Discovering Important Regions of Cytological Slides Using Classification Tree*

Marek Kowal[1], Andrzej Marciniak[1],
Roman Monczak[2], and Andrzej Obuchowicz[1]

[1] Univeristy of Zielona Góra,
Institute of Control and Computation Engineering, Poland
{M.Kowal,A.Marciniak,A.Obuchowicz}@issi.uz.zgora.pl
[2] Department of Pathomorphology, Regional Hospital in Zielona Góra, Poland
R.Monczak@issi.uz.zgora.pl

Abstract. Modern digital microscopy systems allow imaging of biological material with very high accuracy. Paradoxically, this gives rise to many problems because huge amounts of raw data significantly increase the time required by specialist to analyze them. As a result, we obtain a time-consuming diagnostic process and reduction of the number of patients being diagnosed. The paper presents a method of discovering regions of the cytological image, which are essential to correct diagnosis. The purpose of this method is to help pathologists by indicating regions of the image that should be analyzed first. Moreover, method can be used to explore new, previously unknown features discriminating benign from malignant lesions. Multi-level image thresholding is responsible for image segmentation and is the core of the proposed system. Thresholds are evaluated by predictive accuracy on testing dataset. Honey Bee Mating Optimization (HBMO) algorithm is applied to find the optimal threshold set. The developed method was successfully applied to analyze cytological images of biological material collected from a breast tumor.

1 Introduction

Nowadays whole slide digital imaging is becoming a standard in cytological examinations. Glass slides are scanned, converted to digital slides and can be viewed on computer monitor. The scanning accuracy is usually very high, and the output slide has a large size. Visual analysis of the entire slide can then be very time-consuming. To overcome this problem, numerous automatic segmentation algorithms was proposed to extract features of cells or nuclei from cytological images. Unfortunately, it still remains a big challenge due to noise, cell and nuclei overlapping or high variation of image parameters. Most of the approaches lack of generality, and give expected results only for specific images.

In this work we limit our researches and discussion to the issue of breast cancer diagnosis based on fine needle biopsy (FNB) cytological images. The task at hand is to classify a case as benign or malignant. Many researchers have already

* This research was partially supported by National Science Centre in Poland.

R.S. Choraś (ed.), *Image Processing & Communications Challenges 6*,
Advances in Intelligent Systems and Computing 313, DOI: 10.1007/978-3-319-10662-5_9

studied similar issues [13,15,11]. In our previous works we used a segmentation method based on the combination of adaptive thresholding, clustering in color space, fast marching and seeded watershed [10,8,5]. We tested the predictive accuracy on 450 images recorded by an analog video camera mounted atop of microscope. These experiments gives satisfactory results reaching 99.3% of correct classifications. Unfortunately, direct application of these methods for digital slides acquired with use of virtual microscopy system gave unsatisfactory predictive accuracy, mainly due to oversegmentation. To cope with this problem we applied classification driven multi-level thresholding to image segmentation [12]. Although regions extracted by this approach may not correspond to any medically interpretable objects but numerical experiments proved existence of implicit relationship between features of segmented regions and classes of breast cancer [9]. Here we propose a new system for exploring important regions of cytological images. The system is based on classification driven multi-level thresholding. In the first step, the HBMO algorithm is employed to find suboptimal thresholds in the sense of predictive accuracy. Then, classification tree is constructed using features generated by the suboptimal set of thresholds. Importance of image regions is determined based on classification tree structure. The system speed up the analysis of the cytological image by indicating which fragments of the image should be inspected in the first place. It can be also used to discover new, previously unknown features discriminating benign from malignant cases.

The remainder of this paper is organized as follows. In Section 2, material and methods are described. Section 3 gives the description of the experiments and study of the results obtained for implemented method. Concluding remarks are given in Section 4.

