Hemolysis detection based on SVM of Adaboost classification algorithm

xiaonan Shi^{1,*}, Zitong Wang¹, and Zhenmin Zhao²

¹Xi'an University of Science and Technology, IT Academy, Xi'an, Shaanxi Province, China

²Shaanxi Victoria as the digital image Technology Co., Ltd, IT Department, Xi'an, Shaanxi Province, France

Abstract. Aiming at the problem that clinical hemolysis is difficult to be observed and judged, a method of Adaboost learning classification based on SVM is proposed. The method firstly extracts the basic features of the target area of the blood sample, such as the average of the gray level, the standard deviation of the gray level and the appearance frequency of the particles, as the input eigenvectors of the learning, and carries out SVM weak learner learning. Subsequently, Adaboost algorithm is used to measure the weak learner Set linear weighting, so as to enhance the strong learning device; Finally, online testing, calculation of test sample hemolytic degree and classification. The Adaboost learning classification test based on SVM is compared with the macroscopic and red blood cell counting methods. The experimental results show that the learning-based classification testing method achieves higher detection accuracy without subjective factors and has the highest detection efficiency.

1 Introduction

The abnormal destruction of this red blood cell can cause many diseases to occur^[1-2], the diagnosis of hemolysis in clinic is mostly obtained by the observation of the medical staff on hemolysis and condensation reaction^[3-4], although this method is simple and easy, but it is easy to observe the small amount of hemolysis is susceptible to subjective factors, Therefore, only to adapt to the obvious samples of hemolysis, and in order to be able to accurately and quickly make correct judgments, medical staff need to carry out a long period of learning to accumulate experience, work efficiency is relatively low.

Machine learning^[5-6] is the understanding of human learning and thinking mode by cognitive science and physiology, and establishes the cognitive model or computational model of cognition process based on human thought. Among them, machine learning based on SVM can be used as classification and regression analysis based on structural risk minimization^[7], so the SVM classification algorithm is used to classify the hemolytic samples, but because the SVM algorithm has a smooth effect on the data, the prediction of the samples with mutation characteristics is not good, We introduced the adaboost algorithm^[8] to enhance the SVM classifier: The Learning weights of SVM sample learning are adjusted, the repetition training obtains different SVM classifier, until the preset classifier number is reached, and all SVM classifier is weighted linearly, then a strong classifier is obtained. Through comparing the classification algorithm with other methods, it shows that the classification of blood samples based on SVM

algorithm can make effective and accurate hemolysis judgment quickly and effectively.

2 Weak classifier learning based on SVM

2.1 SVM Machine Learning Principle

SVM is a supervised learning model in the field of machine learning. It is typically used for regression analysis, classification and pattern recognition^[9-10]. SVM changes the low-dimensional linear inseparable use cases into the high-dimensional eigenspace. Use cases can be linearly separable. This mainly uses the non-linear mapping algorithm, so a linear algorithm can be used to analyze the nonlinear characteristics of the use cases linearly. In the implementation, SVM learning strategy is to get the maximum interval, and finally can be classified into a convex quadratic programming problem to solve^[11].

Consider first the linear separability case, assuming two different data types are represented by a circle and a cross, respectively, in a two-dimensional plane (as showed in Fig.1). Depending on the principle of linear separable data, we separate the two types of data by a straight line. We can regard this straight line as a hyperplane. The data point on the hyperplane side corresponds to -1 and the other side corresponds to 1.

* Corresponding author: shelly200607@126.com

Fig. 1. SVM basic two classification concept

The categorical discriminant function $f(x) = w^T x + b$ represents the hyperplane, and when $f(x) = 0$, x is the point on the Yiping surface, and the $f(x) > 0$ point corresponds to the $y = 1$'s data point, and the point of $f(x) < 0$ corresponds to the point of $y = -1$, as shown in Figure 1 (b). The greater the "spacing" between the hyperplane and the data points, the higher the accuracy of the classification. To make the classification more accurate, you must select a hyperplane that maximizes this "spacing" value, as shown in Figure 1 (c).

