Image Classification via Support Vector Machine

Xiaowu Sun[1], Lizhen Liu[1], Hanshi Wang[1], Wei Song[1], Jingli Lu[2]
[1] Information and Engineering College, Capital Normal University, Beijing 100048, P. R. China
[2] Agresearch Ltd, New Zealand

***Abstract*—With the rapid growth of images information, how to classify the images has been a main problem, and most of researchers are concerning on the neural networks to realize the images classification. However, the neural networks can not escape from its own limitations including the local optimum or the dependence on the input sample data. In this paper, another new algorithm named support vector machine, whose main idea is to build a hyperplane as the decision surface, is introduced to solve the problems. In the theory part, in order to solve the optimal hyperplane for the separable patterns problem, the method of Lagrange multiplier is transformed into its dual problem. In the application section, where it proves that the support vector machine can solve the problem of classification perfectly, with regard to the input data, the eigenvalues of the images' gray information which are treated by the method of Principal Component Analysis are abstracted as input sample. It is found that the precision of the classification could arrive at 89.66%, which is far higher than the neural networks' 41.38%.**

Keywords—support vector machine; neural networks; image classification; principal component analysis; hyperplane

I. INTRODUCTION

In the information era, the big data, which is described by TB instead of GB now, greatly influences the culture and the life of our community continuously. And at the same time, not only the numerical information and media information but also the figure information, which contains lots of resources, is worth being mined. But the problem we face is that how to classify the information we need from all kinds of data. With the development of the intelligent computation, the neural networks solve that problem by data prediction and pattern recognition[1] and regression, but the neural networks also have their own limitations, for example, it sometimes could fall into the local optimum and it relies too much on the input data. Thus, exploring another effective method to solve those issues is particularly urgent.

In 1995, Vapnik, a well-known Russian mathematician and statistician, raised an innovative algorithm named support vector machine [2], [3] which is introduced on the basis of the traditional statistics theory. Compared with the neural networks [2], [4], [5], the support vector machine could just depend on its own mathematical theoretical foundation to build an effective high-dimensional data model, even if the sample data are inadequate. Moreover, better generalization ability and convergence to the global optimum are the other features of support vector machine. The support vector machines get rid of the bound with building the learning machine from the view of biological bionics, so support vector machine has become another one popular field of machine learning research after the neural networks[6] and promoted the development of machine learning theory.

In this paper, some discussions of the support vector machine's theory are introduced firstly, then the solving process of the support vector machine objective function is shown. Secondly, in order to accomplish the task of classification[7], the images in the experiences are dealt with. At last getting help from the support vector machines, the images are classified successfully.

II. RELATED WORK

On the one hand, the image classification[8] has occurred for long time, and the main thought is to obtain the eigenvalue from the images including the color feature, texture, the shape, the spatial relation and so on. The eigenvalue is used as the input data or the sample data to train a neural network repeatedly, after that, the neural networks trained by the input data recognize the classification with the test-data. The neural networks could solve the classification problem, although the neural networks have its disadvantages. For example, only by virtue of the input samples, the parameters of the networks, which are not adjustable to all of the samples, can be solved. And the solving process obeys the iteration which spends too much time on getting an optimal answer.

On the other hand, compared with multilayer perceptrons and the neural networks, support vector machines present excellent features not only in regression problem but also in image classification. In order to overcome the disadvantages which appear in the neural networks, the support vector machine is expressed as simply linear function. What's more, in the machine learning the curse of dimension has been a challenging problem, but when the thought of Inner-Product Kernel[2],[9] is adopted in the support vector machine, the problem of classification seems much easier.

III. THEORIES OF SUPPORT VECTOR MACHINE

In this part, the theory of the support vector machine is simply depicted firstly, then we describe in detail to explore the optimal hyper-plane for linearly separable patterns problem with Lagrange multipliers and its dual problem.