2 Material and Methods

2.1 Image Database

The database contains 50 slides of the cytological material obtained FNB. The material was collected from 50 patients of the clinic in Zielona Góra, Poland. The set contains 25 benign and 25 malignant lesions cases. Smears from the biological material were fixed in spray fixative and dyed with hematoxylin and eosin. Cytological preparations were then digitalized into virtual slides using the Olympus VS120 Virtual Microscopy System. The average size of the slides is approximately 200 000 $\times$ 100 000 pixels. On each slide a pathologist selected 11 areas which illustrated a sufficient amount of biological material. These areas were archived in RGB (8 bit/channel) TIFF files of size 1583 $\times$ 828pixels. The number of areas per one patient was recommended by the pathologists at the hospital and allows for a correct diagnosis. The image database contains 550 images (11 images per patient). Both malignant and benign sets contain the same number of images (275 images describing 25 cases). All cancers were histologically confirmed and all patients with benign disease were either biopsied or followed for a year.

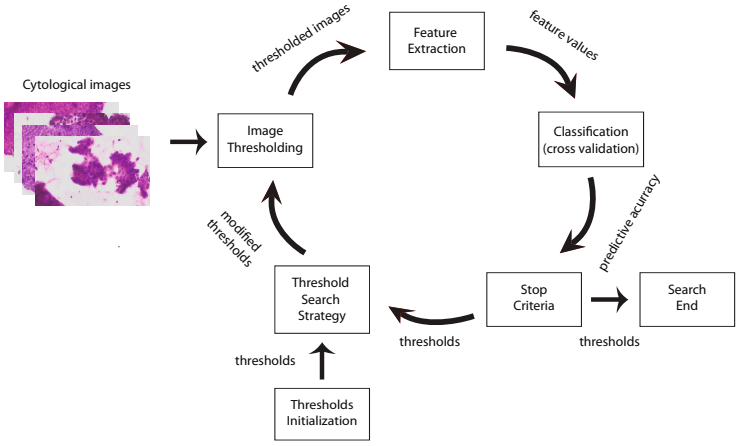

Fig. 1. Classification driven image thresholding

2.2 Multi-level Thresholding

Image thresholding boils down to discrete optimization problem with constraints. The key step of the algorithm is to determine set of thresholds corresponding to the maximum (or minimum) of the evaluation function:

$$T^* = \arg\max_T \{f(T) : T \in X \subset \mathbb{Z}^K\} \tag{1}$$

where $T = [t_1, t_1, \ldots, t_K]$ is a vector of thresholds, $f : \mathbb{Z}^K \rightarrow \mathbb{R}$ is an evaluation function, $X = \{T \in \mathbb{Z}^K : 0 \leq t_1 \leq t_2 \leq \ldots \leq t_K \leq L\}$ is a set of feasible solutions and L is a number of gray levels. In literature, a great variety of evaluation criteria have been already proposed [17,19]. Most of them are based on homogeneity measures. In cytological analysis, images are segmented to find the objects of interest and than their features are extracted to discriminate malignant from benign cases. Unfortunately, homogeneity based criterions can not guarantee obtaining high predictive accuracy because threshold selection is performed independently from the results of classification. To overcome this problem we propose to use a predictive accuracy to evaluate the thresholds. This approach originate from well known wrapper based feature selection strategy where predictive model is employed to score feature subsets [7]. In our approach classification tree is used to evaluate threshold sets [18,2]. The scheme of the approach is shown in Fig. 1. In the first step, thresholds are applied to segment images. Next, segmented regions are measured to compute features. Finally, images are classified and the predictive accuracy is estimated using n-fold cross-validation procedure [4]. Since the method trains a classification tree for each training subset, it is very computationally intensive. To deal with such computational effort we decided to use HBMO searching strategy. The HBMO algorithm belongs to the general class of swarm intelligence methods that models the behaviors of social insects. It is inspired by the marriage behavior of honey-bees [1,6]. The pseudo code of main steps of HBMO is given in Algorithm 4.

