ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Data mining techniques for feature selection in blood cell recognition

T. Markiewicz[1] and S. Osowski[1,2]

1- Warsaw University of Technology, Warsaw, pl. Politechniki 1, Poland
2- Military University of Technology, Warsaw, ul. Kaliskiego 3, Poland

Abstract. The paper presents and compares the data mining techniques for selection of the diagnostic features in the problem of blood cell recognition in leukemia. Different techniques are compared, including the linear SVM ranking, correlation and statistical analysis of centers and variances of clusters corresponding to different classes. We have applied radial kernel SVM as the classifier. The results of recognition of 10 classes of cells are presented and discussed.

1 Introduction

The recognition of the blast cells in the bone marrow of the patients suffering from leukemia is a very important step in the recognition of the development stage of the illness and proper treatment of the patients [2]. There are different cell lines in the bone marrow: the megacaryocytic, erythrocytic, monocytic, lymphocytic and granulocytic. To the most known and recognized cells belong: monoblasts, promonocytes, monocytes, myeloblasts, promyelocytes, myelocyte, metamyelocytes, proerythroblasts, basophilic erythroblast, polychromatic erythroblast, pyknotic erythroblast [2,3]. They differ by the size, texture, shape, density and color. Fig. 1 presents the typical image of the bone marrow containing different types of blood cells (most of them are myeloblasts and erythroblasts).

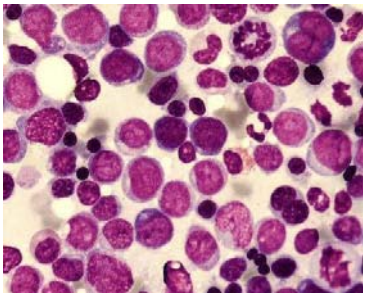

Fig. 1 The typical image of the bone marrow

The process of automatic recognition requires the extraction of individual blood cells, generation of diagnostic features and finally the recognition using chosen classifier. The applied automatic blood cell recognizing system is presented in Fig. 2. The bone marrow image is digitized using digital camera. The next step is the extraction of individual cells from the image. For each cell the feature generation and selection are

performed. The selected features form the vector **x** applied to the input of Support Vector Machine (SVM) network working as the recognizing and classifying system.

Fig. 2 The proposed automatic blood cell recognition system

Up to now no automatic system exists that could recognize and count the blood cells with the accuracy comparable to the human expert. Although some attempts to solve this problem have been presented [3,7] the results are still not satisfactory and a lot of research should be done to achieve the efficiency of the human expert. The most important problem is generation of the features of the blood cells that characterize them in a way enabling the recognition of different blast types with the highest accuracy. This task needs highly efficient feature selection techniques.

The paper studies different methods of feature selection: the linear SVM ranking, correlation between features and the class, statistical analysis of the clusters corresponding to different classes. The results of recognition of 10 blast cell types by using Gaussian kernel SVM are given and discussed. We have chosen SVM after many introductory experiments with other types of classifiers (MLP, neuro-fuzzy, hybrid, etc) since this network has provided the best accuracy of recognition.

2 Feature generation

The first step of blood cell processing is the extraction of individual cells from the image of the bone marrow. We have done it by segmentation using morphological operations of watershed transformation [6]. Recognition of the blood cell on the basis of its image needs generation of the numerical features well describing the differences of images belonging to different classes. In characterizing the images by the numerical values we try to get the features strictly corresponding to these on the basis of which the human expert makes his diagnosis, that is the geometry of cell, texture, color and intensity of the image associated with different cell types. Four families of features have been created in this way [3]. The geometrical features include such parameters as radius, perimeter, area, the area of convex part of the cell, compactness, concavity, symmetry, major and minor axis lengths, etc. These parameters are determined only for the nucleus of the cell. Up to 19 geometrical features have been generated for each cell on the basis of these parameters [3].

The texture refers to an arrangement of the basic constituents of the material and in the digital image is depicted by the interrelationships between spatial arrangements of the image pixels [8]. After some preliminary experiments, we have chosen two texture preprocessing methods, due to Unser and Markov random field [8]. Up to 105 texture features have been generated for the cell image at normal and reduced resolutions.

The next set of features has been generated from the analysis of the intensity distribution of the image. The histograms of the image and gradient matrix of such

intensity have been determined for R, G, B components of the image. On the basis of such analysis we have generated the following features: the mean and variance of the histogram of the image of nucleus and cytoplasm (separately) as well as for the gradient matrix of the image, the skewness and kurtosis of the image of the whole cell as well as for the gradient matrix of the whole cell. Up to 24 statistical features have been generated in this way for two colours (red and green).

