On the importance of nucleus features in the classification of cervical cells in Pap smear images

Marina E. Plissiti and Christophoros Nikou [*]
Department of Computer Science, University of Ioannina, 45110 Ioannina, Greece
{marina,cnikou}@cs.uoi.gr

Abstract

In this work, we investigate the classification of cervical cells by exploiting only the nucleus features and not taking into account the features extracted from the cytoplasm area. This procedure is motivated by the fact that the nuclei areas can be extracted automatically from Pap smear images, in contrast to the cytoplasm segmentation which is not a solved problem yet. Furthermore, we consider the representation of these features in low dimensional spaces using non linear dimensionality reduction techniques, which can properly handle complex nonlinear data, as they better describe their manifold structure. The classification was performed in a supervised manner, using support vector machines (SVM) with several kernel functions. The obtained results indicate that we can achieve high classification performance when we use features extracted only from the nuclei areas.

1. Introduction

Normal and abnormal cells in Pap smear images are identified by evaluating changes in the density and morphology of the structural parts of the cells, which are the nucleus and the cytoplasm. More specifically, the nucleus is the structural part of the cell that presents significant changes when the cell is affected by a disease. These changes are identified through visual interpretation of the slide by an expert. However, this procedure is characterized as tedious and time-consuming and includes a considerable risk of misclassification, depending on the experience of the observer.

Thus, the characterization of the slide is mainly feasible by the evaluation of salient features of the nuclei of the cells, which, in general, is subject to the competence of the observer. In the last years, some efforts have been made in order to obtain reproducible and objective characterization of Pap smear images through computer vision methods, reducing the dependency on human experts. The automated classification of Pap smear images has been an interesting field for the researchers and many methods have been proposed, which involve both intelligent feature extraction techniques and machine learning algorithms, in order to recognize abnormalities in these images.

In this scope, several classification techniques have been proposed in order to recognize and classify a certain image/cell into the corresponding class. The first attempts to classify the cells in Pap smear images were based on the Bayes rule. Thus, in [17], the parametric Bhattacharyya distance is used for the determination of a pair of textural features and the Bayes classifier is then applied for the classification of the samples. The Bayesian classifier was also used in [3].

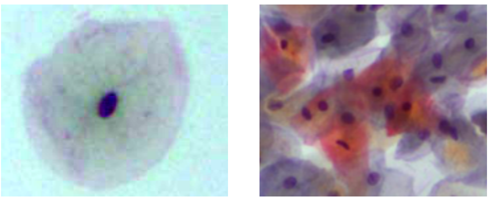
(a) (b)

Figure 1. Images of (a) single cell and (b) cell cluster. The area of the cytoplasm in (a) is well distinguished, in contrast to (b).

Another widespread technique used for the classification of Pap smear images are the artificial neural networks. In [6], a hierarchical hybrid multilayer perceptron network (H^2MLP) is illustrated for the classi-

[*] This work is co-financed by the European Union (European Regional Development Fund- ERDF) and Greek national funds through the Operational Program "THESSALY-MAINLAND GREECE AND EPIRUS-2007-2013" of the National Strategic Reference Framework (NSRF 2007-2013).

fication of cervical images into three categories. The Rank M-Type Radial Basis Function (RMRBF) neural network was implemented in [4] for the classification of microscopic Pap smear images. A feed forward neural network with a single hidden layer is used in [2] and the training method selected was backpropagation with a variable learning rate. A multilayer sigmoid neural network along with Levenberg-Marquardt backpropagation training algorithm was implemented in [10] for the classification of samples for which fuzzy based classification is unclear. Furthermore, some classification techniques based on fuzzy logic have also been proposed [8], [9]. Finally, support vector machines (SVM) were also used for the classification of Pap smear cells into normal and abnormal categories [3], [5]. It must be noted that most of these methods use presegmented images which contain only one cell, so the correct segmentation of the nucleus and the cytoplasm is feasible.

