32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

Using PCA and LVQ Neural Network for Automatic Recognition of Five Types of White Blood Cells

P. R. Tabrizi, S. H. Rezatofighi, M. J. Yazdanpanah

Abstract—**Designing an effective classifier has been a challenging task in the previous methods proposed in the literature. In this paper, we apply a combination of feature selection algorithm and neural network classifier in order to recognize five types of white blood cells in the peripheral blood. For this purpose, first nucleus and cytoplasm are segmented using Gram-Schmidt method and snake algorithm, respectively; second, three kinds of features are extracted from the segmented areas. Then the best features are selected using Principal Component Analysis (PCA). Finally, five types of white blood cells are classified using Learning Vector Quantization (LVQ) neural network. The performance analysis of the proposed algorithm is validated by an expert's classification results. The efficiency of the proposed algorithm is highlighted by comparing our results with those reported in a recent article which proposed a method based on the combination of Sequential Forward Selection (SFS) as the feature selection algorithm and Support Vector Machines (SVM) as the classifier.**

I. Introduction

Recognition and inspection of white blood cells in peripheral blood can assist hematologists in diagnosing many diseases such as AIDS, Leukemia, and blood cancer [1]. Thus, this process is assumed as one of the most prominent steps in the hematological procedure.

Albeit not extensive, some methods have been proposed in the literature for this purpose. For example, in [1], the authors have suggested a system to classify five major groups of white blood cells in peripheral blood categorized into basophil, neutrophil, eosinophil, monocyte, and lymphocyte (Fig. 1). This system has three major parts: image segmentation, feature extraction, and classification. The feature selection step using Sequential Forward Selection (SFS) algorithm has been adjoined to the feature extraction process for ameliorating the classifier performance and accelerating the program trend. Moreover, in the classification step, the performance of two different classifiers, Support Vector Machines (SVM) and Multi-Layer Perceptron (MLP) has been compared, and the superiority of SVM over MLP has been shown. Also, in other papers, different types of artificial neural networks (ANNs) such as feed-forward back-propagation [2], [3], local linear map [4], and fuzzy cellular neural network [5] have been used. It can be construed that classifiers and feature selection algorithms have a crucial role in the performance of a system.

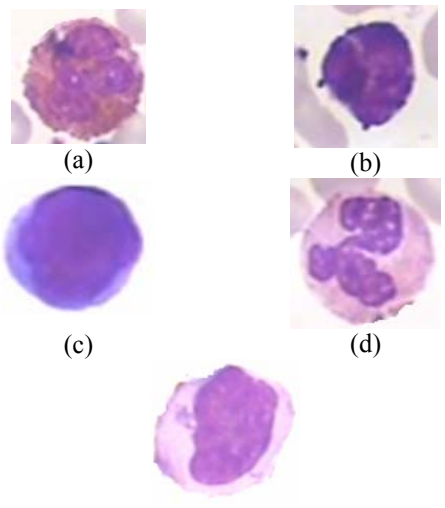


Fig. 1. Five major groups of white blood cells in peripheral blood. (a) eosinophil, (b) basophil, (c) lymphocyte, (d) neutrophil, and (e) monocyte.

In this paper, the main purpose is to apply a combination of feature selection algorithm and neural network classifier in order to improve the performance of the white blood cell recognition system in comparison with [1] along with reducing the dimension of the features. Therefore, at first, nucleus and cytoplasm are segmented, and three kinds of features are extracted using proposed system in [1]. Then the dimension of features is reduced using Principal Component Analysis (PCA) [6]. Finally, five types of white blood cells are classified using Learning Vector Quantization (LVQ) neural network [7], [8].

The rest of the paper is organized as follows. In Section II, we propose our system structure for recognition of five types of white blood cells. The experimental results are presented and discussed in Section III. Finally, Section IV is appropriated to presentation of the conclusions.

Manuscript received April 23, 2010.

P. R. Tabrizi is with the Control and Intelligent Processing Center of Excellence (CIPCE), school of Electrical and Computer Engineering, University of Tehran, P.O. Box 14395/515, Tehran, Iran (e-mail: p.roshani@ece.ut.ac.ir).