The maximum distance classifier's objective function is defined as: $\max \tilde{\gamma}$, while meeting some constraints:

$$y_i(w^T x_i + b) = \hat{\gamma}_i \geq \hat{\gamma}, \quad i = 1, \dots, n \quad (1)$$

Geometric interval is defined as:

$$\tilde{\gamma} = \frac{\hat{\gamma}}{\|w\|} \quad (2)$$

The above objective function is transformed into:

$$\max \frac{1}{\|w\|}, \quad s.t., y_i(w^T x_i + b) \geq 1, \quad i = 1, \dots, n \quad (3)$$

Since the prescribed function is quadratic and has linear constraints. The QP (Quadratic Programming) optimization package can be used to solve the convex quadratic programming problem.

This paper discusses the general linear irreducible situation, in fact, most of the time, the data are linearly not divided, you cannot find the maximum spacing classification hyperplane (Fig. 2 (a)). In a linear and $k(x)$ state, SVM constructs a new discriminant function $f(x) = w^T k(x) + b$ by selecting a kernel function, and sends the data to the high-dimensional space so that it can be linearly divided, as showed in Fig. 2(b).

Fig. 2. SVM classification of non-linear distinction method

Commonly used kernel functions are isotropic d-order polynomials, anisotropic d-order polynomials, radial basis functions, Gaussian radial basis functions, etc. It is interesting to note that the selection of kernel functions during training is often not critical. Several simple kernel function classifier performance is usually comparable^[12].

2.2 Feature extraction of blood samples

In the process of classification of hemolytic images, three features of the target area will be extracted: the average of the gray scale $\bar{g}$, the standard deviation of the gray scale σ , the frequency of the particles $fren$, corresponding to the three components of the eigenvector x in the above text. Suppose the number of samples is N , the i th eigenvector $x_i = (\bar{g}_i, \sigma_i, fren_i)$, the evaluation score and the classification S correspond to the above $f(x)$, the i th evaluation score and the classification S_i . The eigenvector $x_i = (\bar{g}_i, \sigma_i, fren_i)$ is invoked as the input, the evaluation score and the classification S_i as the output. Depending on the training and learning process, the non-linear function of the evaluation score, classification S_i , and eigenvector $x_i = (\bar{g}_i, \sigma_i, fren_i)$ is obtained. That is, the training model is obtained. After getting the learning model, online fusion testing can be done. The image is acquired online, the target area is selected, the three characteristics of the area are calculated and used as the input of the learning model, and the evaluation scores and the classification corresponding to the target area are obtained through the learning model calculation.

Hemolytic target area image sample feature extraction, followed by the average gray-scale features, gray-scale standard deviation characteristics and particle frequency characteristics.

2.2.1 Gray average feature

The gray mean value of image in target region mainly reflects whether the blood distribution is completely insoluble, the illumination is completely through the blood test tube, at this time the average gray value of the target area image is very high; when the part of the blood is dissolved, the illumination part passes through the

blood test tube, at this time the average gray value of the target area image is relatively low the illumination was relatively few. The average gray value of the pixel in the target zone is:

$$\bar{g} = \frac{\sum_{i=1}^n g_i}{n} \quad (4)$$

Where g_i represents the pixel gray value at any point in the target area, i is the i th pixel in the image, and n is the total number of pixels in the target area.