A. The Discriminant Function of Support Vector Machine

The training sample $\{x_i, y_i\}_{i=1}^{N}$, in which N indicates the number of the sample x_i, is the ith input example with y_i corresponding desired target output. And the value of y_i only has two cases, which stand for two patterns. One is $y_i = 1$, the

978-1-4673-8173-4/15/$31.00 ©2015 IEEE

other one is $y_i = -1$. The pattern of the test sample $\{t_i\}_{i=1}^{N}$ can be denoted as $t_i\_label = \mathbf{sgn}(w_0 t_i + b_0)$ using the support vector machine built by the training sample. So the discriminant function is defined by

$$g(x) = w^T x + b \tag{1}$$

where the w represents an adjustable vector, and b means a bias. When $g(x_i) > 0$, the sample x_i is the positive side, and on the contrary, when $g(x_i) < 0$, the sample x_i is the negative side. Accordingly, the equation $g(x) = 0$ could be regarded as a decision surface, which also is called hyper-plane and we sign it by writing

$$w^T x + b = 0 \tag{2}$$

Now, both of the sample x_1 and sample x_2 are on the decision surface, thus

$$w^T x_1 + b = w^T x_2 + b \tag{3}$$

which is equal to

$$w^T (x_1 - x_2) = 0 \tag{4}$$

Hence, the vector w is the normal vector of the decision surface.

B. *Optimal Hyperplane for Linearly Separable Patterns*

Constructing an optimal hyperplane regarded as the decision surface using the input samples to make the two sides' margin largest is the main mission of support vector machine. Furthermore, shown as the Figure1, the support vectors must satisfy the constraint:

$$w^T x_i + b = \pm 1, y_i = \pm 1 \tag{5}$$

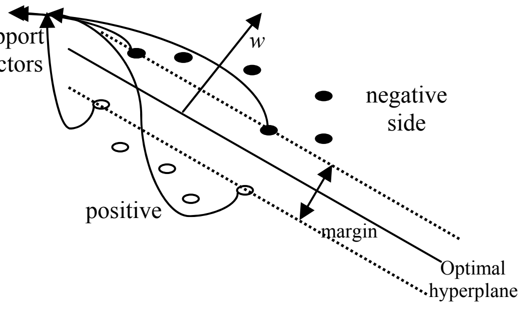

Fig. 1. A simple illustration of the optimal hyperplane.

From Figure1, it can be concluded that the maximized margin exists in the support vectors among the negative side and the positive side. Therefore, the expression of an algebraic measure of distance from the sample x to the optimal hyperplane must be provided. The input sample x also can be expressed as

$$x = x_p + r\frac{w}{\|w\|} \tag{6}$$

The relationship between x_p and x is a normal projection onto the optimal hyperplane and r is an algebraic distance from x to the optimal hyperplane as Figure 2.

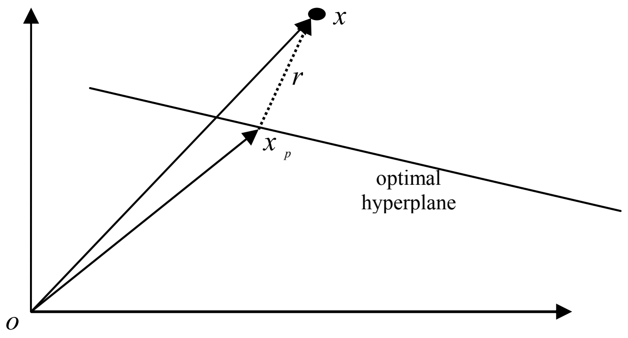

Fig. 2. The expression of the input sample

Both sides of equation (6) are multiplied by w^T , i.e.

$$w^T x = w^T x_p + r\frac{w^T w}{\|w\|} \tag{7}$$

where vector x_p , which satisfies the formula (2), is on the optimal hyperplane, so the equation (7) can be simplified further as

$$w^T x = -b + r\|w\|$$

which is the same as

$$w^T x + b = g(x) = r\|w\| \tag{8}$$

Therefore, the desired algebraic distance r follows that

$$r = \frac{g(x)}{\|w\|} \tag{9}$$

and in the view of formula (5), the maximized margin shown in Figure 1 can be referred to as

$$\rho = \frac{1}{\|w\|} + \frac{1}{\|w\|} = \frac{2}{\|w\|} \tag{10}$$

Equation (10) indicates that the maximum value of ρ is equal to the minimum value of the Euclidean norm of w . Besides, only when the input training samples subject to the following condition, the optimal hyperplane can be built.

$$y_i(w^T x_i + b) \geq 1 \qquad i = 1,2,...,N \tag{11}$$

which also can be determined from the equation (5) and the Figure 1.

In brief, the optimization problem can be fully depicted as:

objective function:

$$\min \quad \phi(w) = \frac{1}{2} w^T w$$

subject to:

$$y_i (w^T x_i + b) \geq 1 \qquad i = 1,2,...,N \tag{12}$$

where the objective function is convex and the constraints are linear concerning w. So Lagrange multiplier method can be introduced to solve the problem.