S. H. Rezatofighi is with the Control and Intelligent Processing Center of Excellence (CIPCE), school of Electrical and Computer Engineering, University of Tehran, P.O. Box 14395/515, Tehran, Iran (e-mail: h.tofighi@ece.ut.ac.ir).

Prof. M. J. Yazdanpanah is with the Control and Intelligent Processing Center of Excellence (CIPCE), school of Electrical and Computer Engineering, University of Tehran, P.O. Box 14395/515, Tehran, Iran (e-mail: yazdan@ut.ac.ir).

II. System Architecture

Fig. 2 illustrates the block diagram of our system. As shown in this figure, the system has three major parts whose details are explained in the next sections.

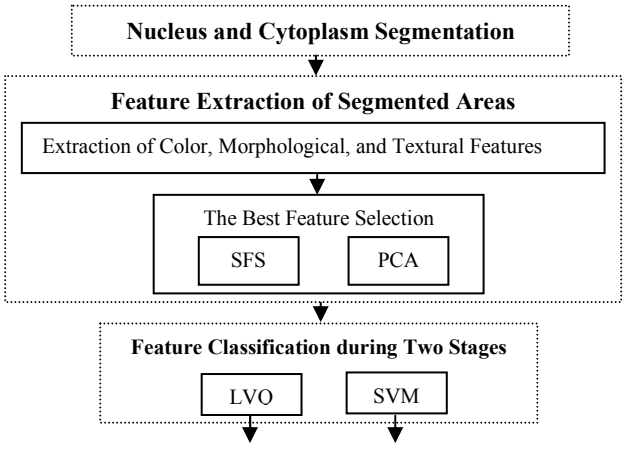


Fig. 2. The block diagram of the proposed system.

A. Nucleus and Cytoplasm Segmentation

In this phase, we apply the algorithm proposed in [1] to segment nucleus and cytoplasm. The Gram-Schmidt method is used to segment nucleus. Using Gram-Schmidt method, a composite image with higher intensity in the nucleus area compared to other areas is obtained. Next, by choosing an appropriate threshold based on the histogram information, the nucleus area is segmented. Moreover, the snake algorithm is applied to segment cytoplasm of the white blood cells. To this end, image size reduction, preprocessing before the snake algorithm, and finding an initial contour for the snake algorithm are applied. The details of the proposed algorithm have been described in [1].

B. Feature Extraction of Segmented Areas

In this phase, we use the best features introduced in [1]. These features are categorized into three groups of morphological, textural, and color features. The morphological features are cytoplasm area and whole cell body perimeter, nucleus area and perimeter, number of the separated parts of nucleus, mean and variance of the cytoplasm and nucleus boundaries, roundness criterion of nucleus and the whole cell, and the ratio between the cytoplasm and nucleus areas. Textural features are also extracted from the nucleus and cytoplasm area by the local binary pattern. At the end, a normalized vector of the average cytoplasm and nucleus color is extracted as the color features.

After feature extraction of nucleus and cytoplasm area, we must apply a feature selection algorithm for improving the classifier performance. For this purpose, PCA and SFS [1] algorithm are applied separately, and their results are compared.

C. Feature Classification during Two Stages

Since in most of the samples, the boundary between nucleus and cytoplasm of basophils cannot be distinguished visually; these cells should not involve in segmentation of cytoplasm's step. Therefore, first they should be recognized from the other samples using features of nucleus area. Next, remaining classes should be classified using features extracted from the cytoplasm area in combination with the features extracted from the nucleus area.

To classify white blood cells, the performances of two classifiers, SVM [1] and LVQ, are evaluated. Learning vector quantization is a nearest-neighbor pattern classifier based on competitive learning [8]. A LVQ network contains an input layer, a Kohonen layer which learns and performs the classification, and an output layer. The input layer contains one node for each input feature; the output layer contains one node for each class. Fig.3 illustrates the structure of the LVQ neural network. In this figure, R, S^1, and S^2 are number of elements in input vector, number of competitive neurons, and number of linear neurons, respectively.

