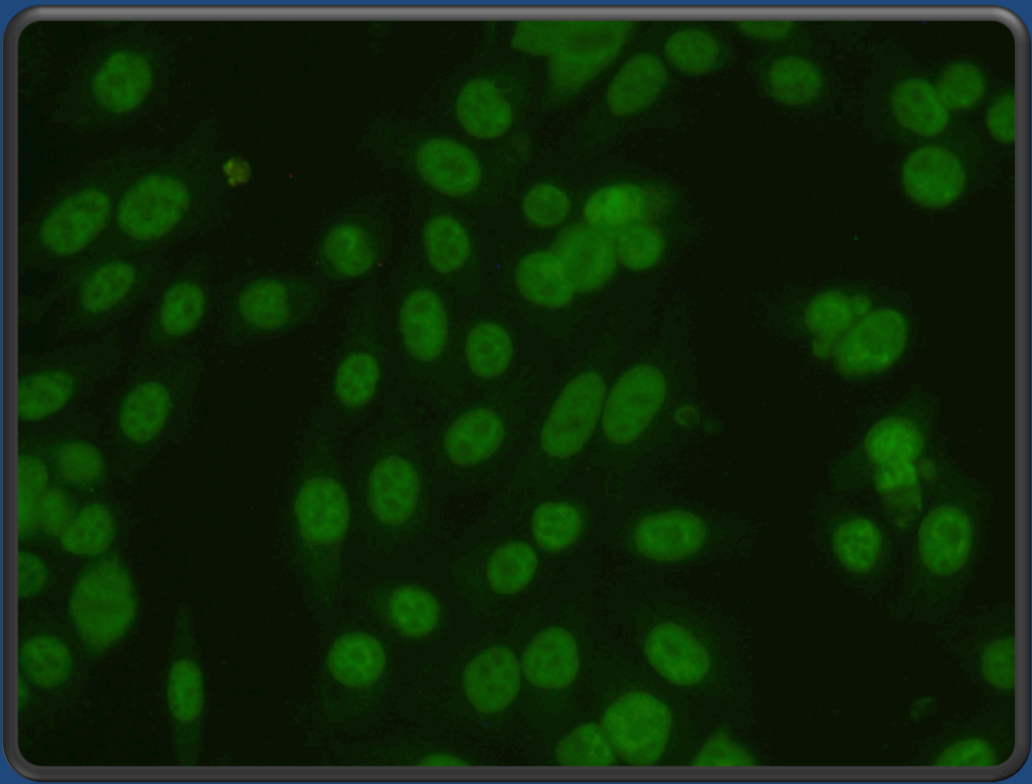
1st International Contest on
HEp-2 Cells Classification

hosted by the
21th International Conference on Pattern Recognition (ICPR 2012)



Organizers

Gennaro Percannella - *(pergen@unisa.it, University of Salerno, Italy)*
Pasquale Foggia - *(pfoggia@unisa.it, University of Salerno, Italy)*
Paolo Soda - *(p.soda@unicampus.it, University "Campus Bio-Medico" of Roma, Italy)*

http://mivia.unisa.it/hep2contest

Table of content

Optimizing a Classification System for HEp-2 Cells by Evolutionary Computation

Daniel Mateos-García, Jorge García-Gutiérrez, J. Cristóbal Riquelme-Santos

Department of Computer Science, University of Seville.
Avda. Reina Mercedes S/N, 41012 Seville (Spain)

mateosg@us.es (Daniel Mateos-García), jorgarcia@us.es (Jorge García Gutiérrez),
riquelme@us.es (J. Cristóbal Riquelme-Santos)

Abstract

In this work, we describe a classification system to automatically recognize the pattern of HEp-2 cells within IIF images. For this purpose we have carried out several steps to preprocess the data, select a proper predictor and generate a model. The use of Evolutionary Computation in the feature selection step and the optimization of the classification system is the main contribution of this work.

