In: J. Crespo, V. Maojo, and F. Martin (Eds.), Medical Data Analysis, Springer Verlag, lncs 2199, 2001, p.219-224.

Classification of HEp-2 Cells using Fluorescent Image Analysis and Data Mining

Petra Perner

Institute of Computer Vision and Applied Computer Sciences
Arno-Nitzsche-Str. 45, D-04277 Leipzig, Germany
ibaiperner@aol.com http://www.ibai-research.de

Abstract. The cells that are considered in this application for an automated image analysis are Hep-2 cells which are used for the identification of antinuclear autoantibodies (ANA). Hep-2 cells allow for recognition of over 30 different nuclear and cytoplasmic patterns, which are given by upwards of 100 different autoantibodies. The identification of the patterns has recently been done manually by a human inspecting the slides with a microscope. In this paper we present results on image analysis, feature extraction, and classification. Starting from a knowledge acquisition process with a human operator, we developed an image analysis and feature extraction algorithm. A data set containing 162 features for each entry was set up and given to a data mining algorithm to find out the relevant features among this large feature set and to construct the classification knowledge. The classifier was evaluated by cross validation. The results show the feasibility of an automated inspection system.

1 Introduction

In this paper, we present results on the analysis and classification of cells using image analysis and data mining techniques. The kinds of cells that are considered in this application are Hep-2 cells, which are used for the identification of antinuclear autoantibodies (ANA). ANA testing for the assessment of systemic and organ specific autoimmune disease has increased progressively since immunofluorescence techniques were first used to demonstrate antinuclear antibodies in 1957. Hep-2 cells allow for recognition of over 30 different nuclear and cytoplasmic patterns, which are given by upwards of 100 different autoantibodies.

The identification of the patterns has up to now been done manually by a human inspecting the slides with the help of a microscope. The lacking automation of this technique has resulted in the development of alternative techniques based on chemical reactions, which have not the discrimination power of the ANA testing. An automatic system would pave the way for a wider use of ANA testing.

We present our results on image analysis, feature extraction, and classification based on images that were taken by an digital image acquisition unit under real clinical conditions (see Sect. 2). Starting from a knowledge-acquisition process with a human operator (see Sect. 3), we developed an image analysis and a feature extraction algorithm, described in Sect. 4 and Sect. 5. A data set containing 162 features for each entry was set up and given to our data mining tool to find out the relevant features

among this large feature set and to construct the structure of the classifier, see Sect. 6. The classifier was evaluated by cross validation. The results show the feasibility of an automatic inspection system (see Sect. 7).

2 Image Acquisition

The images were taken by a digital image-acquisition unit consisting of a microscope AXIOSKOP 2 from Carl Zeiss Jena, coupled with a color CCD camera Polariod DPC [1]. The digitized image were of 8-bit photometric resolution for each color chanel with a per pixel spatial resolution of 0.25 μm. Each image was stored as a color image on the hard disk of the PC. Form there it was accessed for further calculation.

3 Knowledge Acquisition

For our experiment we used fluorescence images from six different classes (see Fig. 1).

In a knowledge-acquisition process [2] with a human operator, using an interview technique and a repertory grid method, we acquired the knowledge of this operator, while classifying the different cell types. Some of this knowledge is shown in table 1. The symbolic terms show that a mixture of different image information is necessary for classification. The operator uses the color intensity as well as some texture information. In addition, the appearance of the cell parts within the cells are of importance, like “dark nuclei”, which also requires spatial information.
We started out to develop the image analysis procedure and constructed a feature set, which seems to be powerful enough to describe this symbolic knowledge. It is left to the data mining experiment to find out the relevant features for classification and to show us gaps in our description of the domain.

Table 1. Some knowledge about the class description given by a human operator

Class	ClassName	Description
Homogeneous nuclei fluorescence	Class_1	Smooth and uniform fluorescence of the nuclei. Nuclei appear sometimes dark. The chromosome fluorescence is weak up to very intense
Fine speckled nuclei fluorescence	Class_2	Dense fine speckled fluorescence
...	...	...
Nuclei fluorescence	Class_6	Nuclei are weakly homogenous or fine grained and can be hardly discerned from the background

4 Image Analysis

The color image has been transformed into a gray level image. Automatic thresholding has been performed by the algorithm of Otsu [3]. The algorithm can localize the cells with their cytoplasmatic structure very well, but not the nuclear envelope itself. We then applied morphological filters like dilation and erosion to the image in order to get a binary mask for cutting out the cells from the image. Since there are only cells of one type in an image, overlapping cells have not been considered for further analysis.