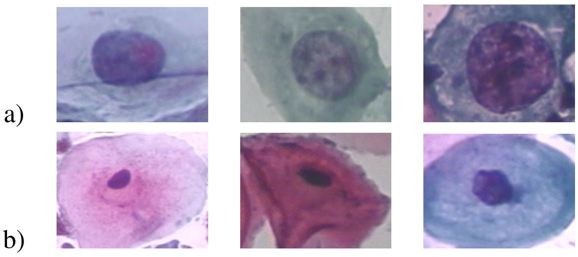

Figure 2. Types of cells included in the Pap smear benchmark [7]. (a) Abnormal cells and (b) normal cells.

The features that are usually evaluated by these methods are based on the intensity, the texture and the shape of the areas of the cytoplasm and the nucleus (Fig. 1(a)). However, these features exhibit different discriminative ability. Furthermore, it must be noted that in Pap smear images with high complexity and cell overlapping (Fig. 1(b)), the automated detection of the cytoplasm boundary is a difficult problem and until now, there is not any reported method or technique that successfully identifies the cytoplasm border of each cell in cell clusters. On the other hand, the automated segmentation of the nuclei areas in complex images containing extensive cell overlapping has been accomplished by several studies [12], [13].

Based on these facts, two open problems arise in the automated classification of a Pap smear image: a) the classification process fundamentally relies on the exploitation of the features extracted from the areas of the nuclei, since these areas can be automatically detected in an image acquired directly from an optical microscope, and b) the feature selection process must provide an efficient feature subset, which will exhibit the best discriminative ability for the recognition of normal and abnormal cells.

In this paper, we investigate both of the aforementioned problems. A detailed description of the classification of the cells in normal and abnormal categories using the nuclei features combined with supervised classification schemes is included in the following sections. Furthermore, the contribution of non linear dimensionality reduction schemes in the classification performance is examined. The work presented herein complements the study in [11], where only unsupervised classifiers (fuzzy C-means, spectral clustering) were examined in a similar framework.

2. Meterials and Methods

2.1 Study Group

In our experiments we have used the cell features included in the Pap-smear benchmark database presented in [7] . The database consists of 917 images containing a single cell each (Fig. 2). The detailed description of the database is depicted in Table 1.

2.2 Feature extraction and their projection in low dimensional spaces

The features of each cell that are available in the Pap smear benchmark database include nine common features for the nucleus and the cytoplasm areas and two more features that combine both areas. More specifically, the common features are the following: area, brightness, shorter diameter, longest diameter, elongation, roundness, perimeter, maxima, minima (the number of pixels with the maximum/minimum intensity value in a 3×3 neighborhood of the specific area). In terms of the features that are extracted from both the cytoplasm and the nucleus area, the nucleus position and the nucleus/cytoplasm (size) are calculated. Thus, nine out of twenty features concern the nucleus area and they can be calculated independently.

For the construction of a suitable feature subset with high discriminative ability, we have investigated the dimensionality reduction through non-linear techniques. These techniques are advantageous in comparison with their linear counterparts, because they can properly handle complex nonlinear data. In our study we have investigated the performance of four nonlinear techniques: Kernel-PCA (K-PCA) [15], Isomap [16], Locally Linear Embedding (LLE) [14] and Laplacian Eigenmaps [1].

Table 1. Cell categories in the Pap-smear benchmark database [7].

NORMAL	#cells
Superficial squamous epithelial	74
Intermediate squamous epithelial	70
Columnar epithelial	98
TOTAL	**242**
ABNORMAL	#cells
Mild squamous non-keratinizing dysplasia	182
Moderate squamous non-keratinizing dysplasia	146
Severe squamous non-keratinizing dysplasia	197
Squamous cell carcinoma in situ intermediate	150
TOTAL	**675**

3. Results and Discussion

In our experiments, we have investigated the effectiveness of the above dimensionality reduction schemes in the classification performance obtained in a supervised manner. For this reason, the SVM classifier was used and several kernel functions were tested. More specifically, we have used the linear, the polynomial and the radial basis function (RBF) kernels. For the evaluation of the classification performance, the harmonic mean (H-mean) of the *sensitivity* and the *specificity* indices was calculated. The sensitivity measures the proportion of abnormal cells which are correctly identified as such by the classification algorithm, and the specificity measures the proportion of the normal cells that are correctly characterized as such.