2.2.2 Gray standard deviation characteristics

The characteristic of gray standard deviation of image in target area mainly reflects the degree of blood-soluble compatibility. That is, the degree of uniform intensity of the target area. When the blood distribution is completely insoluble, the illumination is completely through the blood test tube, at this time the target area image average gray value is very large, and the gray standard deviation is small, sometimes even 0; when the distribution of blood is completely dissolved, light through the blood test tube is very few, at this time the target area image average gray value is very small, and gray standard deviation is also relatively small; when the blood is partially fused, The illumination part passes through the test tube, at this time the target area image Gray standard difference is big, when the target area average value is bigger, the gray standard difference is bigger, the match blood consistency is better; Conversely, when the target region average time, the gray standard deviation is bigger, the match blood consistency is worse That is, when the target area gray standard difference is compared with the hour, the gray standard deviation affects the matching degree of the gray mean, which plays a positive role; when the target area gray standard deviation is larger, the gray standard deviation plays a reverse role in the influence of the gray mean on the matching degree of the blood distribution. Grayscale Standard deviation σ :

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (g_i - \bar{g})^2}{n-1}} \quad (5)$$

2.2.3 Particle frequency characteristics

In the hemolysis process, the blood sometimes condenses into granules, extract these granules from the target area as a particle characteristic, count the number of these characteristic particles, and calculate the frequency of occurrence of these particles (particle frequency = number of particles / target area image pixel Number). When the blood is completely incompatible, less blood coagulation, the frequency of particles in the target area is relatively small, but the gray value of the target area is very large; when the blood is completely matched, the blood coagulation is less, the frequency of particles in the target area But the

gray value of the target area is relatively small at this time; when the blood is partially matched, the blood condenses into a granular form, and the frequency of the particles in the target area is relatively large. When the gray value of the target area is relatively large, the frequency of the particles is higher , The better the degree of matching with blood; the contrary, when the average of the target area is relatively small, the greater the frequency of particle appearance, the worse the degree of matching blood. That is to say, when the frequency of the particle in the target area is relatively small, the frequency of the particles affects the mean value of the gray level to match the degree of blood matching, which plays a positive role in promoting. When the frequency of particles in the target area is relatively large, Impact on the degree of matching with the blood played the role of reverse offset. Calculate the particle frequency of the target area as follows:

- (a) Gaussian template image denoising for target area image;
- (b) The image computed Sobel gradient image after denoising;
- (c) The Sobel gradient image is trimmed to a threshold value of two-valued image;
- (d) Statistics of binary image connected domain number, that is, the number of statistical particles n_p ;
- (e) Calculate the frequency $fren = n_p / n$ of the particles, in which n_p is the number of particles, n the number of pixels in the target region image.

2.3 Learning training based on SVM algorithm

1) Training Database collection: First, collects the image, select the target region of the blood distribution in the image, then extract the eigenvector of the target area, and then give the hemolytic degree score evaluation and classification S of the target area. The image sample database and the data collection $x_i = (\bar{g}_i, \sigma_i, fren_i)$ are set up in order to collect the images continuously.

2) Support Vector Computation: Assuming that we have N independent and homogeneously distributed blood samples, x_i is an input eigenvector of dimension 3 and s_i is the corresponding hemolysis fraction and class label, the goal of SVM is to find in the feature space The linear discriminant that satisfies the maximized category boundary, where the kernel function is written as $k(x)$ (feature space is $\mathbf{R}^3 \rightarrow \mathbf{R}^s$), then the discriminant function is:

$$f(x) = w^T k(x) + b \quad (6)$$

The ω is the weight coefficient, and the b represents the offset of the separation hyperplane. Using Lagrange to odd, the above discriminant function can be deformed as follows.

$$f(x) = \sum_{i=1}^N a_i s_i k(x_i, x) + b \quad (7)$$

The objective function is:

$$\min \frac{1}{2} \|w\|_2^2, \quad s.t., s_i(w^T k(x_i) + b) \geq 1, \quad i = 1, \dots, N \quad (8)$$

The $w \in \mathbb{R}^S$ is a vector of weighting coefficients, and the $b \in \mathbb{R}$ represents the offset of the detached hyperplane.

