Automated cell phenotype image classification combining different methods

Loris Nanni¹  loris.nanni@unibo.it
Chun-Nan Hsu²  chunnan@iis.sinica.edu.tw
Alessandra Lumini¹  alessandra.lumini@unibo.it
Yu-Shi Lin²  bathroom@iis.sinica.edu.tw
Chung-Chih Lin²

¹Department of Electronic, Informatics and Systems (DEIS), Università di Bologna, Via Venezia 52, 47023 Cesena, Italy.
²Institute of Information Science, Academia Sinica, Taipei, Taiwan.

Keywords: cell phenotype image classification; subcellular location; machine learning techniques; ensemble of neural networks; AdaBoost.

Abstract

In this paper our aim is to study how an ensemble of classifiers can improve the performance of a machine learning technique for cell phenotype image classification. We want to point out some of the advantages that an ensemble of classifiers permits to obtain respect a stand-alone method. Finally, the preliminary results on the 2D-HeLa dataset, obtained by the fusion between a random subspace of Levenberg-Marquardt neural networks and a variant of the AdaBoost, are reported. It is interesting to note that the proposed system obtains an outstanding 97.5% Rank-1 accuracy and a >99% Rank-2 accuracy.

1. Introduction

(Peng, 2008) has defined as “bio-image informatics” all the bioinformatics applications and problems where the aim is to extract biological information from image data.

The automated cell phenotype image classification problem is an interesting “bio-image informatics” application. It can be used for establishing knowledge of the spatial distribution of proteins within living cells. It is a very useful tool since it permits to screen systems for drug discovery or for early diagnosis of a disease.

In the last years several methods that achieve an accuracy >90% are proposed, e.g. (Nanni & Lumini, 2008; Lin et al., 2007), which is higher than the performance of an human expert.

The main part of the proposed approaches are based on the use of image statistics (Chebira et al., 2008) for extracting a feature vector from an input image and a general purpose classifier for the classification step. Recently, to improve the performance some authors proposed to combine different set of features (Hamilton et al., 2007; Nanni & Lumini, 2008; Lin et al., 2007). Another proposed method for improving the performance is to use ensemble of classifiers instead of a stand-alone method (Lin et al., 2007; Chebira et al., 2007). In particular in (Chebira et al., 2007) a multi-resolution approach is proposed and it gains a very high accuracy rate (95%) in the 2D HeLa dataset.

2. Machine Learning and ensemble of classifiers

A very important subset of the Pattern Recognition techniques is the Machine learning, in those methods the parameters of a given approach are learned from a given dataset, named training set. Notice that if the learning is performed when the labels of the patterns that belong to the training set are known it is called supervised learning (Duda et al., 2000). While when the learning is performed in a unlabelled dataset it is called unsupervised learning. In the unsupervised learning the aim is to find a good partition of the dataset, e.g. according a clustering algorithm, where each partition contain patterns with similar characteristics.

A generic architecture of a Pattern Recognition System may be defined in five steps, as reported in Figure 1.
The main problem with supervised learning when few examples of the same class (e.g. few images of the same class) are available is that the error rate in the test set is higher than the training error. If enough samples are available it is possible to extract from the training set a validation set and to use the validation set to find the parameters of a given method. In Figure 2 we plot the average error rate of the training set, of the test set and of the validation set.

From the machine learning point of view, we can answer to the question “What is Learning?” as a method for changing the structures of the program to maximize (or minimize) a given objective function. A simple example is the back-propagation neural network that modify its parameters to minimize the errors obtained considering the training patterns.

In the Pattern Recognition literature there is a huge number of works that show that ensembles of classifiers enhance the performance of a stand-alone method.

The main general methods (a good survey for ensemble of classifier in bioinformatics is (Nanni & Lumini, 2009a)) for building an ensemble of classifiers are:

- **Patterns perturbation**, i.e. each different classifier is trained with a different set of training patterns. The most well-known methods based on pattern perturbation are Bagging and Boosting. In Bagging each base classifier is trained using a different bootstrap replicate, while in Boosting the classifiers are constructed iteratively considering the performance of the previous set of classifiers.

- **Features perturbation**, i.e. each different classifier is trained with a different set of features. A well know feature perturbation method is the Random Subspace Method (RS) (Ho, 1998). In RS each classifier is trained using a randomly selected feature subset.

- **Parameters perturbation**, i.e. each different classifier is trained with a different set of classifier's parameters.

To combine a set of classifiers several different combination strategies have been proposed (Kittler et al., 1998). One of the most used is the sum rule (used also in this paper), it consists in summing the scores provided by the different classifiers.

