An improved SVM web page classification algorithm

Xun-yi Ren, Chen Shi, Dan Zhang and Wen-si Wang
College of Computers, Nanjing University of Posts and Telecommunications, Jiangsu 210003, China

Abstract. This study explored an indepth support vector machine (SVM) classification algorithm suitable for high-dimension computing. An improved SVM algorithm was proposed aiming at two kinds of classical kernel functions of support vector machine. The global kernel function and the local kernel function were weighted into a new mixed kernel function, and the genetic algorithm was used to optimize the function parameters. The improved algorithm avoided the limitation of a single kernel function, taking into account the generalization and learning abilities of the text classification algorithm. Simulation experiments of the collected web pages confirmed the effectiveness of improved SVM algorithm.

1. Introduction
At present, the mainstream text classification algorithms include Naive Bayes algorithm, decision tree algorithm, K-nearest neighbor algorithm, and support vector machine (SVM). The SVM algorithm is more suitable for calculating high dimension, and has better generalization and extension capabilities. Therefore, it is the most widely used algorithm in the era of big data. However, the SVM is associated with the problem of high-dimensional space complexity and how to better solve the quadratic programming problem. This study aimed to improve the SVM algorithm based on the aforementioned two aspects to further enhance the performance of its classification.

Statistical learning theory is based on a small sample of a machine learning domain knowledge. The theory of structural risk minimization principle is used to solve machine learning problem, and the VC dimension theory is used to operate the machine based on the generalization ability of learning [1]. Therefore, before exploring the support vector machine, it is essential to first understand the two concepts of VC dimension and structure risk minimization principle, and then introduce the thought and nature of SVM algorithm to a solid foundation for improving SVM algorithm.

2. SVM algorithm
2.1. VC dimension theory
The basic concept of VC dimension is the set of functions for a test. If the functions in the set can separate k sample elements according to all possible 2k patterns, it is said that the function set can break k sample elements [2]. In general, an n-dimensional space Rn can have n spatial points that are linearly independent. In this case, the VC dimension of the n-dimensional hyperplane is n + 1 [3].

For the full function of the test function set, a probability of at least $1 - \eta$, $0 \leq \eta < 1$ exists between the test risk $R(\omega)$ and the empirical risk $R_{emp}(\omega)$:
\[ R(\omega) \leq R_{emp}(\omega) + \left( h \left( \log \frac{2l}{h} + l \right) - \log \frac{\eta}{4} \right)^{\frac{1}{l}} \]  

(1)

In the inequality, the number of samples is \( l \), and \( h \) is the VC dimension of the test function set. It is evident from the formula that the test risk is determined by the two parts of the inequality on the right-hand side. One is the empirical risk and the second is the confidence risk, expressed by the following formula:

\[ \phi \left( \frac{h}{l} \right) = \left( h \left( \log \frac{2l}{h} + l \right) - \log \frac{\eta}{4} \right)^{\frac{1}{l}} \]  

(2)

Therefore, formula 2.1 can be simplified as:

\[ R(\omega) \leq R_{emp}(\omega) + \phi \left( \frac{h}{l} \right) \]  

(3)

At this point the relationship is clearer. When the confidence is relatively large, the gap between empirical risk and test risk is widened, leading to the machine learning "over-fitting" situation. Then, it is necessary to control the VC dimension to control the confidence risk in a smaller range and the empirical risk of the minimum value, thereby reducing the test risk and enhancing the generalization performance of the machine learning model.

If the ratio of VC dimension and sample number was appropriate, a better learning model of learning algorithm and neural network could be achieved in accordance with the specific features of the sample itself. This helped in choosing a different VC dimension and minimizing the risk of experience [4]. The subsets were arranged from small to large according to the VC dimension, and then a subset of the function with the same confidence in the risk was selected for the minimum empirical risk [5].

2.2. SVM in the nuclear function

This set of support vectors is a defined set of feature subsets [6]. The SVM algorithm was originally used to solve the problem of linearly separable patterns. The SVM solution can accurately separate two different categories and can also split the plane distance category of the largest spacing, which is the so-called SVM seeking the accurate classification of optimal hyperplane.

