IJAMML 7:1-2 (2017) 13-22 September and December 2017
ISSN: 2394-2258
Available at http://scientificadvances.co.in
DOI: http://dx.doi.org/10.18642/ijamml_7100121908

# COMPARATIVE STUDY OF THREE KERNEL PARAMETER OPTIMIZATION METHODS OF SUPPORT VECTOR MACHINES BASED ON PRACTICAL EXAMPLES

**Yan Deng**

School of Science, China University of Geosciences, Beijing, P. R. China

## Abstract

Based on practical examples, three kernel parameter optimization methods of Support Vector Machines (SVM), i.e., grid method, particle swarm optimization, and genetic algorithm, are compared. The results are as follows: (1) Grid method lasts the shortest time in optimization process, and as the number of samples increases, it shows more obvious advantages. (2) Optimization results of particle swarm optimization and genetic algorithm are both unstable, and take longer time on average. Due to their wider search scope, it is expected that more suitable kernel parameters can be obtained through many experiments. (3) With the increasing of the number of samples, kernel parameter optimization time will increase sharply. For large sample classification model, it is necessary to explore new SVM training algorithm for large-scale data set.

*Corresponding author.
*E-mail address*: dengyan@cugb.edu.cn (Yan Deng).


2010 Mathematics Subject Classification: TP181.
Submitted by Jianqiang Gao.
The project is supported by the Fundamental Research Funds for the Central Universities.
Received November 10, 2017

*Keywords*: support vector machines (SVM), kernel parameter; parameter optimization.

---

## 1. Introduction

Support vector machines are a new machine learning algorithm, based on VC dimension theory and structural risk minimization principle [1]. Due to the better generalization ability, and unique advantages in solving some machine learning problems such as small sample, nonlinear and high-dimensional, SVM have been widely used in pattern recognition, regression estimation, probability density function estimation and other fields. At present, the research on SVM classification mainly focuses on kernel function and kernel parameter selection. In this paper, three common kernel parameter selection methods of SVM, i.e., grid method, particle swarm optimization algorithm, and genetic algorithm, are selected. The advantages and disadvantages of three kernel parameter selection methods are pointed out through detailed analysis of their parameters optimization and the comparison of instance data, trying to provide references for the selection of SVM in future.

## 2. The Principle of SVM

Given a set of linearly separable points $(x_i, y_i)$, where $i = 1, 2, \cdots, n$, $x_i \in R^m$, and the $y_i \in \{-1, 1\}$ as their labels. The basic idea of SVM is to find an optimal separating hyperplane,

$$g(x) = w^T \cdot x - b = 0,$$

which can accurately separate the two classes of points that have been linearly separated. And that is guaranteed to the maximal margin, which means to the strongest generalization ability. Then the question of

finding the optimal hyperplane turns into the settlement of the following optimization problem, that is,

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

$$\text{s.t. } y_i(w^T x_i + b) \geq 1,\ i = 1, 2, \cdots, n.$$

To solve the problem above, the Lagrange multiplier $\alpha_i$ is introduced. Substitute this optimization problem for the Wolfe dual problem, that is,

$$\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i x_j)$$

$$\text{s.t.} \quad \alpha_i \geq 0,\ i = 1, 2, \cdots, n,$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0.$$

The problem is a convex quadratic programming problem, and exists the only global optimal solution. Then, the decision function of the optimal separating hyperplane is,

$$f(x) = sgn[\sum_{i=1}^{n} \alpha_i y_i (x \cdot x_i) + b].$$

By KKT conditions, only a small number of $\alpha_i$ are not zero in decision function, and correspond to the support vector.

For the linear inseparable cases, use a nonlinear mapping $\varphi(\cdot)$ to map the sample space into the high dimensional feature space and then gain the optimal separating hyperplane in this new feature space. In this case, the decision function is,

$$f(x) = sgn[\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b].$$

In this formula, the kernel function $K(x, x_i) = (\varphi(x) \cdot \varphi(x_i))$ meets the Mercer conditions, and different kernel functions correspond to different

SVM. In this paper, the radial basis kernel function $K(x, x_i) = \exp(-g \cdot \|x - x_i\|^2)$ is used and the kernel parameters to be determined are the kernel function parameter $g$ and the penalty parameter $c$ [2-4].

## 3. Kernel Parameter Optimization Methods of SVM

There is no international conclusion about the optimization selection of the parameters of SVM now. The common method is to select $c$, $g$ within a certain range, take the training sets as the original data sets, obtain the validated classification accuracy of the training sets in this $c$, $g$ group by using the K-CV method, and select the $c$, $g$ with the highest classification accuracy as the best parameter.

