

Classification of leukemia diseased cells

Chun Hua Shi

Langfang Health Vocational College, Hebei, China, Email:545682864@qq.com

Abstract. In this paper, the characteristics of leukemia diseased cells, diseased cells existing classification methods were studied and analyzed, in order to solve classification categories leukemia diseased cells more characteristic dimensions of the problem of high, finite automata and kernel methods, and experimental verification the advantages of this method.

1 Introduction

Leukemia (Leukemia) is a malignant tumor of the hematopoietic system, currently in the country is a threat to children and young adults leukemia life and health of the most common malignancies. Blood disease occurs in modern society, how to properly detect blood disease and propose effective and reliable treatment options based on the detection result, it has become a major issue in today's medicine. Blood disease is due to human hematopoietic dysfunction caused by blood cells in quantity, shape, proportion and other aspects have changed, so the correct identification of blood cells are correctly detected blood disease premise. At present, for the analysis of microscopic images of leukemia also not much hope to design a blood-based leukemia microscopic image identification and classification system.

2 Blood diseases and typing

According to current international combined cell morphology, tissue staining, flow cytometry and immunohistochemistry leukemia typing will be divided into:

1. Acute leukemia: acute non-lymphocytic leukemia and acute lymphocytic leukemia:
2. Chronic leukemia:
 - 1) chronic myelogenous leukemia (CML): bone marrow myeloid hyperplasia, intermediate stage cell hyperplasia.
 - 2) chronic lymphocytic leukemia (CLL): bone marrow mature lymphocytes $\geq 40\%$.

The test sample is collected from chronic myelogenous leukemia (also known as chronic myeloid leukemia) of the blood film. Chronic myeloid leukemia blood film will appear segmented neutrophils and rod-shaped

granulocytes, which is cell category this test to study.

3 Existing classification

By selecting features will be recognized classification objects merge, confirm their category is called classification process. The process is based on the appropriate decision rule, the feature space is divided into different types of samples. In some practical process, the pre-given condition, the category attribute is considered often have similarities, classification error is inevitable, therefore, the classification process can only be done in a certain error rate. Obviously, the smaller the better classification error rate. However, the classification error rate and is subject to many conditions, such as classification, classifier design and selection of samples and extracted characteristics and other factors will affect the classification results.

method	characteristic	Pros and cons
Recently mean classifier	Assigned to the category average of the most recent sample category	Rapid detection of the dependent measures taken
Nearest Neighbor Classifier	Assign samples to a class of its nearest training samples category	Robust performance, testing is slow, for dependent measures taken
Bayes classifier based law (minimum risk)	Assigned to the type of sample maximum a posteriori estimate the probability of	Based on a simple Gaussian distribution assumption classifier (linear or quadratic), density estimation error sensitive
Fisher linear classifier	Using the mean square error (MSE) criterion linear classifier	Quick and easy one-class classification, similar laws Bayes classifier, classification criterion function using a covariance matrix
Binary decision tree classifier	Achieve classification by finding a series of samples based on	Iterative training process, overtraining sensitive, we need to have a termination

^a Corresponding author: yuanquanbo@126.com

	pattern features Lanna value	condition is detected quickly
sensor	An iterative optimization of linear classifiers	Sensitive training parameters
Multilayer Perceptron Neural Network	Based iterative minimum mean square error for the Priority 2 layer or multi- layer perceptron (neurons) using a transfer function Sigmoid	Parameter sensitivity training, training is slow, non-linear classification function, overtraining sensitive, need to be adjusted

Table 3-1 several commonly used classification methods

It can be divided into statistical decision method, syntactic structure, fuzzy judgment method and artificial intelligence four categories according to their method. Former two methods is the classic pattern recognition technology, the introduction of fuzzy mathematics research in these two methods, greatly improving the classification results. Artificial Neural Network 1980 years revival, more globally relevant feature in the field of pattern recognition has made many using traditional methods difficult to achieve success. Table 3-1 lists commonly used in blood cell classification method.

As can be seen from the above discussion and the work of their predecessors, in the blood cell classification category more due to the classification, feature dimension is high, so the problem classifier selection, structural design, and classification speed, accuracy, etc. are required considering the research prospects.

