populations), b) without feature reduction (WFR) (by updating only the tree weights) and c) using the proposed random forest without using biological (stroma and nuclear membrane) features (WBF). The P-R curves for each of these experiments are also presented in Fig. 2(d). We find that the area under the curve of the proposed method is the largest which indicates that both feature reductions and tree weights play significant roles in classification performances. It also indicates the importance of using the biological features. Some sample sub-images are shown in Fig. 1 along with classification results of our method and IDSIA [4]. It is notable that there is significant improvement in training and testing times compared to [4]. The training time for the proposed method is 12 minutes and the average testing time per image is 22 sec using a PC with 3.2 GHz intel Xeon processor and 16 GB memory.

4 Conclusions

We propose a novel approach for mitosis detection utilizing domain knowledge. Intensity features of the nucleus, texture features of possible stroma and features of nuclear membrane are classified with a novel regenerative random forest that performs automatic feature selection. We prove that the regeneration process converges producing maximum classification accuracy during training. We achieve lower time complexity comp ared to state-of-the-art techniques. In future, we want to use more biological features to further improve the classification performance. A generalized regenerative random forest will also be a significant direction of future research.

References

1. (2012). `http://ipal.cnrs.fr/ICPR2012/?q=node/5` Available as on (February 18, 2015)
2. Barlow, P.W.: Changes in chromatin structure during the mitotic cycle. Protoplasma 91(2), 207–211 (1977)
3. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
4. Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Mitosis detection in breast cancer histology images with deep neural networks. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part II. LNCS, vol. 8150, pp. 411–418. Springer, Heidelberg (2013)
5. Haralick, R.M., Shanmugam, K., Dinstein, I.H.: Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics (6), 610–621 (1973)
6. Huang, C.H., et al.: Automated mitosis detection based on exclusive independent component analysis. In: 21st ICPR, pp. 1856–1859. IEEE (2012)
7. Khan, A.M., et al.: A gamma-gaussian mixture model for detection of mitotic cells in breast cancer histopathology images. Journal of Pathology Informatics 4 (2013)
8. Kuru, K.: Optimization and enhancement of h&e stained microscopical images by applying bilinear interpolation method on lab color mode. Theoretical Biology and Medical Modelling 11(1), 9 (2014)

9. Malon, C.D., Cosatto, E.: Classification of mitotic figures with convolutional neural networks and seeded blob features. Journal of Pathology Informatics 4 (2013)
10. Paul, A., Mukherjee, D.P.: Enhanced random forest for mitosis detection. In: Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing, p. 85. ACM (2014)
11. Veta, M., et al.: Breast cancer histopathology image analysis: a review. IEEE Trans. Biomed. Engineering 61(5), 1400–1411 (2014)