A training data set is constructed randomly from the entire data set and the remaining patterns were used as test set. The performance of the classification is calculated using the trained SVM classifier in the test set and the final performance is obtained by the mean of the results, after the execution of this experiment 10 times, in a 10-fold cross validation scheme.

In order to estimate the discriminative ability of the features we have implemented two general experiments. More specifically, the SVM classifier was trained using patterns from two different feature sets: one containing both cytoplasm and nucleus features (20 features) and the other containing only nucleus features (9 features). Several experiments were performed and the performance of the classification techniques was measured using patterns of increasing dimension varying from 1 to 20 features for the first subset and from 1 to 9 for the second subset. The best results for each classifier are presented in this work.

The parameters used in each experiment were obtained after several tests. In K-PCA, the Gaussian and the polynomial kernel were used. In Isomap, LLE and Laplacian Eigenmaps, different numbers of nearest neighbors ranging from 4 to 20 were also tested for the construction of the distance graph. In SVM, the tolerance threshold was set to 0.001 for all the kernel functions, for the polynomial kernel the degree was set to 3 and for the RBF kernel the kernel width was set to 1.

The classification results are depicted in Table 2 in terms of the H-mean measure. As we can observe, the initial feature set without the use of dimensionality reduction schemes, leads to the weakest classification performance. Furthermore, the use of all the features, including those of the cytoplasm areas, exhibits better classification results. However, the use of non-linear dimensionality reduction schemes not only leads to a significant improvement of the classification but it achieves better classification results using only the features extracted from the nucleus area.

More specifically, in Table 2, we can observe that the SVM with RBF kernel leads to the higher classification results for both feature sets, reaching 96.89% for the nuclei features, and 96.30% for the cytoplasm and nuclei features, when the K-PCA with Gaussian kernel is used. The polynomial SVM reaches the best classification results using the nuclei features in combination with K-PCA with polynomial kernel (91.68%), which is 3.14% higher than the best classification results using all the features (88.54%). Finally, the linear SVM exhibits the lowest performance in comparison with the other versions of the classifier. However, even in this case, the use of the nuclei features results in better classification performances. The results of the supervised training presented herein are in general more accurate compared with the corresponding results of unsupervised clustering techniques [11], where, indicatively, the best performance was 90.42% (provided by K-PCA). Therefore, it may be concluded that the use of non-linear dimensionality reduction schemes, improves the classification performance and achieves the successful separation of normal and abnormal cercival cells, by exploiting only the nuclei features.

4. Conclusions

In complex Pap smear images the only areas that can be extracted automatically are the nuclei [12], [13], in contrast to the cytoplasm. From the nuclei areas we can calculate salient nuclei features that contribute in the correct classification of normal and abnormal cells. In this study, we have investigated the performance of supervised classifiers, when they are trained with a feature set based exclusively on the features extracted from

Table 2. Performance of SVM classification in terms of H-mean (%).

Kernel Function	Linear		Polynomial		RBF	
Feature sets	All	Nuclei	All	Nuclei	All	Nuclei
No dimensionality reduction	84.52	83.91	86.73	84.98	87.12	86.60
K-PCA (polynomial)	84.75	**86.99**	87.92	**91.68**	87.14	91.25
K-PCA (Gaussian)	84.70	85.85	88.14	91.37	**96.30**	**96.89**
Isomap	84.87	86.42	86.82	89.11	88.20	89.29
LLE	**85.00**	85.96	**88.54**	91.49	87.23	91.25
Laplacian Eigenmaps	84.64	85.56	86.77	87.29	87.75	91.92

the nuclei areas, and we have compared the results with the corresponding results obtained with the use of both the cytoplasm and nuclei features in the training procedure. As it was deduced by our experiments, the use of non-linear dimensionality reduction techniques has a positive effect in the correct classification of the cells, when only the nuclei features are used. The above results have significant importance, since they imply that we can achieve remarkable classification performance with the use of the nuclei features alone.