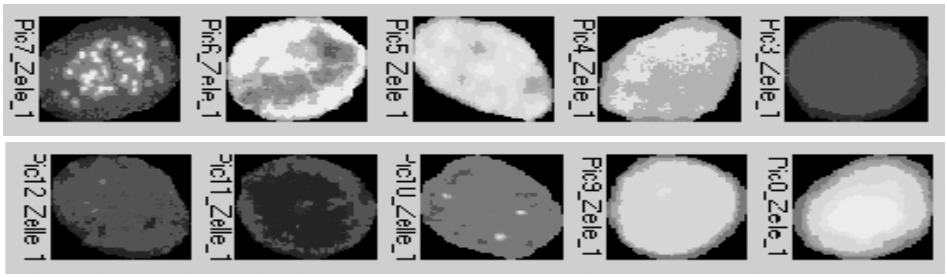

Fig. 1. Examples of Cell Images for 10 different Classes

The gray levels ranging from 0 to 255 are quantized into 18 intervals *t*. Each subimage *f(x,y)* containing only a cell gets classified according to the gray level into *t* classes, with *t={0,1,2,..,18}*. For each class a binary image is calculated containing the value "1" for pixels with a gray level value falling into the gray level interval of class *t* and value "0" for all other pixels. The quantization of the grey level into 18 intervals was done based on a logarithmic characteristic curve. We call the image *f(x,y,t)* in the following class image. Object labeling is done in the class images with the contour following method [4]. Then features from these objects are calculated for classification.

4 Feature Extraction

For the objects in each class image features are calculated for classification. The first one is a simple Boolean feature which expresses the occurrence or none occurrence of objects in the class image. Then the number of objects in the class image is calculated. From the objects the area, a shape factor, and the length of the contour are calculated. However, not a single feature of each object is taken for classification, but a mean value for each feature is calculated over all the objects in the class image. This is done in order to reduce the dimension of the feature vector. We also calculate the frequency of the object size in each class image. The list of features and their calculation is shown in table 2.

Table 2. List of Features and their Calculation

Description	**Name**	**Type**	**Formula**
Object occurred in class image *t*	***Gray_t***	boolean	yes or no
Number of objects in class image *t*	***Count_t***	numerical	n(t)
Mean area of objects in class image *t*	***Area_t***	numerical	$\overline{A}(t) = \frac{1}{n(t)} \sum_{i=1}^{n(t)} A_i(t)$
Relative mean area of objects in class image *t* to area of cell	Rarea_t	numerical	$RA(t) = \frac{\overline{A}(t)}{A_{cell}}$
Mean shape factor for objects in class image *t*	*Form_t*	numerical	$\overline{F}(t) = \frac{1}{n(t)} \sum_{i=1}^{n(t)} 10 \cdot \frac{A_i(t)}{u_i(t)}$ with $\boldsymbol{u_i(t)}_{,\ contour}$ being the length of the *i-th* object in class image *t*.
The contour length of a single object is $u = l + \sqrt{2} \cdot m$ with *l* being the number of contour pixels having odd chain coding numbers and *m* being the number of contour pixels having even chain coding numbers.			
Mean contour lenght of objects in class image *t*	L***ength_t***	numerical	$\overline{u}(t) = \frac{1}{n(t)} \sum_{i=1}^{n(t)} u_i(t)$

6 Learning of Classifier Knowledge

For each of the six classes, we had 19 data sets; each data set contained 162 features, obtained from each of the eigthteen class images. The whole data set has 105 samples. Based on that data set we acquired the knowledge for classification. We used a binary [5] and n-ary decision tree induction algorithm [6] realized in our data mining tool *DECISIONMASTER* [7]. The n-ary decision tree can split up a numerical feature into more than two intervals which leads sometimes to a better performance than the one of a binary decision tree. The learning algorithm selects from the data set the most promising features and constructs the structure of the classifier during the learning phase. The resulting decision tree is shown in Fig. 2. The true error rate was estimated by cross validation [8], which works properly on small sample sets. The error rate for both classifier is shown in table 3. The unpruned binary decision tree shows the best result.