4 Support vector machine kernel function method of classification of blood cells

In order to solve the multi-category classification, characteristic dimension of the problem of high blood cell classification in the above, we use support vector machine classification. Support vector machine can solve high-dimensional feature of classification.

4.1. Linearly separable support vector machine

Support vector machine can be further divided into linear separable support vector machine, kernel function vector machines. We first examine the classification question whether linear classification problems, whether to use linear support vector machine method can be solved.

Using linear support vector machine plus soft margin optimization method, using 120 samples for training, with 30 samples were tested and the results are shown in Table 4-1.

C	The total number of SV	The number of Class 1 SV	The number of Class 2 SV	The number of Class 3 SV	Accuracy
1	62	21	30	11	91.057
2	57	21	26	10	91.057
3	47	17	23	9	91.87

Table 4-1 classification linear SVM

The first column in Table 4-1 interval optimization soft penalty factor, from the above table, we found that the greater the penalty factor, the fewer number of support vectors obtained, it means that the promotion of learning machines get better. Moreover, the penalty factor, the higher the accuracy of the test. This reflects that the system is approximately linear separable, but linear classifier and can not get the best results.

4.2 Kernel Methods

If you want to use the kernel function method, then choose what kind of non-linear classifiers have the best effect? Select nonlinear classifier, in fact, it is to choose the type of kernel function and its parameters. Present research kernel is not very thorough, it has not been studied all possible kernel function and its advantages when classification. Now people can only prove out several common kernels on classification issues have a better effect. These core functions are:

1. polynomial kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = [(\mathbf{x}_i \cdot \mathbf{x}_j) + 1]^d \quad d = 1, 2, \dots \tag{5-1}$$

2. Sigmoid Kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh[v(\mathbf{x}_i \cdot \mathbf{x}_j) + c] \tag{5-2}$$

3. Gauss radial basis kernel function

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp \left\{ -\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2} \right\} \quad (5-3)$$

We experiment with polynomial kernel and a Gaussian kernel to do a test classification, respectively, and have chosen a few parameters for each kernel. Still 120 training samples, 30 test samples. The results are shown in Table 4-2.

Table 4-2 and Table 4-1 comparison, it can be seen, with the kernel function improved classification accuracy than the method of linear classification soft interval. Moreover, the radial basis function better than polynomial kernel function results. After you have selected the radial basis function, we choose RBF kernel function of several parameters. Gamma is where RBF kernel function parameters, C is the soft margin optimization penalty factor. After testing a number of parameters, we found that when the Gamma = 2, C = 2 - delivering up to the maximum classification accuracy, when Gamma and C and then increase, not in the correct classification rate increases. So learning machine obtained at this time is the most suitable machine learning WBC we can get.

		C	Gamma	nSV	The number of Class 1 SV	The number of Class 2 SV	The number of Class 3 SV	Correct rate
RBF		3	3	68	23	28	17	93.33
		2	2	51	21	22	16	93.33
		1	2	67	25	27	15	92.68
		1	1	61	20	27	14	90.44
		2	1	54	19	22	13	91.7

Table 4-2 polynomial kernel and RBF kernel of the classification results

	Degree	C	Gamma	nSV	The number of Class 1 SV	The number of Class 2 SV	The number of Class 3 SV	Correct rate
Polynomial	2	1	1	47	17	21	9	91.87
	2	2	1	42	15	19	8	92.68
	3	1	1	42	13	21	8	92.68
	2	1	2	41	14	19	8	92.68

When the prediction of a learning machine with the best forecast of 93.496% accuracy rate. Its confusion matrix (Confusion Matrix) as shown in Table 4-3, where the i-th row j-th column represents the first i kind of cells into the wrong kind of probability j cells.

	Lobulated nucleus	The rod	Damaged
Lobulated nucleus	0.95	0.05	0
The rod	0.075	0.9	0.025
Damaged	0	0.05	0.95

Table 4-3 predicted confusion matrix

5 summary

By support vector machines and kernel methods on cell sorting experiments, the results illustrate the applicability of using support vector machine kernel function method on blood cell, indicating that it has a good prospect in leukemia classification pattern recognition .

References

- [1] Lijing Qin blood cytology Beijing: Chinese Medicine Press, 2001.
- . [2] Lin Fengru, Ren Jinhai World Health Organization on the classification and diagnosis of acute leukemia clinical meta, 2004, 19 (22): 1315 - 1318.