The minimization (8) can be accomplished by: Satisfying $\sum_{i=1}^N s_i a_i = 0$ and $a_i \geq 0$ conditions, and maximizing Lagrangian functions relative to a_i :

$$L(w, b, a) = \sum_{i=1}^N a_i - \frac{1}{2} \sum_{i,j=1}^N a_i a_j s_i s_j k(x_i, x_j) \quad (9)$$

Any Lagrange multiplier a_i has corresponding training eigenvector x_i . Vectors that contribute to maximizing spacing have nonzero a_i , called support vectors. The rest of the training vectors is not the final discriminant function, so we just store the corresponding training vectors of 0 a_i .

3 Strong classifier learning based on adaboost algorithm

The AdaBoost algorithm is an algorithm that can upgrade a weak learner to a strong learner [13-14]. The basic idea of the adaboost algorithm is to learn a weak learner from the training sample, then change the weight of the sample according to the learning result, then learn a weak learner according to the modified training sample, and continue to do it repeatedly until the number of the learners reaches the preset value, in the T A weak learner is weighted to get a strong learner.

In this paper, we apply the standard SVM algorithm as a weak learner to the Adaboost algorithm to improve the prediction accuracy of the algorithm, as showed in Fig.3.

Fig. 3. Adaboost algorithm based on SVM

Algorithm flow:

Enter: Training Set

$$D = \{(x_1, s_1), (x_2, s_2), \dots, (x_m, s_m)\} \quad ; \text{Training rounds } T;$$

Step1. Let all the samples have the same weight in

$$\text{the initial state } w_i^1 = \frac{1}{m};$$

Step2. For $t = 1, 2, \dots, T$ do

(1) training base classifier $SVM_t(x)$ is obtained

according to the training set with weight w_i^t distribution;

(2) Calculate the classification error rate of $SVM_t(x)$ on the training data set:

$$e_t = P(SVM_t(x_i) \neq s_i) = \sum_{i=1}^m w_i^t I(SVM_t(x_i) \neq s_i) \quad (10)$$

(3) Calculate the coefficient of $SVM_t(x)$:

$$\beta_t = \frac{1}{2} \ln \left(\frac{1 - e_t}{e_t} \right) \quad (11)$$

(4) update the training data set weights:

$$w_i^{t+1} = \frac{w_i^t}{Z_t} \exp(-\beta_t s_i SVM_t(x_i)), i = 1, 2, \dots, m \quad (12)$$

among them

$$Z_t = \sum_{i=1}^m w_i^t \exp(-\beta_t s_i SVM_t(x_i)) \quad (13)$$

Step3. Construct a linear combination of weak classifiers

$$G(x) = \sum_{i=1}^T \beta_i SVM_i(x) \quad (14)$$

4 Experiment Analysis

After establishing the target Area sample library, by collecting 5 blood sample images, selecting the detection target area, calculating the gray average gray character of the target area, the gray standard deviation feature and the particle frequency characteristic, the target region corresponding classification and its fractional value are predicted by the learning model presented in this paper, and the results are as shown in Fig.4 and Fig.5:

Fig. 4. Blood samples with a 60-point test result

Fig. 5. Blood samples with a 90-point test result

In the process of data experiment, the sample library is divided into two parts, some of which are used to train Adaboost machine learning model based on SVM, the other part of the sample as a control, through the feature extraction and load learning model to predict the part of the sample score and classification, Compared with the actual mark of the score classification. A total of 1600

samples were constructed from 11 groups of samples, 100 samples from each group were taken to form 11 groups of 1100 samples for SVM-based Adaboost learning model training, and the remaining 11 groups of 500 samples were used for prediction, experimental control and training as shown in Table 1 As shown.

Table 1. Determination of hemolytic results of different methods (- said no hemolysis, + said hemolysis)

Group	Visual observation method		Erythrocyte counting method		This test method	
	A	B	A	B	A	B
1	-	-	17.09 ± 3.04	20.11 ± 2.79	20	20
2	-	-	21.79 ± 4.69	23.34 ± 5.61	20	20
3	-	-	50.94 ± 2.42	49.23 ± 6.30	50	50
4	-	+	58.12 ± 8.66	60.79 ± 8.12	60	60
5	+	+	77.64 ± 5.63	76.89 ± 7.64	80	80

In this paper, the method based on SVM based on AdaBoost learning is compared with the naked eye and RBC counting [15]. Measured data as showed in table 1, through the experimental results found that the production of hemolytic samples can be observed by the naked eye, using this algorithm. RBC classification algorithm to determine the degree of hemolysis in 60 Above.