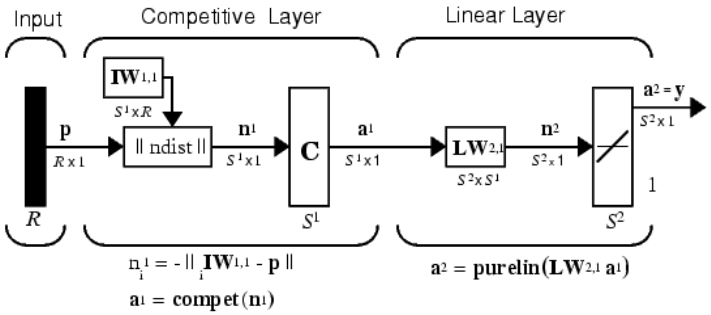


Fig. 3. LVQ structure [8].

In this paper, we apply LVQ3 algorithm, the advanced version of Kohonen's LVQ algorithm, to learn LVQ network, and use original data to initialize reference vectors in competitive layer. A brief description of the algorithm is given in the following paragraph.

For each input sample, the two closet weight vectors, w_1 and w_2, are first found by using the Euclidean distance criterion. The requirement for updating these vectors is that the distances of two closest weight vectors, d_1 and d_2, satisfy the window condition, as (1). If one of the two closest vectors, w_1, belongs to the same class as the input vector x, and the other vector, w_2, belongs to a different class, the weights update as (2) and (3) using the learning rate $\alpha(t)$, respectively. Moreover, if x, w_1, and w_2 belong to the same class, the weights update as (4).

$$\begin{cases} \min[\frac{d_1}{d_2}, \frac{d_2}{d_1}] > (1-\varepsilon)(1+\varepsilon) \\ \varepsilon : Const. \end{cases} \quad (1)$$

$$w(t+1) = w(t) + \alpha(t)[x(t) - w(t)], \quad (2)$$

$$w(t+1) = w(t) - \alpha(t)[x(t) - w(t)], \quad (3)$$

$$\begin{cases} w(t+1) = w(t) + \beta(t)[x(t) - w(t)], \\ \beta(t) = m\alpha(t), \\ 0.1 \le m \le 0.5. \end{cases} \quad (4)$$

III. EXPERIMENTAL RESULTS

The proposed system was tested on 251 blood smear slide images containing 302 white blood cells. These images were acquired by light microscope from stained peripheral blood using the Digital Camera-Sony-Model No. SSC-DC50AP with magnification of 100. The resolution of the images is 720*576 pixels. The digital images were classified by a hematologist into the normal leukocytes: basophil, eosinophil, lymphocyte, monocyte, and neutrophil.

In this paper, the performance of two feature selection algorithms, SFS and PCA, and two classifiers, SVM and LVQ are compared. This comparison is carried out during two stages, as introduced in [1], which the basophils are discriminated from the other types of white blood cells in first stage and the four remaining classes are recognized in second stage. To assess the performance of our proposed system in recognizing the white blood cells, an accuracy criterion is used. We use approximately 30% of dataset at the first stage and 50% of dataset at the second stage as training data. Fig. 4 (a) and (b) illustrate the classification results in different dimensions related to the basophil and the four remaining groups of white blood cells classification, respectively. Some points are construed from these figures:

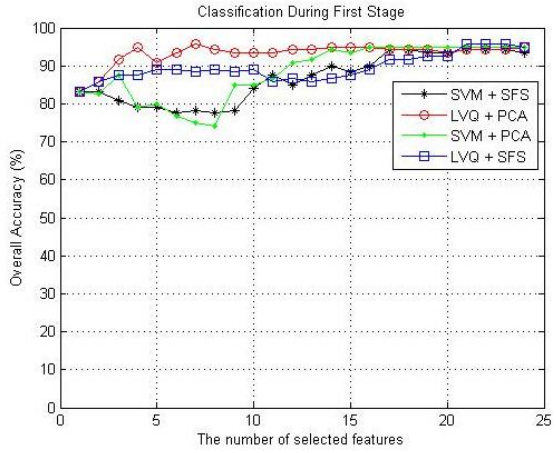
The first point is that the PCA has a better performance in most of the feature dimensions than the SFS when the LVQ is used as classifier. Also, the combination of SVM and PCA has a better performance than the combination of SVM and SFS when feature dimension is increased; because PCA can reduce the number of features by discarding the ones which have small variance and retains only those terms that have large variance using a linear transformation matrix, whereas SFS does not apply a transformation matrix, and selects the best features based on Fisher's discriminant ratio (FDR) [1]. The objective of the FDR is to select the features not only by maximizing the between-class scatter of data, but also by minimizing the within-class scatter. As a result, the FDR has a good performance in fewer dimensions of features when the SVM is used as the classifier, and is not an optimal criterion for the SFS when the LVQ is used as the classifier; accordingly, in the rest of the paper, we only consider the features obtained from the PCA when the LVQ is utilized as the classifier.