First the Lagrange function is constructed as

$$L(w,b,\lambda) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \lambda_i \left[y_i (w^T x + b) - 1 \right] \tag{13}$$

whose expansion is

$$L(w,b,\lambda) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \lambda_i y_i (w^T x_i + b) + \sum_{i=1}^{N} \lambda_i \tag{14}$$

where the parameter λ_i is Lagrange multipliers.

Then, differentiating the formula (14) with respect to w and b, we get two conditions of optimization problem:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{N} \lambda_i y_i x_i = 0$$
$$\frac{\partial L}{\partial b} = \sum_{i=1}^{N} \lambda_i y_i = 0 \tag{15}$$

which implies that

$$w = \sum_{i=1}^{N} \lambda_i y_i x_i \tag{16}$$

If we analyze the equation (13), we conclude that the parameter $\lambda_i \geq 0$ and the polynomial $\lambda_i (w^T x_i + b) - 1 \geq 0$, so there must be the formula (17) right.

$$\max(-\sum_{i=1}^{N} \lambda_i \left[y_i (w^T x_i + b) - 1 \right]) = 0 \tag{17}$$

and from the formula (17), we can draw the a new result:

$$\max_{\lambda} L(w,b,\lambda) = \frac{1}{2} w^T w \tag{18}$$

Accordingly, the objective function can be reformulated as

$$\min_{w,b} \max_{\lambda} L(w,b,\lambda) \tag{19}$$

In order to solve the problem more conveniently, we first provide the objective function model of the dual problem:

$$\max_{\lambda} \min_{w,b} L(w,b,\lambda) \tag{20}$$

Substituting formula (15) into (20), we get

$$\begin{aligned} \min_{w,b} L(w,b,\lambda) &= \frac{1}{2} w^T w - w^T \sum_{i=1}^{N} \lambda_i x_i y_i - b \sum_{i=1}^{N} \lambda_i y_i + \sum_{i=1}^{N} \lambda_i \\ &= \frac{1}{2} w^T w - w^T w - b \times 0 + \sum_{i=1}^{N} \lambda_i \\ &= \sum_{i=1}^{N} \lambda_i - \frac{1}{2} w^T w = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j (x_i^T \bullet x_j) \end{aligned} \tag{21}$$

Hence, the complete dual problem is

$$\max_{\lambda} \left\{ \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j (x_i^T \bullet x_j) \right\}$$
$$s.t. \qquad \begin{cases} \sum_{i=1}^{N} \lambda_i y_i = 0 \\ \lambda_i \geq 0 \end{cases} \tag{22}$$

So the optimum solution of λ can be acquired by the formula (22), then substituting the value of λ into formula (16), the optimum value of w can be solved with the optimal hyperplane achieved.

C. Optimal Hyperplane for Nonseparable Patterns

But the actual classification is that not all of the patterns are separable and in the nonseparable patterns there are some input samples violating the formula (12) leading to classification errors. There are two ways as shown in the Figure3 where the error may arise. The data point x_1 belongs to the classification error, but the data point x_2 is on the right side of the decision surface.

In order to solve the optimal hyperplane for nonseparable patterns, the ξ called slack variable is introduced to describe the restrictive condition (11) as

$$y_i \left(w^T x_i + b \right) \geq 1 - \xi_i, \qquad i = 1,2,\ldots,N \tag{23}$$

If the variables $0 < \xi_i \le 1$, the data point x_i is attributed to the right classification. On the contrary, if the variables $\xi_i > 1$, it means that the point is on the wrong side.

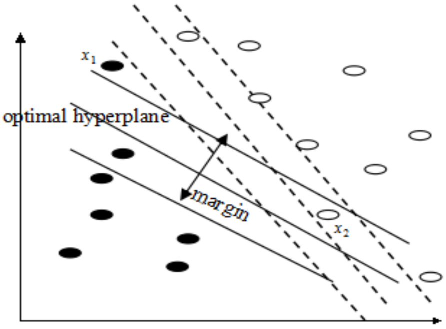

Fig. 3. The description of classification error

We may do this by minimizing the functional to minimize the mis-classification error.