However, between hemolytic and insoluble hemolysis or uncertain samples, each individual has different judgments, the impact of subjective errors. Although the use of red blood cell counting method from the subjective factors, but the experimental process is numerous, longer experimental period, erythrocyte itself prone to rupture during the experiment, resulting in measured values will produce errors, so the practicality of detection of hemolysis is not high. The proposed Adaboost classification algorithm based on SVM is not affected by the subjective factors, and the hemolysis judgment which is similar to the erythrocyte counting method is provided simply and effectively.

5 Summary

Clinically hemolysis detection mostly obtained by the naked eye, so the individual subjective factors, and the detection efficiency is low, especially the hemolytic degree is not easy to observe the situation, the observation often fails. In this paper, a new Adaboost classification method based on SVM is proposed. This method learns a large number of samples through feature extraction and sample hemolysis and classification evaluation, and obtains a strong classification model suitable for hemolysis detection. Through the experimental verification, the detection algorithm in this paper can reach the accuracy of manual detection completely, and the detection speed is fast and the detection efficiency is high, which saves a lot of

manpower costs and significantly improves the work efficiency and the detection precision.

References

1. Qian Yinjuan newborn hemolysis test clinical significance [J]. Chinese Medical Journal, 2016,14 (23): 60-62.
2. Ren Guoqing, Niu Hairing. Blood disease in patients with serum lactic dehydrogenase concentration changes in clinical significance [J]. Modern Journal of Integrated Traditional and Western Medicine, 2010, 19 (1): 100.
3. Effect of hemolysis on the results of procoagulant test [J] .Journal of Sichuan University Medical College, 2010,25 (05): 449-452. LIU Wen, LEI Jiahong, TANG Zhong, TANG Ting,
4. Huang Hongwei. Automatic biochemical analyzer detection of hemolysis specimens on the accuracy of results [J]. China Medical Instrumentation Information, 2016,22 (06): 49-50.
5. Simon J.D. Prince. Computer vision, models, learning and inference[M]. 2012 by Simon Prince; to be published by Cambridge University Press 2012.
6. LUO Yu. Application Research of Support Vector Machine in Machine Learning [D]. Southwest Jiaotong University, 2007.
7. N.Yugo, Y. Kazuyosh, A. Makoto. Support vector machine and its bias correction in high-dimension, low-sample-size setting[J]. JOURNAL OF STATISTICAL PLANNING AND INFERENCE, 2017, 191 : 88-100.
8. Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting[C]// European Conference on Computational Learning Theory. Springer, Berlin, Heidelberg, 1995:23-37.
9. M.Sonka, V. Hlavac, R. Boyle. Image processing, analysis, and machine vision (Third Edition) [M]. Springer, Boston, MA, 2008.
10. Chang tiantian. Support vector machine learning algorithm of a number of issues [X]. Xidian University, 2010.
11. Ma Xiaoyong. Expression and classification of cardiopulmonary bypass based on deep learning [D] .Zhejiang University Master's degree thesis, 2016.
12. M. Hearst. Trends and controversies: support vector machines. IEEE Intelligent Systems, 13:18-28, 1998.
13. Zhou Zhihua. Machine Learning [M]. Beijing: Tsinghua University Press. 2016
14. Yan X, Wang Y, Song Q, et al. Salient object detection via boosting object-level distinctiveness and saliency refinement[J]. Journal of Visual Communication & Image Representation, 2017.
15. Zhang Jianfeng, Xiang, Shand. Study on hemolysis detection method of Xuesaitong injection [J].

Chinese modern Chinese Herbal medicine, 2011, 13
(1): 34-38.