Table 3. Error Rate for the Classifier estimated with Cross Validation

Method	**Unpruned Tree**	**Pruned Tree**
Binary Tree Induction	13.33 %	15.24 %
N-Ary Tree Induction	20.00 %	20.00 %

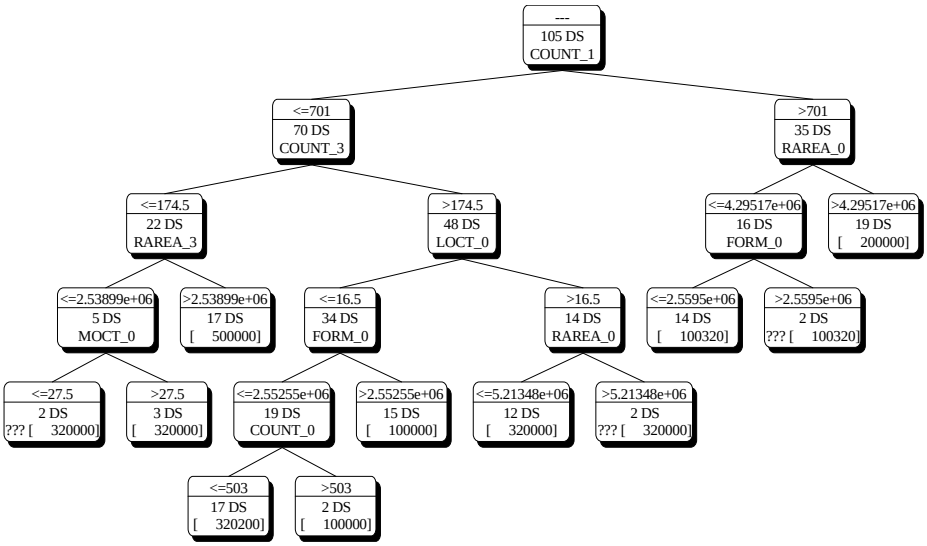

Fig. 2. Resulting Decision Tree

7 Results and Discussion

The learning algorithm only recognizes 10 features as relevant features. The most discriminating feature is the feature *count_1*, which means that there occur or do not occur objects in the class image one. This goes conform to what we can see in the class images for all cell types. There are classes having lighter objects within the cell and there are other classes having darker objects within the cell. The next important feature is the number of objects in class three and the relative area of objects in class_0, which can also be confirmed by looking at the class images.

The error rate of 13.33 % is a very good result. For the type of application presented in this paper there has not existed any automated feature extraction and classification system up to now.

Recently, we are collecting more samples for all classes as soon as the class occurred in practice. Beyond that we want to improve the accuracy of the system by defining new features and incorporating these features in our data mining experiment.

8 Conclusions

In this paper, we have shown the feasibility of an automatic system for Hep-2 cell analysis and classification. The classification problem is a multi-class problem. In our application 6 classes have to be distinguished. We developed an image analysis and a feature extraction algorithm, which classifies the subimage containing only the cells into 18 class images and calculates from each class image features for classification. The resulting data set has 162 features for each entry. Based on this data set, it is possible to obtain the relevant features with data mining techniques. Only 10 features are necessary in order to classify nine classes with an accuracy of 86,67%. We evaluated our result with cross validation. We have shown that data mining techniques are powerful techniques for determining the relevant features as well as the classifier structure. Furthermore, they can work on small data sets.

Further work will be done to improve the classification accuracy. Therefore, we will develop new features that can better describe the basic properties of the different classes.

Acknowledgement

The work presented in this paper is part of the project *LernBildZell* funded by the German Ministry of Economy.

References

1. U. Sack, Digital Image Acquisition Unit for Fluorescent Images, IBaI Report 2000
2. P. Perner, A knowledge-based image inspection system for automatic recognition, classification, and process diagnosis, Machine Vision and Applications (1994) 7:135-147.
3. Otsu, N., A threshold selection method from gray-level histograms, IEEE Trans., vol. SMC-9, Jan. 1979, p. 38-52.
4. H. Niemann, Pattern Analysis and Understanding, Springer Verlag 1990.
5. J.R. Quinlain,"Induction of Decision Trees," Machine Learning 1(1986): p. 81-106.
6. P. Perner and S. Trautzsch, Multinterval Discretization for Decision Tree Learning, In: Advances in Pattern Recognition, A. Amin, D. Dori, P. Pudil, and H. Freeman (Eds.), LNCS 1451, Springer Verlag 1998, S. 475-482
7. Data Mining Tool Decision Master http://www.ibai-solutions.de.
8. S.M. Weiss and C.A. Kulikowski, Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems. Morgan Kaufmann, San Mateo, 1990.