The last set of features is related to the morphological operations performed on the images (erosion, dilation, opening and closing). These parameters include the area and number of separated objects of the image after application of some morphological operations. Up to 16 morphological parameters have been generated in this way. All features have been normalized, dividing their original values by the corresponding maxima.

3 Feature selection techniques

To get the best results of recognition we have to apply the proper set of features. There are many techniques of feature selection (Schurmann, 1996, Guyon, et al., 2003). To the most popular belong principal component analysis, projection pursuit, correlation existing among features, correlation between the features and the classes, analysis of mean and variance of the features belonging to different classes, application of linear SVM feature ranking, etc. In this paper we have applied and compared some of them, including two linear SVM ranking methods, correlation analysis and the statistical analysis of clusters corresponding to the different classes.

3.1 The selection based on the mean and variance of the data

The most often used criterion of feature selection is the analysis of variance and means of the data samples belonging to each class [1,5]. The variance of the feature describing the cells, belonging to one class should be as small as possible. Moreover to distinguish between different classes, the positions of means of feature values for the data belonging to different classes should be separated as much as possible. However the particular feature may be very good for recognition between two chosen classes and useless for some others. Therefore the class oriented features should be considered to get the optimal choice of them. The multiclass problem should be solved by separating the task into two-class recognition sub-problems, as it is done in SVM classifiers (one against one mode of operation).

The systematic policy for feature selection is to combine the variance and mean together to form single quality measure [1,5]. He have done it by defining so called discrimination coefficient $S_{AB}(f)$. For two classes A and B the discrimination coefficient of the feature f was defined as follows

$$S_{AB}(f) = \frac{|c_A(f) - c_B(f)|}{\sigma_A(f) + \sigma_A(f)} \qquad (1)$$

In this definition c_A and c_B are the mean values of the feature f in the class A and B, respectively. The variables σ_A and σ_B represent the standard deviations determined for both classes. The large value of $S_{AB}(f)$ indicates good separation ability of the feature f for these two classes.

3.2 **The linear SVM methods**

The interesting method of feature selection is the application of the linear Support Vector Machine [1]. The simplest way of application of SVM linear network for feature selection is training the network using only one feature. The predictive power of the single feature for a classification task is characterized by the value of error function minimized by a one-dimensional linear SVM trained to classify learning samples on the basis of only one feature of interest. The smaller this error the better is the quality of the feature.

The ranking of the features may be also done for all features working together. The method is based on the idea, that the absolute values of the weights of a linear classifier trained on the whole set of features produces a feature ranking [1]. The feature associated with the larger weight is more important than that associated with the small one. We have used here the linear kernel SVM classifier working in one against one mode. All values of weights have been arranged in decreasing order and only the most important have been selected for each pair of classes and then used in the final classification system.

3.3 **Correlation analysis**

The discriminative power of the candidate feature *f* for the recognition of the particular class among *K* classes can be also measured by the correlation of this feature with the class [5]. Let us assume that the target class *k* is one among the classes forming target vector **d.** Let us assume that the feature *f* is described by its unconditional and conditional means $m_c = E\{f\}$ and $m_{ck} = E\{f|k\}$. Assume that the variance *var*(*f*) of feature *f* is known. The correlation between *f* and **d** is derived from the covariance vector **cov**(*f*,**d**), related by the respective variance. The discriminative power of feature *f* is measured as the squared magnitude of the vector **corr**(*f*,**d**), i.e., $S(f) = |\mathbf{corr}(f,\mathbf{d})|^2 = |\mathbf{cov}(f,\mathbf{d})|^2 / \mathrm{var}(f)\,\mathrm{var}(\mathbf{d})$. Denoting by P_k the probability of kth class and taking into account that $\mathrm{var}(\mathbf{d}) = \sum_{k=1}^{K} P_k(1-P_k)$ we get $\mathbf{cov}(f,\mathbf{d}) = [P_1(m_{c1}-m_c),...,P_K(m_{cK}-m_c)]^T$. The discriminative power of feature *f* is then defined in the form [5]

$$S(f) = \frac{\sum_{k=1}^{K} P_k (m_{ck} - m_c)^2}{\mathrm{var}(f)\sum_{k=1}^{K} P_k(1-P_k)} \qquad (2)$$

Using this measure we can arrange the features in decreasing order from the highest to the smallest discriminative value.