Algorithm 4. HBMO algorithm

```
Initialize randomly drones D and queen Q
while stop condition not met do
   while speed is above the threshold AND spermatheca is not full do
      Select randomly drone D_rand from set D
      Compute the probability of mating for drone D_rand
      if the drone D_rand passes the probabilistic condition then
         Add its sperm to the queens spermatheca
      end if
      Decrease the speed of queen
   end while
   for i = 1 : number of broods do
      Select randomly sperm P_rand from spermatheca
      Generate brood B_i based on P_rand and Q
   end for
   for j = 1 : number of mutations do
      Select randomly brood B_rand
      Apply mutation to B_rand
   end for
   Find the brood B_best with the highest value of objective function
   if objective function of B_best > objective function of queen Q then
      Replace current queen Q with B_best
   end if
end while
```

2.3 Discovering Important Regions of Image

There are many different approaches to the problem of discovering important regions of the image [3,14,16]. In our method, training images are segmented using suboptimal thresholds T^* determined by procedure described in Section 2.2. Then, features are computed for all images. Discriminative ability of individual features can be estimated by classification tree [2,18]. The full binary decision tree based on Gini diversity index as splitting criterion is generated. Then, the estimates of input feature importance are computed for tree by summing changes in the risk due to splits on every feature. At each node, the risk is estimated as node impurity and weighted by the node probability. Variable importance associated with this split is computed as the difference between the risk for the parent node and the total risk for the two children. Features with insignificant importance values are removed from the tree. Pruning is repeated until a significant deterioration in the predictive accuracy is observed. Structure of pruned tree is used to evaluate the region importance. Each node in the classification tree is associated with a single feature, and thus also with a single region in the image. The higher level the node occupies in the tree, the more important is region associated with him. The contours of important regions are mapped to original cytological image and in this way can be presented to pathologists. Sample results are presented in Fig. 5.

3 Numerical Experiments

To find optimal threshold set for 550 images described in Section 2.1, we performed the procedure described in Section 2.2. Six types of features are computed for each segmented region: area (A), perimeter (P), euler number (EN), variance in R channel (varR), variance in G channel (varG) and variance in B channel (varB). Thus each image is described by $n_f = 3(K + 1)$ features. For classifying images we use classification tree. Predictive accuracy is estimated using n-fold cross-validation procedure. There are 50 folds (the number of cases), and each fold consist of 11 images that belong to a single case. The images belonging to the same case are never at the same time in the training and testing set. The predictive accuracy is defined as the percentage ratio of successfully classified images to the total number of images. Remembering that HBMO belongs to class of stochastic global optimization algorithms, 10 runs of algorithm was performed for each number of thresholds. Computational effort was constant for each experiment and set to 500 evaluations. To find the optimal number of thresholds K^*, the searching procedure was performed for $K = 2, \ldots, 10$ thresholds. The

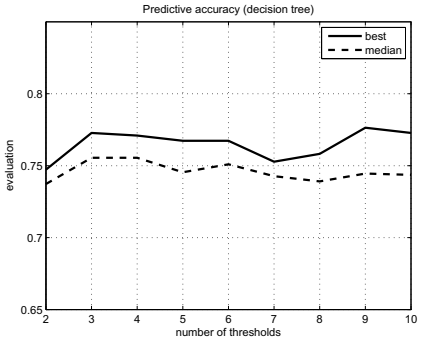

Fig. 2. The best and median classification results depending on the number of thresholds

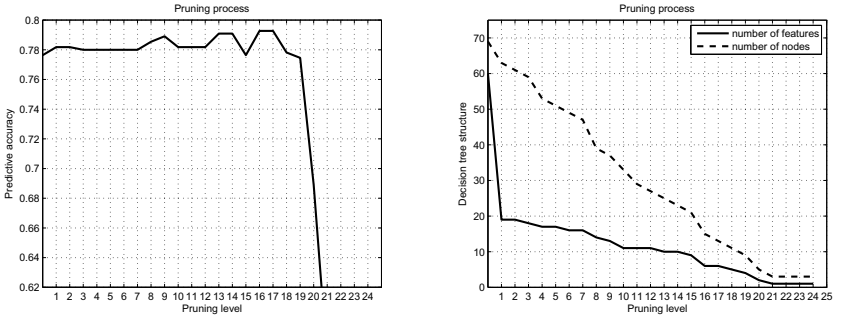

Fig. 3. Decision tree performance (left), and structure (right) over pruning process