### A. Grid method

Grid method is to specify the value ranges of the parameter $c$, $g$, make the value of $c$, $g$ discrete grid and then find within the grid according to a certain step. This is the simplest and the most direct method of finding suitable parameters. We can conduct a rough search in a wide range and then a detailed one at the optimal value [5].

### B. Particle swarm optimization (PSO)

The basic idea of particle swarm optimization (PSO) stems from the simulation of predatory behaviour of birds. PSO compares the search space of problems to the flight space of birds, and each bird is abstracted into a particle without mass and size. According to the idea of cooperation and competition of individuals among populations, the entire movements of particle swarm show the characteristics of bird foraging. Let the scale of particle swarm be $m$, the position vector of each particle in $n$-dimensional space be $x_i = (x_{i1}, x_{i2}, \cdots, x_{in})$, $i = 1, 2, \cdots, m$ and the velocity vector of each particle be $\nu_i = (\nu_{i1}, \nu_{i2}, \cdots, \nu_{in})$. Each particle adjusts its flight path and approaches the optimal point according to its own flight experience and group flight experience. Assume the optimal

position of each particle is $P_i = (P_{i1}, P_{i2}, \cdots, P_{in})$ and the optimal position of whole particle swarm is $P_g = (P_{g1}, P_{g2}, \cdots, P_{gn})$. Then, each particle can update its speed and location according to Equation (1):

$$v_{i+1} = wv_i + c_1 r_1 (P_i - x_i) + c_2 r_2 (P_g - x_i), \quad (1)$$

$$x_{i+1} = x_i + v_{i+1}. \quad (2)$$

Here, $w$ is the function of inertia weight; $c_1$ and $c_2$ are learning factors; $r_1$ and $r_2$ are random numbers in the interval $(0, 1)$.

PSO algorithm towards the process of the SVM parameter optimization is as follows:

(1) PSO initialization. A random set $\{c, g\}$ is used as the initial position of particles. Set the number of populations as 20, the maximum number of iterations as 200, the acceleration factor as $c_1 = 1.5$, $c_2 = 1.7$.

(2) For each particle, conduct SVM training according to current $\{c, g\}$ and take the classification accuracy of the SVM model through the cross validation method as the fitness of each particle.

(3) Update the optimal fitness of the particle itself and its population $P_i$ and $P_{g'}$ and adjust the velocity and position of the particle by the PSO optimization equations (1) and (2) to get a new particle position, that is to get a new SVM parameter value $\{c, g\}$.

(4) If not meet the maximum number of iterations or the ending conditions, return to step (2), or output the optimal parameter values.

(5) Use the final optimal parameters to retrain SVM classifier, build PSO-SVM model and use test sample set to perform generalization ability testing [5].

**C. Genetic algorithm (GA)**

Genetic algorithm (GA) is a randomized search method based on the principle of natural selection and natural genetic mechanism in biology. It simulates the natural process of gene recombination and evolution and encodes the parameters of the problem to be solved, that is to become genes [5].

Several genes form a chromosome, and then many chromosomes perform the operations like natural selection, cross pairing and mutation. The final optimization results are obtained after repeated iterations (i.e., inheritance of generations). GA is a kind of global search optimization method, which has the characteristics of simple and universal, strong robustness, and high optimization efficiency. The main factors of GA are: coding of parameters, design of fitness function, setting of algorithm control parameters, and processing of constraints. The optimization process of GA for SVM parameters is as follows:

(1) Parameter encoding and initialization. Parameter $g$ and $c$ use binary encoding. Set the size of group as 20, terminate algebra as 100. Generate initial values of a set of $\{c, g\}$ randomly, and then constitute initial groups.

(2) For each individual, conduct SVM training according to current $\{c, g\}$ and take the classification accuracy of the SVM model through the cross validation method as the fitness of each individual.

(3) Generate a new set of chromosomes by selecting, crossing, mutation and other genetic operator operations, that is to obtain new SVM parameter values.

(4) If not meet the maximum number of iterations or the ending conditions, return to step (2), or output the optimal parameter values.

(5) Use the final optimal parameters to retrain SVM classifier, build GA-SVM model and use test sample set to perform generalization ability testing.

## 4. Case Analysis

In order to compare the respective characteristics of three different optimization methods, data sets of speech signal features in *Thirty cases of MATLAB neural network analysis* [5] are selected as instance data. Extract 24-dimensional speech signal respectively from four different kinds of music, folk songs, Guzheng, rock and pop music and classify these four effectively by SVM. Before the classification testing, optimize the speech features by using principal component analysis method. Select 17 speech features as input variables, randomly choose training samples and test samples, normalize the speech features and use the libsvm toolbox. Respectively use three parameters optimization methods, grid method, particle swarm optimization, and genetic algorithm, and then classify.