The concept of kernel function is as follows [7]. Suppose in a space \( \mathbb{R}^n \) a function \( K(x, y) \) satisfies the Mercer conditions:

\[ K(x, y) = \phi(x) \cdot \phi(y) \]  

(4)

This is called the kernel function. The Mercer condition means that if \( g(x) \in L_2(\mathbb{R}^n) \), \( K(x, y) \in L_2(\mathbb{R}^n \times \mathbb{R}^n) \), for any, \( g(x) \neq 0 \) and \( \int g(x)^2 \, dx < \infty \), the following formula is used:

\[ \int \int K(x, y) g(x) g(y) \, dx \, dy \geq 0 \]  

(5)

For all training elements \( x_1, x_2, \ldots, x_n \in \mathbb{R}^n \), \( K(x, y) = \phi(x) \cdot \phi(y) \) exists for the positive definite matrix. The commonly used kernel functions have polynomial kernel functions:

\[ K(x, y) = \left[ (x \cdot y) + c \right]^p, \quad p = 1, 2, \ldots N \]  

(6)

The Gaussian kernel function is also known as the radial basis function:

\[ K(||x - xc||) = \exp\left(-\frac{||x - xc||^2}{2 \cdot \sigma^2}\right) \]  

(7)
It is the center of the kernel function. \( \sigma \) is the width of the kernel function parameters, which can be used in another form of the function [8]:

\[
K(\|x-x_c\|) = \exp\left(-\gamma \|x-x_c\|^2\right)
\]

(8)

Sigmoid kernel function:

\[
K(x,y) = \tanh(\beta_0 \langle x \cdot y \rangle + \beta_1)
\]

(9)

where \( \beta_0 \) is a scalar and \( \beta_1 \) is the amount of displacement.

3. SVM algorithm improvement

The kernel function directly affects the ability of SVM algorithm [9]. In the commonly used kernel functions, the polynomial kernel function \( K(x,y) = \left(\langle x \cdot y \rangle + c\right)^p \) is a rotation-invariant kernel function. That is, it can be represented by a function \( K(x,y) = f(\langle x \cdot y \rangle) \), where \( f: D \rightarrow \mathbb{R} \) is a unary real function [10]. Therefore, the polynomial kernel function belongs to the global kernel function, and this typical global kernel function has the advantage of being able to extract the characteristics of the overall sample. The disadvantage is that the interpolation ability is relatively low. The parameters of the polynomial kernel function, that is, homogeneous, exponential parameters of the polynomial function of the kernel function, are shown in figure 1. The test point is 0.3. It can be deduced from the figure that the SVM with polynomial kernel function has a high generalization level. If the number of input samples is different and the difference in sample change is large, the polynomial kernel function still has a strong classification ability.

![Figure 1. Global kernel function diagram.](image)

The Gaussian kernel function \( K(\|x-x_c\|) = \exp(-\frac{\|x-x_c\|^2}{2\cdot \sigma^2}) \) is a translation invariant kernel function [11]. That is, it can be expressed by a function \( K(x,y) = f(x,y) \). Therefore, the Gaussian kernel function is a local kernel function, which is superior in terms of the overall performance for local learning classification.

The value of the parameter \( \sigma \) of the Gaussian kernel function is set to 1, 2, 3, and 4, and the test point is 0. The function of the Gaussian kernel function is shown in figure 2. It is deduced that the Gaussian kernel function is complementary to the polynomial kernel function, which is more conducive to learning the local range of the data. In this case, the Gaussian kernel function is inversely proportional to the radius of the Gaussian kernel function.
In the realization of the web page classification algorithm, the learning and generalization performances of the classifier are mutually restricted. The number and type of test samples are usually unbalanced in a high-dimension space. Satisfactory results could not be achieved in this study based on the aforementioned detailed analysis of the properties and characteristics of polynomial kernel function and Gaussian kernel function, combined with the performance of global kernel function and local kernel function. A new mixed kernel function is constructed:

$$K_{mix} = t \exp \left( -\frac{||x-y||^2}{2\sigma^2} \right) + (1-t)\left[ (x \cdot y) + c \right]^p$$

The mixed kernel function comprises an effective feasible Gaussian kernel function and the polynomial kernel function. The new mixed kernel function is beneficial to extract the training sample taking into account the global characteristics of the training samples, thereby theoretically establishing a more stable SVM model.

After constructing a new hybrid kernel function, the SVM's ability depends mainly on the penalty factor, parameters of the mixed kernel function, and weight coefficient. These parameters control the search of the VC vector of the feature space after the feature vector is mapped by the SVM algorithm. Therefore, how to make these three parameters achieve the most matching value is the key to improving the SVM algorithm before the need to analyze these parameters.

In the nonlinear SVM the following formula is used in addressing the optimization problem:

$$\begin{align*}
\min_{w,b} & \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \\
\text{s.t.} & \quad y_i \left( \langle w \cdot \phi(x_i) \rangle + b \right) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, n
\end{align*}$$

$\xi_i$ is the loss function, and the value of the penalty factor $C$ determines the penalty generated by the empirical error. A lower $C$ value reduces the complexity of SVM, indicating that the empirical risk of SVM is increased. However, a larger $C$ value generates the maximum threshold allowed by the learning feature subspace, indicating that the SVM's generalization and learning abilities can be improved by the fact that the SVM cannot increase the learning complexity. A balance needs to be maintained between the two penalty factors for a better performance of the SVM classifier.