The second point is that reduction in dimension of features aggravates the classification rates as expected. However, increasing dimension of features over a threshold may reduce the generalization ability of a classifier. Besides, systems are more practical in fewer dimensions of features. Therefore, an optimal dimension of features is required. The optimal dimensions of features for the combination of LVQ and PCA are 7 and 5 for the basophil and four remaining groups of white blood cells classification, respectively. In contrast, the optimal dimensions of features reported in [1], were 15 and 10 for the basophil and four remaining groups of white blood cells classification, respectively. These selections are because after selecting these features, the differentiation between the features of each class is at a maximum, and the overall accuracy does not have considerable escalation for both LVQ and SVM [1] classifiers. Thus, the combination of PCA and LVQ has a better performance in selected feature dimensions than those reported in [1].

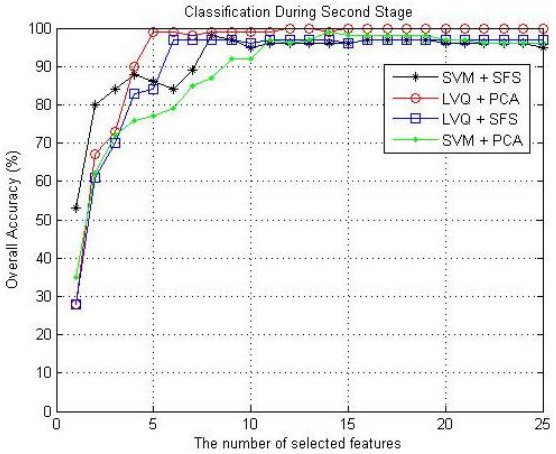
The third one is that the SVM classifier has more oscillation in the overall accuracy in comparison with the LVQ. These changes are because the SVM is not trained well as a result of its Gaussian Kernel function [1]. In fact, SVM needs diverse Kernel functions in different dimensions of feature. Obviously, a fixed Kernel function is used for SVM in [1]. Therefore, SVM [1] has more oscillation than LVQ. Totally, SVM is very sensitive to type of Kernel function applied to separate features in a space which has high dimension. The best Kernel function must be selected by trial and error experiments. Similarly, LVQ depends on initial weights. In this paper, we choose its initial weights using original data; indeed the search for the best selection may be time consuming when the number of classes and data is increased. But due to the confidence of the best selection in training, it may be preferred.

The last point is that the combination of PCA and LVQ has better performance in comparison with the combination of SFS and SVM. The main reason for superiority of the combination of PCA and LVQ is that LVQ and SVM have different structures. SMV is a classifier which classifies data using a supervised algorithm and a nonlinear transformation. In contrast, LVQ classifies data into subclasses and classes during two stages of classification using a supervised and an unsupervised algorithm; consequently LVQ can have better exploitation of data structure by considering the existence of clusters within each class.

According to the above conclusion, selecting 7 and 5 features during two stages of classification and utilizing PCA and LVQ is the best way to have optimal performance, especially in lower dimension of features which is the main purpose of this research to achieve lower processing time. Tables 1-4 show confusion matrix, accuracy, and overall accuracy for features classified using the combination of PCA and LVQ or SFS and SVM [1] during two stages of classification.



(a)



(b)

Fig. 4. The overall accuracy results for the combination of PCA and SFS [1] with SVM [1] and LVQ in order to classify (a) basophil, (b) four remaining groups of white blood cells

TABLE 1. CONFUSION MATRIX, ACCURACY, AND OVERALL ACCURACY OF PCA AND LVQ CLASSIFIER WHEN 7 FEATURES ARE SELECTED.