$$\phi(\xi) = \sum_{i=1}^{N} Q(\xi_i - 1) \tag{24}$$

And the function Q is defined according to the above contents

$$Q(\xi) = \begin{cases} 0, & \xi \le 0 \\ 1, & \xi > 0 \end{cases} \tag{25}$$

But, the function $\phi(x)$ is discontinuous, which will bring the trouble into the solving process, so that the thought of function approximation is imported, so the formula (24) can be rewritten into

$$\phi(\xi) = \sum_{i=1}^{N} \xi_i \tag{26}$$

The nonseparable classification case can be concluded as:

$$\min \quad \phi(w,\xi) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i$$

$$\text{s.t.} \quad \begin{cases} y_i(w^T x_i + b) \ge 1 - \xi_i \\ \xi_i \ge 0 \end{cases} \qquad i = 1,2,\ldots,N \tag{27}$$

where C is the parameter, controlling the weight between complexity of the machine and the number of nonseparable points. Using Lagrange multipliers method and proceeding in a manner similar to the described method in the last section, with the Kuhn-Tucker conditions, we may formulate the dual problem for nonseparable patterns as:

$$\max_{\lambda} \left\{ \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j (x_i^T \bullet x_j) \right\}$$

$$\text{s.t.} \quad \begin{cases} \sum_{i=1}^{N} \lambda_i y_i = 0 \\ 0 \le \lambda_i \le C \end{cases} \tag{28}$$

At last, the optimum solution for the Lagrange multipliers can be got perfectly, and the weight vector w is given by

$$w = \sum_{i=1}^{N_s} \lambda_{0,i} y_i x_i \tag{29}$$

in which N_s indicates the number of support vectors.

IV. Experiments Results and Analysis

In this section, some figures divided into 5 categories are taken into the experiment, where it proves that the support vector machine can solve the problem of classification more perfectly than the neural networks[11],[12].

Firstly, the gray value, which ranges from 0 to 255, of the image is constructed as the input vector α_i. 48 images are shown in those experiments, so 48 different input vectors[10] are provided, all of which are expressed as

$$\alpha_i = [a_0, a_2, \ldots, a_{255}]$$

But, the number of the elements in the input vector is too large to be used in the image classification, so in order to simplify the input vectors α_i, a method called Principal Component Analysis is introduced again. As illustrated in the Figure 5, the 256 components develop into 3 components when effected by the component array, which is denoted as $B_{256\times3}$. If we note the array which consists of the 256 components from the 48 sample figures as $A_{48\times256}$, combined with the component array $B_{256\times3}$, the array $C_{48\times3}$ satisfying the formula (37) can be regarded as the new input sample

$$C_{48\times3} = A_{48\times256} \bullet B_{256\times3} \tag{30}$$

For example, an input sample with 256 components, shown in the figure 6, is transformed into a simple input sample with 3 components, when the method of Principal Component Analysis is adopted. The advantage of this method is that the data set can be represented by a reduced number of effective features[13] and yet retain most of the intrinsic information content of the data; in other words, the data set undergoes a dimension reduction.

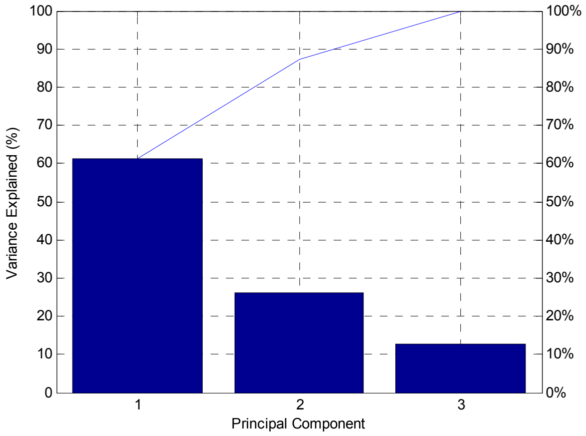

Fig. 4. The Principal Component

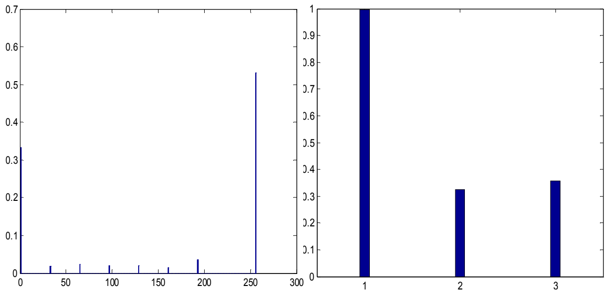

Fig. 5. The input sample α_1 with Principal Component Analysis

The results of the experiments are obvious that the accuracy of the classification by support vector machine is more excellent than the neural networks'. On the whole, the support vector machine's accuracy for this 5 patterns is 89.66%, which is far higher than the neural networks' 41.38%. And the support vector machine reveals better effect on the single classification as the Figure 7 shown.