The mixed kernel function includes two parameters. One is the radial radius $\sigma$ in the Gaussian kernel function, and the other is the order $p$ in the polynomial kernel function. If $\sigma \to 0$, then the training text is all support vector, which helps in improving the accuracy of classification, but at the same time is generated over the fitting situation; if $\sigma \to \infty$, then all the training text completely belongs to a subject class that meets the purpose of generalization, but has low accuracy. For the order $p$ in the
polynomial kernel function, if the number of training samples is large, it leads to huge spatial dimension, resulting in the occurrence of dimension disaster.

The new hybrid kernel function constructed in this study also has an important parameter, that is, the weight coefficient \( t \), which ranges from 0 to 1. It is obvious from the formula that if \( t \to 1 \), then the Gaussian kernel function contributes more to the whole SVM algorithm model, whereas if \( t \to 0 \), then the overall generalization advantage of the polynomial kernel function is more obvious in the SVM algorithm model. Therefore, it is also important to select the value of the optimal mixed kernel function weight \( t \) while improving the SVM algorithm, so that the best performance of the constructed new mixed kernel function is achieved in the classification process.

A key step in the improvement of SVM algorithm is the selection of SVM parameters. The traditional method is the grid search. This study proposes a cross-validation method of genetic algorithm (GA). The essence of the problem is to show the potential of an individual and then use some of the genes encoded by the individual to form a population. The chromosome is the main carrier of genetic material, with a collection of multiple genes [12].

At the beginning of the GA, \( N \) initial populations are randomly generated. All the initial populations exist in the data structure of the gene string type, and then the iteration is carried out from the \( N \) string populations. Normally, the range of \( N \) is \([20,100]\). The next step is to calculate the fitness of each individual in the population, which represents the performance of each individual. The maximum number of evolutionary iterations is set to \( T \), and the current number of evolutionary iterations is \( t \). If \( t \leq T \), then \( t \) is incremented and the following process is continued until \( t > T \); the result is the optimal solution. Replication in the algorithmic process is to extract good individuals from the current population to the next generation, while others are passed to the next generation through crossover and mutation operations, where the extraction criteria are the degree of individual adaptation previously calculated. The crossover operation is to randomly transform the chromosomes of all the individuals within the population and then generate new individuals; the newly generated individuals are individuals who have inherited the previous generation of good genes. The last variation of this process is done based on genetic mutations that produce new individuals by mutation.

In this study, GA was used to optimize the SVM hybrid kernel function parameters as follows:

1. Randomly generate \( N \) initial populations and use binary coding, mixed kernel function parameters, weight coefficient, and penalty factor to be encoded, resulting in chromosome gene string.
2. Encoded gene string is decoded into corresponding parameters and then is inputted into the SVM to train the learning.
3. The accuracy of training set is determined in the cross-validation as a fitness function to calculate the degree of adaptation of each individual in the group.
4. The conditions are satisfied to find the best combination of SVM parameters, or the process is continue to the following steps.
5. Another individual is replaced according to the rules of optimal individual replication.
6. A crossover probability is selected to cross the generated new individual.
7. A mutation probability is selected to mutate the new individual and go back to step (2) to continue the loop.

4. Simulation experiments
In this experiment, Matlab was used as simulation tool to introduce LIBSVM open-source software package to build and train the SVM model. The software is used to integrate the kernel function, including Gaussian kernel function and polynomial kernel function. Search optimization algorithm is also integrated. For the newly constructed mixed kernel function, LIBSVM supports custom kernel functions. Therefore, a code can be added to implement the new kernel function in the package, and then the kernel function selection parameter can be set to be customizable. LIBSVM package also integrates the GA function. Therefore, the GA can be used for optimizing new kernel function parameters.

The data used in the simulation experiment was obtained using the WebCrawler; 450 web page texts were selected, including military, scientific and technological information, health, and other eight
categories. According to the category of mixed choice, 300 pages of text were used as training samples, and the remaining 150 were test samples. All the text data must first be processed by the data preprocessing module. If the weight difference in the processing of the data is large, it adversely affects the speed and prediction accuracy of the learning and training, making the data normalization indispensable. The TF-IDF algorithm itself is a normalized calculation method. Therefore, all the processed data was normalized in the range of 0 to 1.