3.4 **Collective evaluation of feature sets**

After ranking the features we get them arranged in the decreasing order, from the best to the least significant. However the open question remains: what is the size of the optimal set of these ordered features? We need a rough but quick assessment of the discriminative power of the whole set of features on the basis of the learning data. Let

us denote by V_k^2 the mean-squared within-class distance between samples of the class k and the corresponding mean $\mathbf{m}_k$, $V_k^2 = \mathrm{var}\{f|k\}$ and with V^2 the mean of these within-class squared distances, $V^2 = \sum_{k=1}^{K} P_k V_k^2$. Denote by D^2 the mean-squared between-class distance between different classes centres $\mathbf{m}_k$ and $\mathbf{m}_j$. At probabilities P_k and P_j of corresponding kth and jth classes we get [5]

$$D^2 = \frac{1}{(1-\sum_{k=1}^{K} P_k^2)} \sum_{k=1}^{K} \sum_{j=1}^{K} P_k P_j \left| \mathbf{m}_k - \mathbf{m}_j \right|^2 \qquad (3)$$

The measure of separability of the whole set of features can be derived as $S = D^2/(D^2+V^2)$. It ranges from $S \to 0$ (lack of separability) to $S \to 1$ (optimal separability).

4 Results of numerical experiments

The numerical experiments of recognition of 10 classes of blood cells typical for leukemia have been performed using SVM of Gaussian kernel and one against one mode of operation. These classes include: basophilic erythroblast (1), polychromatic erythroblast (2), pyknotic erythroblast (3), mono- and myelo-blasts called usually blasts (4), promyelocyte (5), myelocyte (6), metamyelocyte (7), neutrophilic band (8), neutrophilic segmented (9) and lymphocyte (10). The problem is really difficult, because of large number of recognized classes and also of close similarity of the representatives of the cells belonging to different classes. At the same time there is a large variation among cells belonging to the same family of cells.

The number of considered candidate features, generated according to the procedure presented in section 2, was equal 164. To obtain the best set of features we have applied 4 described above methods of feature selection. The number of used features has been varied for different pairs of classes and was adjusted automatically on the basis of collective evaluation of features. The composition of the best features was strongly dependent on the applied selection technique. The numerical experiments of recognition have been performed for optimal sets of features for each 2-class recognition subproblems (at 10 classes it means 50 SVM individual classifiers). For comparison the same experiments have been repeated for the sets of worst features and also for the whole set of features. The number of learning data used in experiments was equal 925 and identical number of data has been used in testing.

Choice of features	SVM (single feature)	SVM (collective features)	Mean and variance	Correlation
Set of best features	18.7%	19.2%	18.7%	18.6%
Set of worst features	35.2%	32.9%	31.8%	33.3%
All features	23.7%			

Table 1: The overall recognition error rate at different selection methods of features

Table 1 presents the overall error rate of blood cell recognition on the testing data at different selection methods. The testing data have not been used previously in learning. The benefits of application of the feature selection is evident.

In our data set there are cells close to each other in their development stage. For example the cells of classes 1, 2 and 3 represent the succeeding stages of the erythrocytic development line, the cells from 4 to 9 - the granulocytic line. Recognition between two neighbouring cells in the same development line is dubious even for human expert, since the images of both cells are very alike and thus difficult to recognize. Close analysis of our misclassification cases reveals that most errors have been committed at the recognition of the neighbouring cells in their development lines. The summary of performance of the classifiers at neglecting the errors following from the neighbourhood of cells in their development line is presented in Table 2.

Choice of features	SVM (single feature)	SVM (collective features)	Mean and variance	Correlation
Set of best features	5.1%	5.1%	5.3%	5.3%
Set of worst features	10.7%	12.3%	9.6%	10.1%
All features	5.6%			

Table 2: The overall recognition error rate at different selection methods of features neglecting the neighborhood errors

5 Conclusions

The results of experiments prove the important role of feature selection. Irrespective of the applied method, the selection improves significantly the accuracy of recognition. The observed improvement rate for 10 classes of cells is up to 20% with respect to the case of applying all features. Among the best feature selection methods is the correlation between the feature and the class as well as the linear SVM ranking based on single feature application.

References

[1] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, JMLR, 3, 1158 – 1182, 2003

[2] K. Lewandowski, A. Hellmann, Haematology atlas, Multimedia Medical Publisher, Gdansk, 2001

[3] S. Osowski, T. Markiewicz, B. Mariańska, L. Moszczyński, Feature generation for the cell image recognition of myelogenous leukemia, IEEE Int. Conf. EUSIPCO, Vienna, pp. 753-756, 2004

[4] B. Schölkopf and A. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002

[5] J. Schurmann, Pattern classification, a unified view of statistical and neural approaches, Wiley, 1996

[6] P. Soille, Morphological image analysis, principles and applications,, Berlin: Springer, 2003

[7] N. Theera-Umpon, P. Gader, system-level training of neural networks for counting white blood cells, IEEE Trans. SMS-C, 32:48-53, 2002

[8] T. Wagner, "Texture analysis" (in Jahne, B., Haussecker, H., and Geisser P., (Eds.), Handbook of Computer Vision and Application), 275-309, Academic Press,1999