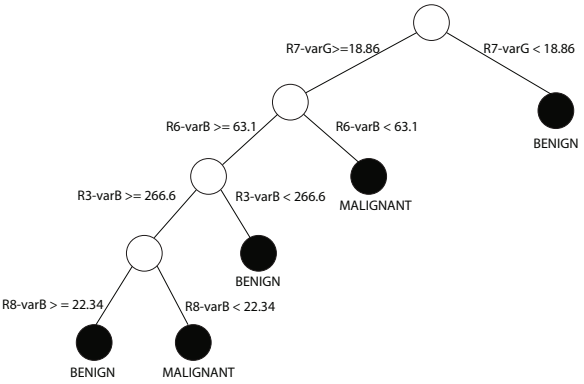

Fig. 4. The decision tree after pruning process (Rn-feature denotes the feature value for the $n-th$ region)

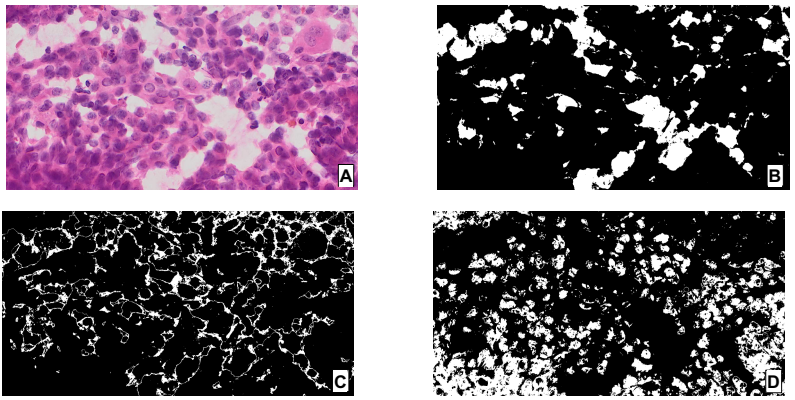

Fig. 5. Important regions: (A) original image, (B) binary mask for R7, (C) binary mask for R6, (D) binary mask for R3

relationship between the number of thresholds and predictive accuracy is shown in Fig. 2. The best classification accuracy 77.64% has been achieved by following thresholds $T = [57, 75, 113, 163, 174, 200, 225, 243, 249]$. Then, these thresholds are used to generate training data in order to induce the final classification tree. Next, tree is pruned to remove insignificant features. This process is described by Fig. 3. Pruned tree is shown in Fig. 4. The importance of image regions is determined based on position of nodes in the tree. Sample results are plotted in the form of binary masks presented in Fig. 5.

4 Conclusions

This study shows that proposed system of breast cancer diagnosis can discriminate benign from malignant cases with predictive accuracy equal to 77.64%.

Such result was achieved for image classifications accuracy. But single patient is described by 11 images and by applying majority voting for all of the classifier outcomes we can compute predictive accuracy for patient. This simple modification increases the predictive accuracy to 94%. However, it is too early to conclude that the system is ready to assist medical tasks by suggesting a diagnosis to pathologists. Higher predictive accuracy can be achieved if better quality feature space will be found. It seems that particularly important can be textural features of segmented objects. Further research will be performed to check this hypothesis. In its current form, the system can help less experienced pathologists by showing them image regions that are important from the point of view of breast cancer diagnosis. Our system can help also experienced pathologist to explore new unknown features discriminating benign from malignant cases. The main advantage of the proposed system is its generality because it can be relatively easily adapted to the analysis of microscopic images of any type. Although the training of the system is relatively computationally expensive but it is done off-line and only once. Image analysis is performed by tuned system very quickly.