References

[1] M. Belikn and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[2] A. N. Bondarenko and A. V. Katsuk. Extracting feature vectors of biomedical images. In *Proceedings of the 9th Russian-Korean International Symposium on Science and Technology*, pages 579 – 583, 2005.

[3] T. Chankong, N. Theera-Umpon, and S. Auephanwiriyakul. Cervical cell classification using Fourier transform. In *Proceedings of 13th International Conference on Biomedical Engineering*, volume 23, pages 476–480, 2009.

[4] F. J. Gallegos-Funes, M. E. Gomez-Mayorga, J. L. Lopez-Bonilla, and R. Cruz-Santiago. Rank M-type radial basis function (RMRBF) neural network for Pap smear microscopic image classification. *Apeiron*, 16(4):542–554, 2009.

[5] P. C. Huang, Y. K. Chan, P. C. Chan, Y. F. Chen, R. C. Chen, and Y. R. Huang. Quantitative assessment of Pap smear cells by PC-based cytopathologic image analysis system and support vector machine. In *Proceedings of the 1st International Conference on Medical Biometrics*, 2007.

[6] N. A. M. Isa, M. Y. Mashor, and N. H. Othman. An automated cervical pre-cancerous diagnostic system. *Artificial Intelligence in Medicine*, 42:1–11, 2008.

[7] J. Jantzen, J. Norup, G. Dounias, and B. Bjerregaard. Pap-smear benchmark data for pattern classification. In *Proceedings of Nature inspired Smart Information Systems (NiSIS)*, pages 1–9, 2005.

[8] K. B. Kim, S. Kim, and K. B. Sim. Nucleus classification and recognition of uterine cervical Pap-smears using fuzzy art algorithm. In *Proceedings of 6th International Conference on Simulated Evolution and Learning, Lecture Notes in Computer Science*, volume 4247, pages 560–567, 2006.

[9] K. B. Kim, D. H. Song, and Y. W. Woo. Nucleus segmentation and recognition of uterine cervical Pap-smears. In *Proceedings of the 11th RSFDGrC 2007, Lecture Notes in Computer Science*, volume 4482, pages 153–160, 2007.

[10] Z. Li and K. Najarian. Automated classification of Pap smear tests using neural networks. In *Proceedings of the International Joint Conference on Neural Networks*, volume 4, pages 2899–2901, 2001.

[11] M. E. Plissiti and C. Nikou. Cervical cell classification based exclusively on nucleus features. In *Proceedings of the 9th International Conference on Image Analysis and Recognition, Lecture Notes in Computer Science*, volume 7325, pages 483–490, 2012.

[12] M. E. Plissiti, C. Nikou, and A. Charchanti. Automated detection of cell nuclei in Pap smear images using morphological reconstruction and clustering. *IEEE Transactions on Information Technology in Biomedicine*, 15(2):233–241, 2011.

[13] M. E. Plissiti, C. Nikou, and A. Charchanti. Combining shape, texture and intensity features for cell nuclei extraction in pap smear images. *Pattern Recognition Letters*, 32(6):838–853, 2011.

[14] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[15] B. Scholkopf, A. Smola, and K. R. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[16] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[17] R. F. Walker, P. Jackway, B. Lovell, and I. D. Longstaff. Classification of cervical cell nuclei using morphological segmentation and textural feature extraction. In *Proceedings of the 2nd Australian and New Zealand Conference on Intelligent Information Systems*, pages 297 – 301, 1994.