### 3. Texture descriptors

The texture descriptors are used to describe each image with a vector, the “ideal” descriptor should be invariant to image scale, image rotation and robust to changes in illumination.

In the literature two different types of descriptors have been proposed and tested (Nanni & Lumini, 2008b):

- **global descriptors**, the features are extracted from the whole image
- **local descriptors**, the features are extracted separately considering local regions.

In the following sub-sections, the used descriptors are briefly presented (for a more detailed review see (Nanni & Lumini, 2009) or (Nanni et al., 2009b)).

#### 3.1 Invariant Local Binary Patterns (LBP)

LBP is calculated considering the joint distribution of grey values of a circularly symmetric neighbor set of $P$ pixels, of a pixel $c$, on a circle of radius $R$. In this paper, LBP is used as local descriptors (see (Nanni & Lumini, 2008) for more details).

Here we concatenate the features extracted considering different parameters configurations: $(P=8; R=1), (P=16; R=2), (P=24; R=3)$. 

---

**Figure 1.** Generic architecture of a Pattern Recognition System.

**Figure 2.** The learning curves show the average error rate. From: Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification. Copyright 2001 by John Wiley & Sons, Inc.
3.2 Threshold adjacency statistics (TAS)

Threshold adjacency statistics (TAS) has been proposed by (Hamilton et al., 2007). The features are obtained by calculating three 9-bin histograms, starting from three different binary images, considering the number of adjacent white pixels for each white pixel in the binary image. Each binary image is built using a different threshold. In this paper, TAS is used as global descriptors.

3.3 Haralick texture features (HAR)

This set of features is calculated by extracting 13 statistical measures (i.e. energy, correlation, inertia, entropy, inverse difference moment, sum average, sum variance, sum entropy, difference average, difference variance, difference entropy and two information measures of correlation) from the Spatial Grey Level Dependence Matrices (SGLD).

Each element \((i,j)\) of the SGLD matrices is calculated considering the probability that two pixels with a specified distance \(d\) have grey levels \(i\) and \(j\) along the direction \(\theta\). In this work, \(\theta = \{0,45^\circ,90^\circ,135^\circ\}\) and \(d=1\). In this paper, HAR is used as local descriptors.

3.4 Edge features (EDGE)

These features (ranging from SLF1.9 to SLF1.13 of the Subcellular Location Features descriptors of the Murphy’s research group) are calculated from the binary image of the cell \(I_{TH}\) (obtained by thresholding) and its edge image \(I_{ED}\) (obtained by the Canny method). In this paper, EDGE is used as global descriptors.

3.5 Wavelet features (WAVE)

We have used the Coiflets wavelet transform to extract a set of 30 features. These features include the average energy, square root of the sum of the square of each element of the three high-frequency components (i.e. horizontal, vertical and diagonal) calculated up to the \(10^{th}\) level decomposition. In this paper, WAVE is used as global descriptors.

4. Experimental Results

In this work, we train a random subspace, each of the 100 subspaces contain a subset of the 66% of the original features, of Levenberg-Marquardt neural networks (Dude et al., 2000) and the variant of the AdaBoost (named AdaBoost.ERC) used in (Lin et al., 2007). The training is performed using the 142 features obtained by the concatenation of the five descriptors reported in Section 3 (LBP, HAR, TAS, EDGE, WAVE) and the 180 features proposed from Murphy’s Lab for 2D HeLa experiments1. In this way, each method is trained using a set of 322 features.

The dataset used for evaluating the system proposed in this paper is the 2D HeLa dataset (Chebira et al., 2007). It is composed of 862 single-cell images; this dataset is available for download from http://murphylab.web.cmu.edu/. All the experimental results have been obtained using a 10-fold cross validation. Moreover, in order to deal with the inherent non-uniform lighting present in cell images, some of the descriptors are extracted from an enhanced version of the image, obtained by applying the contrast limited adaptive histogram equalization.

As a performance indicator, the Rank \(k\) recognition accuracy is used. It is the ratio of testing images for which the correct class is found in the top \(k\) ranked candidate classes. This indicator is also used in (Nanni et al., 2009b).

In Table 1 the Rank \(k\) recognition accuracies obtained by a Random subspace of LMNC (RS-LM), by an AdaBoost.ERC (A-ERC), and by the fusion with the sum rule (named FUS) between a Random subspace of LMNC and an AdaBoost.ERC are reported.