1. Method

The proposed pattern recognition system has been carried out following the next steps (see Figure 1):

1. We have chosen a standard number of pixels for all the data. In our case the considered resolution for the training and testing steps has been of 100x100. The system resizes the image data to the mentioned resolution.
2. With respect to the training data, every image was transformed to one instance and included into a unique training file with ARFF format [4].
3. The number of features (pixels) of the training file was reduced with the CFS evaluator and considering the Best First algorithm as searching method [1]. In a second phase, a Genetic Algorithm (GA) improved the quality of the final set of features. The fitness of the GA is the accuracy of the classification subsystem that will be described in the next step. To obtain the average of the accuracy, the training file was randomly divided into two folds and then they were used as training and testing data twice (2 x holdout 85%-15%).
4. The classification subsystem used for label predictions is J48 (C4.5) [2]. This predictor has been boosted by the AdaBoost M1 method [3]. This decision has been made by discarding the rest of the classifiers that integrate the framework used (Weka [4]) and taking into account the observed accuracy of each one. Furthermore, a GA optimized the AdaBoost parameters. Regarding the GA design, each individual represents a combination of values of these parameters, and the fitness is the accuracy of the classification subsystem using holdout (2 x holdout 85%-15%) with the pre-processed training data (features selected).
5. The final classifier system was trained with the full pre-processed training data and the optimized AdaBoost + C4.5 stack. In the testing step, each testing image will be resized, transformed in an ARFF file with only one instance and unknown label, and filtered with the same feature set calculated in the training step.

References

[1] M. A. Hall (1998). Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand.

[2] Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

[3] Yoav Freund, Robert E. Schapire: Experiments with a new boosting algorithm. In: Thirteenth International Conference on Machine Learning, San Francisco, 148-156, 1996.

[4] Weka Machine Learning Project. Weka. URL http://www.cs.waikato.ac.nz/ml/weka/

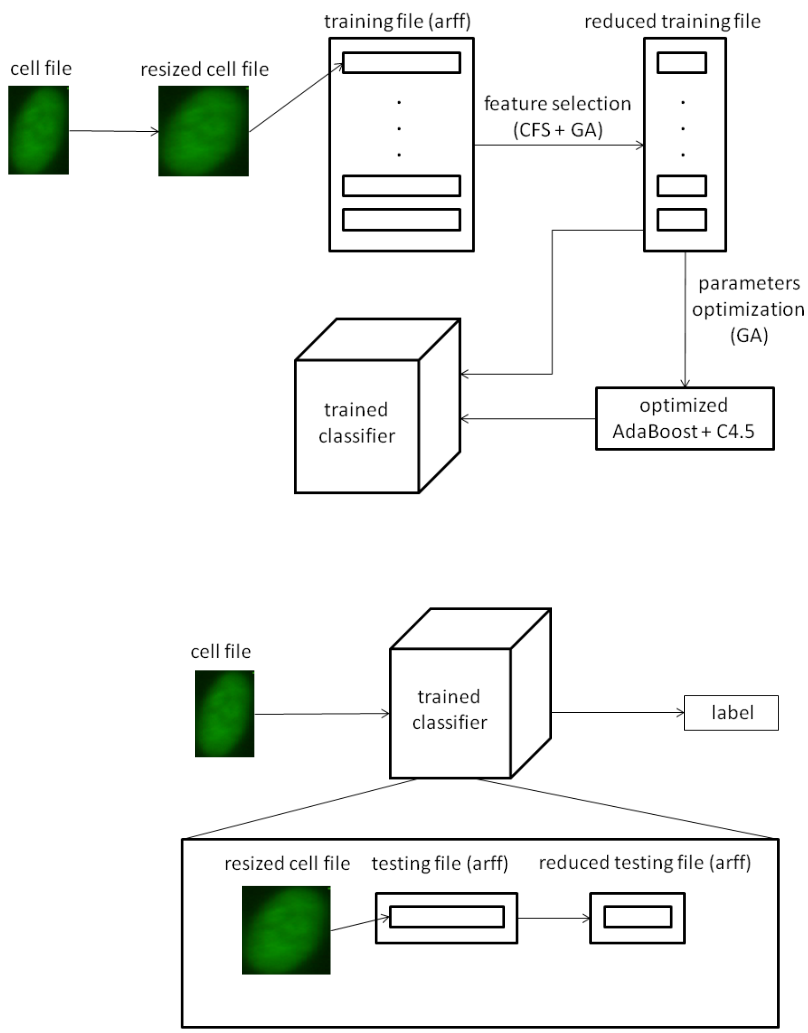


Figure 1. Training and testing processes