	Recognized Basophil	Recognized Non-Basophil	Accuracy
Basophil	43	7	86%
Non-Basophil	4	169	97.69%
Overall Accuracy			95.06%

TABLE 2. CONFUSION MATRIX, ACCURACY, AND OVERALL ACCURACY OF SFS AND SVM CLASSIFIER WHEN 15 FEATURES ARE SELECTED [1].

	Recognized Basophil	Recognized Non-Basophil	Accuracy
Basophil	50	0	100%
Non-Basophil	23	150	86.71%
Overall Accuracy			89.69%

TABLE 3. CONFUSION MATRIX, ACCURACY, AND OVERALL ACCURACY OF PCA AND LVQ CLASSIFIER WHEN 5 FEATURES ARE SELECTED.

	Recognized Eosinophil	Recognized Lymphocyte	Recognized Monocyte	Recognized Neutrophil	Accuracy
Eosinophil	19	0	0	0	100%
Lymphocyte	0	29	0	0	100%
Monocyte	0	1	23	0	95.83%
Neutrophil	0	0	0	28	100%
Overall Accuracy					99%

TABLE 4. CONFUSION MATRIX, ACCURACY, AND OVERALL ACCURACY OF SFS AND SVM CLASSIFIER WHEN 10 FEATURES ARE SELECTED [1].

	Recognized Eosinophil	Recognized Lymphocyte	Recognized Monocyte	Recognized Neutrophil	Accuracy
Eosinophil	19	0	0	0	100%
Lymphocyte	0	27	2	0	93.1%
Monocyte	0	0	23	1	95.83%
Neutrophil	1	0	0	27	96.43%
Overall Accuracy					96%

According to the above tables, the overall accuracy with the combination of PCA and LVQ are superior to those of the combination of SFS and SVM, while the dimension of features is lower than the method introduced in [1].

IV. CONCLUSION

In this paper, we utilized the combination of PCA and LVQ classifier in order to recognize five groups of white blood cell in the peripheral blood. Although there is no significant difference between the proposed method in this paper and [1], we could achieve better performance in lower dimension of features using the combination of an appropriate feature selection algorithm and a better classifier. Hence, our method has less running time, and is more practical in order to execute in hematological laboratories.

Future work will be addressed to apply wavelet neural network to increase our accuracy rate.

REFERENCES

[1] S. H. Rezatofighi, K. Khaksari, and H. Soltanian-Zadeh, "Automatic recognition of five types of white blood cells in peripheral blood," in *Proc. International Conf. Image Analysis and Recognition (ICIAR 2010))*, Povoa de Varzim, Portugal, 2010, to be published.

[2] N. T. Umpon, and P. D. Gader, "System-Level training of neural networks for counting white blood cells," *IEEE Trans. Syst., Man, Cybern*, vol. 32, no. 1, pp. 48-53, 2002.

[3] X. Long, W. L. Cleveland, and Y. L. Yao, "A new preprocessing approach for cell recognition," *IEEE Trans. Inf. Technol. Biomed*, vol. 9, no. 3, pp. 407-412, 2005.

[4] T. W. Nattkemper, H. J. Ritter, and W. Schubert, "A neural classifier enabling high-throughput topological analysis of lymphocytes in tissue sections," *IEEE Trans. Inf. Technol. Biomed*, vol. 5, no. 2, pp. 138-149, 2001.

[5] W. Shitong, and W. Min, "A new detection algorithm (NDA) based on fuzzy cellular neural networks for white blood cell detection," *IEEE Trans Inf. Technol. Biomed*, vol. 10, no. 1, pp. 5-10, 2006.

[6] K. I. Diamantaras, *Neural Networks and Principal Component Analysis*, CRC Press Inc., 2002, ch.8.

[7] L. V. Fausett, *Fundamentals of Neural Networks: Architectures, Algorithms and Applications*, Prentice Hall, 1993, pp. 187-195.

[8] B. H. Demuth, M. Beale, and M. T. Hagan, *Neural Network Design*, PWS Publication, 1996, pp. 14.16-14.21.