The first group: select 100 sets of data as a training set, 40 sets of data as a testing set. Each method runs 10 times. The results of running time and classification accuracy of parameter optimization are shown in Table 1.

**Table 1.** Comparison of parameter optimization results based on 140 data sets

| | Grid method | PSO | GA |
|---|---|---|---|
| Average running time (s) | 9.2 | 48.9 | 63.9 |
| Average classification accuracy (%) | 88.9 | 87.3 | 90.5 |
| The shortest time of a single run (s) | 8.9 | 46.8 | 18.6 |
| The longest time of a single run (s) | 9.6 | 51.1 | 212.2 |
| The lowest classification accuracy (%) | 82.5 | 80.0 | 82.5 |
| The highest classification accuracy (%) | 92.5 | 93.75 | 95 |

The second group: select 200 sets of data as a training set, 80 sets of data as a testing set. Each method runs 10 times. The results of running time and classification accuracy of parameter optimization are shown in Table 2.

**Table 2.** Comparison of parameter optimization results based on 280 data sets

| | Grid method | PSO | GA |
|---|---|---|---|
| Average running time (s) | 35.0 | 185.0 | 163.0 |
| Average classification accuracy (%) | 89.2 | 90.8 | 88.5 |
| The shortest time of a single run (s) | 32.5 | 176.4 | 76.5 |
| The longest time of a single run (s) | 37.4 | 199.0 | 479.7 |
| The lowest classification accuracy (%) | 85.0 | 85.0 | 83.75 |
| The highest classification accuracy (%) | 92.5 | 95.0 | 93.75 |

As is shown in Table 1 and Table 2, the training data of the second group is twice larger than the first group, but their classification accuracy are not significantly different, which is in line with the characteristics of SVM for small sample models. Judging from the average classification accuracy of the two groups, the three methods have no significant difference and are relatively close.

As for the average running time, grid method is relatively less time-consuming, much better than the other two algorithms, especially when the number of samples increases, its advantage in time becomes more obvious. It is difficult to compare PSO and GA, because the two results are contrary. In terms of the longest and shortest times of a single run of the 10 tests, grid method is the shortest, and grid method and PSO are relatively stable while GA fluctuates widely. With the same number of samples, the optimization time of GA can even differ by more than 10 times. The results also show that when the number of samples is large, the time to find suitable parameters increases sharply. For the classification model of large samples, the SVM training algorithm for large scale data sets are needed.

Judging from the highest and lowest classification accuracy of the 10 tests, the lowest classification accuracy is improved when the training samples are large in number. But the highest is not necessarily high. The results of the grid method are relatively stable while that of PSO and GA

fluctuate severely, which is characteristic of all the randomized methods. And the highest classification accuracy of PSO and GA are higher than that of the grid method.

## 5. Conclusions and Discussions

(1) Among the three methods, grid method takes the shortest time, especially when the number of samples increases, its advantage in the running time becomes more obvious. To obtain a better training model, we can conduct a large-scale grid search to select multiple sets of optimal parameters first, and then a fine search of small steps around the obtained optimal value. In addition to step size, grid parameter optimization method doesn't need to select other parameters and its effect is more stable.

(2) PSO and GA are both random methods. In view of their wide search range, it is expected to get more suitable kernel parameters through multiple tests. However, the results of randomized methods are unstable and the average running time is longer. Many optional parameters as well as the number of iterations and the populations will affect parameter optimization, which is more flexible and brings some difficulties in practice.

(3) As can be seen from the test results, when the number of samples is large, the time to find suitable parameters increases sharply. For the classification model of large samples, the SVM training algorithm for large scale data sets are needed. This is also an international hot topic in SVM study now.

## References

[1] X. G. Zhang, Introduction to statistical learning theory and support vector machines, Acta Automatica SINICA 126 (2000), 32-42.

[2] E. Amid, S. R. Aghdam and H. Amindavar, Proc. World Acad. Sci. Eng. Tech. 67 (2012), 1303-1307.

[3] J. Chen and G. R. Ji, The 2nd International Conference on Computer and Automation Engineering, IEEE, Singapore (2010), 242-246.

[4] V. N. Vapnik, The nature of statistical learning theory, V. N. Vapnik Ed., New York: Springer Verlag (2000), 193-245.

[5] F. Shi, X. CH. Wang, L. Yu et al., Thirty Cases of MATLAB Neural Network Analysis, Beijing: Press of Beijing University of Aeronautics and Astronautics, 2010.

■