<table>
<thead>
<tr>
<th></th>
<th>RS-LM</th>
<th>A-ERC</th>
<th>FUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank-1</td>
<td>96.6%</td>
<td>96.5%</td>
<td>97.5%</td>
</tr>
<tr>
<td>Rank-2</td>
<td>99.2%</td>
<td>99.1%</td>
<td>99.4%</td>
</tr>
</tbody>
</table>

It is clear from Table 1 that the proposed ensemble outperforms several state-of-the-art methods as (Chebira et al., 2007; Hamilton et al., 2007; Nanni & Lumini, 2008; Nanni & Lumini, 2009), where the rank-1 is always lower than 96%. As a further experiment, we investigated the complementarities among the three best performing texture descriptors (i.e. LBP, TAS, HAR). For simplicity, we use only the random subspace of Levenberg-Marquardt neural networks. In Table 2 we report the error independence, considering the Yule’s Q-statistic (Kuncheva & Whitaker, 2003), between the classifiers trained using those features. The Q-statistic is defined as:

\[
Q = \frac{N_{11} N_{00} - N_{01} N_{10}}{N_{11} N_{00} + N_{01} N_{10}}
\]

where \(N_{ab}\) is the number of instances in the test set, classified correctly \((a=1)\) or incorrectly \((a=0)\) by the first classifier, and correctly \((b=1)\) or incorrectly \((b=0)\) by the latter classifier. \(Q\) varies between –1 and 1; \(Q=0\) for statistically independent classifiers. Classifiers that tend to recognize the same patterns correctly will have \(Q>0\), and those which commit errors on different patterns will have \(Q<0\). The results reported in Table 2 show that the different descriptors permit to train a set

---

1. http://murphylab.web.cmu.edu/data/2Dhela_images_download.html
of partially uncorrelated classifier. This result could motivate the importance to combine different texture descriptors.

Table 2. Yule’s Q-statistic between RSNN trained using different feature vectors.

<table>
<thead>
<tr>
<th>Feature Extraction</th>
<th>Yule’s Q-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP</td>
<td>TAS</td>
</tr>
<tr>
<td>LBP</td>
<td>HAR</td>
</tr>
<tr>
<td>TAS</td>
<td>HAR</td>
</tr>
</tbody>
</table>

As a final test, see Table 3, we show how it is important a deep study of the configuration of each descriptor to improve the performance. Now we test several methods for extracting features using the Haralick’s method:
- HAR1, which concatenates the features extracted by considering the two angles of 0° and 90°, with d=1, as proposed in (Nanni & Lumini, 2008b);
- HAR2, which adds the mean and the range of the 13 measures, thus increasing the the number of features to 26, as proposed in (Chebira et al., 2007);
- HAR3, which is the same as HAR1 but with four angles 0°, 45°, 135°, and 90°;
- HAR4, which is the same as HAR3 but with d=1 and d=3.

Table 3. Rank-1, obtained by the random subspace Levenberg-Marquardt neural networks, using different configurations of the Haralick feature extractor.

<table>
<thead>
<tr>
<th>HAR1</th>
<th>HAR2</th>
<th>HAR3</th>
<th>HAR4</th>
</tr>
</thead>
<tbody>
<tr>
<td>82.1</td>
<td>84.6</td>
<td>84.5</td>
<td>89.7</td>
</tr>
</tbody>
</table>

All the tests reported in this paper are performed using Matlab, the following toolboxes are used:
- LBP is available at http://www.ee.oulu.fi/mvg/page/source_code;
- Contrast-limited Adaptive Histogram Equalization is available as the function adapthisteq.m of the MATLAB;
- HARA is available at http://read.pudn.com/downloads114/sourcecode/graph/479658/Haralick.m.htm (Haralick Texture Features Matlab Toolbox v0.1b shalinig@ece.utexas.edu);
- software of several works of the Murphy’s research group is available from http://murphylab.web.cmu.edu/.

5. Conclusions

In this work we have shown that combining different descriptors and different classifiers permit to obtain a very reliable cell phenotype image classification method.

Two systems are combined: a random subspace of Levenberg-Marquardt neural networks; a variant of the AdaBoost, named AdaBoost.ERC. The most interesting result is that the proposed ensemble obtains an outstanding 97.5% Rank-1 accuracy and a >99% Rank-2 accuracy in the 2D-Hela dataset, which is the most used benchmark for comparing automated cell phenotype image classification approaches.

References


Nanni L, & Lumini A (2008), A reliable method for cell phenotype image classification, Artificial Intelligence in Medicine, 43 (2) Pages 87-97.