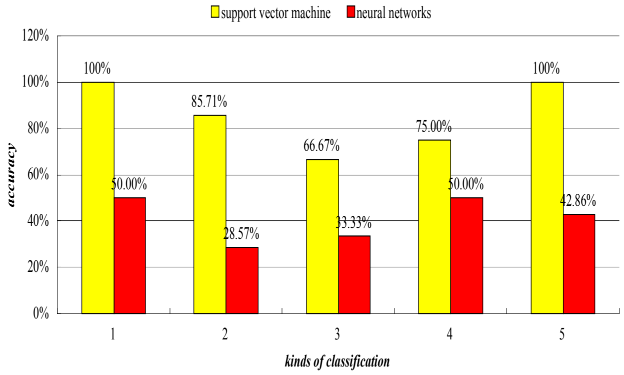

Fig. 6. The Results of Experiments

V. Conclusions

In this paper, we propose a support vector machine model based image classification. With the method of Principal Component Analysis, we abstract the eigenvalue from the gray information of the images to build a support vector machine to classify the images. And the support vector machine can build the decision surface to decide which classification the image should belong to, supported by the Inner-Product Kernel. It has proved that the support vector machine has more advantages than the traditional method, neural networks, which means that extensive experiments have demonstrated the effectiveness of our method compared with several typical methods.

Acknowledgment

This work was supported in part by the Humanity & Social Science general project of Ministry of Education under Grants No.14YJAZH046; National Science Foundation of China under Grants No. 61303105 and 61402304; the Beijing Natural Science Foundation under Grants No. 4154065; the Beijing Educational Committee Science and Technology Development Planned under Grants No.KM201410028017and Beijing Key Disciplines of Computer Application Technology.

References

[1] Van Gemert JC, CJ Veenman, AW Smeulders,JM Geusebroek.Visual word ambiguity. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010

[2] Simon Haykin, Neural Networks A Comprehensive Foundation (Second Edition), Tsinghua University Press, 2001.

[3] Arindam Chaudhuri, Kajal De. Fuzzy Support Vector Machine for bankruptcy prediction[J]. Applied Soft Computing Journal. 2010 (2)

[4] TANG YuHua, ZHANG BaiDa, WU JunJie, HU TianJiang, ZHOU Jing, LIU FuDong. Parallel architecture and optimization for discrete-event simulation of spike neural networks[J]. Science China (Technological Sciences). 2013(02)

[5] A Preliminary Study on Artificial Neural Network[A]. Proceedings of 2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference (ITAIC 2011) VOL.02[C]. 2011

[6] TANG YuHua, ZHANG BaiDa, WU JunJie, HU TianJiang, ZHOU Jing, LIU FuDong. Parallel architecture and optimization for discrete-event simulation of spike neural networks[J]. Science China (Technological Sciences). 2013(02)

[7] Tianzhu Zhang,Si Liu,Changsheng Xu,Hanqing Lu,M 4 L: Maximum margin Multi-instance Multi-cluster Learning for scene modeling[J]. Pattern Recognition. 2013 (10)

[8] Sugata Banerji, Atreyee Sinha,Chengjun Liu. New image descriptors based on color, texture, shape, and wavelets for object and scene image classification[J]. Neurocomputing . 2013

[9] Nakajima S, Binder A, Muller C, Wojcikiewicz W, Kloft M, Brefeld U. Multiple Kernel Learning for Object Classification. Technical Report on Information-Based Induction Sciences. 2009

[10] S. Y. Chen.Kalman Filter for Robot Vision: A Survey. IEEE Transactions on Industrial Electronics. 2012

[11] HAN GuoFeng, LIU XiaoLi, WANG EnZhi. Experimental study on formation mechanism of compaction bands in weathered rocks with high porosity[J]. Science China (Technological Sciences). 2013(10)

[12] Multi-Decision-Tree Classifier in Master Data Management System[A]. Proceedings of 2011 International Conference on Business Management and Electronic Information (BMEI 2011) VOL.03[C]. 2011

[13] Juan C Caicedo, Ebroul Izquierdo."Combining Low-level Features for Improved Classification and Retrieval of Histology Images,". Transactions on Mass-Data Analysis of Images and Signals. 2010