References

1. Abbass, H.A.: MBO: Marriage in honey bees optimization - a haplometrosis polygynous swarming approach. In: Proceedings of the 2001 Congress on Evolutionary Computation, vol. 1, pp. 207–214. IEEE (2001)
2. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and regression trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey (1984)
3. Choraś, R., Andrysiak, T., Choraś, M.: Integrated color, texture and shape information for content-based image retrieval, pattern analysis and applications. Pattern Analysis and Applications 10(4), 333–343 (2007)
4. Devijver, P.A., Kittler, J.: Pattern Recognition: A Statistical Approach. Prentice-Hall, London (1982)
5. Filipczuk, P., Krawczyk, B., Woźniak, M.: Classifier ensemble for an effective cytological image analysis. Pattern Recognition Letters 34(14), 1748–1757 (2013)
6. Horng, M.H.: A multilevel image thresholding using the honey bee mating optimization. Applied Mathematics and Computation 215(9), 3302–3310 (2010)
7. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence 97(1), 273–324 (1997)
8. Kowal, M., Filipczuk, P.: Nuclei segmentation for computer-aided diagnosis of breast cancer. Int. J. Appl. Math and Comp. Sci. 24(1), 19–31 (2014)
9. Kowal, M., Filipczuk, P., Marciniak, A., Obuchowicz, A.: Swarm optimization and multi-level thresholding of cytological images for breast cancer diagnosis. In: Burduk, R., Jackowski, K., Kurzynski, M., Wozniak, M., Zolnierek, A. (eds.) CORES 2013. AISC, vol. 226, pp. 611–620. Springer, Heidelberg (2013)
10. Kowal, M., Filipczuk, P., Obuchowicz, A., Korbicz, J., Monczak, R.: Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images. Computers in Biology and Medicine 43(10), 1563–1572 (2013)
11. Malek, J., Sebri, A., Mabrouk, S., Torki, K., Tourki, R.: Automated breast cancer diagnosis based on GVF-Snake segmentation, wavelet features extraction and fuzzy classification. Journal of Signal Processing Systems 55(1-3), 49–66 (2009)

12. Marciniak, A., Kowal, M., Filipczuk, P., Korbicz, J.: Swarm intelligence algorithms for multi-level image thresholding. In: Korbicz, J., Kowal, M. (eds.) Intelligent Systems in Technical and Medical Diagnostics. AISC, vol. 230, pp. 301–311. Springer, Heidelberg (2013)
13. Mazurek, P., Oszustowska-Mazurek, D.: From the slit-island method to the ising model: Analysis of irregular grayscale objects. Int. J. Appl. Math and Comp. Sci. 24(1), 49–63 (2014)
14. Mikolajczyk, K., Schmid, C.: Scale & affine invariant intrest point detectors. Int. J. Computer Vision 60(1), 63–86 (2004)
15. Niwas, S.I., Palanisamy, P., Sujathan, K., Bengtsson, E.: Analysis of nuclei textures of fine needle aspirated cytology images for breast cancer diagnosis using complex daubechies wavelets. Signal Processing 93(10), 2828–2837 (2013)
16. Ogiela, M.R., Trzupek, M., Tadeusiewicz, R.: Intelligent image content semantic description for cardiac 3d visualizations. Int. J. Engineering Applications of Artificial Intelligence 24(8), 1410–1418 (2011)
17. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Sys. Man. and Cyber. 9, 62–66 (1979)
18. Quinlan, J.R.: Induction of decision trees. Machine Learning 1(1), 81–106 (1986)
19. Sezgin, M., Sankur, B.: Survey over image thresholding techniques and quantitative performance evaluation. J. Electronic Imaging 13(1), 146–165 (2004)