In addition, the parameter \( c = 0 \) of the new mixed kernel function of formula 3.17 was set. That is, the homogeneous polynomial kernel function was taken and then the parameter of the Gaussian kernel function was set to \( r = \frac{1}{2 \sigma^2} \). Consequently, four parameters needed to be optimized: the penalty factor \( C \), Gaussian kernel function parameter \( r \), polynomial order \( p \), and weight coefficient \( t \). Binary coding was used to set the initial population number \( N = 50 \). The selection operator value was 0.9, the crossover probability was set to 0.45, and the mutation probability was set to 0.1. In this study, the precision value of 10 was chosen as the fitness function of GA. The number of iterations is \( T = 50 \). The cross-validation of GA was used to select the best parameter change process, as shown in figure 3.

![Figure 3. GA is optimized for hybrid kernel parameters.](image)

The figure clearly shows that the number of iterations reached 25 times. Also, the value of the penalty factor did not converge, the remaining three parameters reached convergence, and the best fitness function cross-validation accuracy reached 100%. Therefore, the value of the penalty factor of the 26th iteration was taken for the best parameter value. The result of the optimization was \( C = 21.05379 \), \( r = 0.03562 \), \( t = 0.81658 \), and \( p = 1.50175 \). All the optimal parameter values were classified into the new kernel function classification model. The correct rate was 98.67%.

The Gaussian kernel function and the polynomial kernel function were integrated in the LIBSVM package. The type of SVM kernel function in LIBSVM was set by the parameter \( s \). The default value was 0, which represented the linear kernel function. In the experiment, the values were set to 1 and 2, each representing a polynomial kernel function and a Gaussian kernel function, respectively. The grid search algorithm function SVMcgForClass was used to optimize the parameters. The comparison of the optimal GA parameters with the correct rate results is shown in table 1.

| Parameter | Gaussian kernel function | Polynomial kernel function | Mixed kernel function |
|-----------|--------------------------|---------------------------|----------------------|
| Penalty factor \( C \) | 3.41328                  | 3.41328                   | 21.05379             |
| \( \delta \) | 1.5374                   | 1.5374                    | 0.03562              |

Table 1. Comparison of the optimal parameters with the correct rate results.
5. Conclusions
This study explored the SVM algorithm that could deal with the large-scale data in the web effectively. After studying and analyzing the kernel function used in the commonly used SVM, a new mixed kernel function was constructed by combining the global kernel function and local kernel function with different characteristics, and then the optimal solution of the parameter was found through the GA. Finally, the influence of the aforementioned two improvements on the classification performance of web pages was verified by experimental simulation. Experiments showed that the proposed method improved the classification performance of web pages in most cases.

References
[1] Vapnik V N 1995 The nature of statistical learning theory IEEE Trans. Neural Netw. 8(6) 1564
[2] Jin-shu Su, Bo-feng Zhang, Xin Xu 2006 Advances in machine learning based text categorization J. Software 17(9) 1848-59
[3] Jian-hui Wang, Hong-wei Wang, Yun-fa Hu 2005 A simple and efficient algorithm to classify a large scale of texts J. Comput. Res. Develop. 42(1) 85-93
[4] Hua Zhao, Tie-jun Zhao, Shu Zhang 2006 Topic detection research based on content analysis J. Harbin Inst. Technol. 10(38) 1740-49
[5] C. Hsu, C. Lin. 2002 A comparison on methods for multi-class support vector machines IEEE Trans. Neural Netw. 13 415-25
[6] Jian-fen Wang, Yuan-da Cao 2001 The application of support vector machine in classifying large number of catalogs J. Beijing Inst. Technol. 21(2) 225-29
[7] X Z Li, J M Kong 2014 Application of GA-SVM method with parameter optimization for landslide development prediction Nat. Hazards Earth Syst. Sci. 14(3) 525-33
[8] Xing-hua Fan, Mao-song Sun 2006 A high performance two-class Chinese text categorization method Chin. J. Software 29(1) 124-31
[9] C H Wu, G H Tzeng, R H Ling 2009 A novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression Expert Syst. Appl. 36(3) 4725-35
[10] Sebastiani F. 2002 Machine learning in automated text categorization Acm Comput. Surveys 34(1) 1-47
[11] Jian-tao Sun, Dou Shen, Yu-chang Lu 2004 Web document classification techniques J. Tsinghua University: Sci. Technol. 44(1) 65-68
[12] Qing-jie Liu, Gui-ming Chen, Xiao-fang Liu 2012 Genetic algorithm based svm parameter composition optimization Comput. Appl. Software 29(4) 94-1