

An Experimental Study of the Hyper-parameters Distribution Region and Its Optimization Method for Support Vector Machine with Gaussian Kernel

Zhigang Yan^{1,2}, Yuanxuan Yang^{1,2} and Yunjing Ding^{1,2}

¹ School of Environmental Science and Spatial Informatics, China University of Mining and Technology, Xuzhou City, Jiangsu Province 221116, P.R. China

² Jiangsu Key Laboratory of Resource and Environmental Information Engineering, China University of Mining and Technology, Xuzhou City, Jiangsu Province, 221116, P.R.China

Corresponding author: Zhigang Yan, zhg-yan@163.com

Abstract

Support vector machine (SVM) is a kind of machine learning method, but the selection of parameters has important effects on the generalization ability of SVMs. In this study, the relation between the error penalty parameter C , kernel parameter σ and the generalization ability of SVMs is discussed. Parameter C adjusts the similarity among within-class members, while parameter σ adjusts the similarity between classes. Moreover, C and σ balances each other mutually within a certain range, which forms a fan-shaped optional parameter distribution region. The optimal parameter area should be located near the center of the sector where both C and σ are small. According to this, a method is suggested to first search a suitable area with coarse grids, and then determine the optimal parameter within the area with a fine bilinear grid. Experimental results show that the new parameter selection method can not only avoid local optima, and thus excluding the cases in which C and σ are big and unstable, but also can be extremely fast in searching process. Compared with other parameter selection methods, the performance of SVMs cannot be influenced, or even better in some cases.

Keywords: SVM, Gaussian Kernel, Generalization Ability, Bilinear Grid Search Method

1. Introduction

SVM is a kind of prevalent machine learning method developed by Vapnik and his colleagues [1]. The problem of parameter selection has always been one of hot issues in theories and applications of SVM. The existing techniques for adjusting the parameters in SVMs can be summarized into two kinds: one is based on analytical techniques; the other is based on heuristic searches. The first kind of techniques determines the parameters with gradients of some generalized error measures [2-6]. Keerthi, S. *et al.*, introduced an efficient method for gradient-based adaptation of hyper-parameters in SVM models [2]. Chapelle *et al.*, presented a gradient descent algorithm to select parameters which can automatically tuning multiple parameters for SVMs [3]. N.E. Ayat *et al.*, introduced a new SVM model selection methodology based on the minimization of an empirical error estimate [4]. And the second kind of techniques determines the hyper-parameters with modern heuristic algorithms. Recently, many machine learning practitioners use optimization-based methods to speed up the search process in parameter selection. For example, Liang-Hsuan Chen *et al.*, proposed a

genetic algorithm (GA) based parameter selection method which uses different estimators of generalization ability as the fitness function [7]. Pai suggested a parameter selection method based on simulated annealing (SA) [8]. These methods can not only shorten calculation time, but also can reduce the dependence on initial parameter values. However, GA and SA are both complex. Peng proposed a selection method based on particle swarm algorithm (PSO) [9], which is easy to implement and has relatively high optimization efficiency. However, the drawback of PSO is that it may get trapped in local optima. Therefore, for multimodal (multiple local optima exist) SVM parameter selection problems, the optimization efficiency of PSO may decrease.

According to the analysis of the methods above, until now, most SVM users select parameters empirically by trying a finite number of parameter values and keeping those whose testing error is least. This procedure requires a grid search over the range of the parameter values and needs to locate the interval of feasible solution and a suitable sampling step. It is a hard work since a suitable sampling step and the grid interval may not be easy to determine without prior knowledge for the parameter selection. So, fast determining the optimal range of the SVM parameters has great significance and practical value. It can not only avoid getting trapped in local minima, but also reduce the searching time. In this paper, the parameter selection problem of SVMs with Gaussian Kernel $K(x, y) = \exp(-\|x - y\|^2 / 2\sigma^2)$ is discussed. The relation between the error penalty parameter C , kernel parameter σ and the performance of SVM is analyzed. The optimal ranges of C and σ are given. Furthermore, a new parameter selection method is proposed based on the analysis.

2. The Relation between the Performance of SVMs with Gaussian Kernels and Parameter C and σ

The parameters that can affect the performance of SVMs with Gaussian kernel are error penalty parameter C and kernel parameter σ . So far, the exact conclusion for the relation between parameters (C , σ) and the generalization ability of SVMs with Gaussian kernel is given in reference [10]. For a two-class SVM, let the sample numbers of each class be l_1 and l_2 , respectively, where $l_1 > l_2 + 1 > 2$, no completely identical samples exist. Assume the class with a larger sample number to be the positive class, and the other to be negative class. Large amounts of experiments in reference [10] have shown that for different values of (C, σ^2) , there are several possible cases, which are demonstrated in Figure 1.

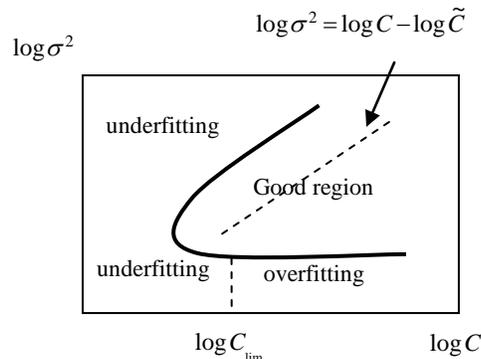


Figure 1. A rough Boundary Curve of the Region of (C, σ^2) according to their Performance

(1) Fixing σ^2 , $C \rightarrow 0$, then we have $C < \bar{C}$, $f(x) > 0, \forall x \in X$. That is, all samples are assigned to positive class and the SVM is underfitting;

(2) Fixing σ^2 , $C \rightarrow \infty$, then the SVM is overfitting, f tends to be a hard margin classifier;

(3) Fixing C , $\sigma^2 \rightarrow 0$, then we have the kernel function $k(x, y) \rightarrow 0$, for $C \geq C_{\text{lim}}/2$. All training samples are correctly classified, which leads to overfitting; for $C < C_{\text{lim}}/2$, f will assign the whole feature space to positive, which is underfitting;

(4) Fixing C , $\sigma^2 \rightarrow \infty$, then we have the kernel function $k(x, y) \rightarrow 1$. f will assign the whole feature space to positive, which is underfitting;

(5) When (C, σ^2) interacts, there is an optimal region. The combination of (C, σ^2) which has the best generalization ability will concentrate on the adjacency of line $\log(\sigma^2) = \log(C) - \log(\bar{C})$. When $(C, \sigma^2) \rightarrow \infty$, $\bar{C} = C/\sigma^2$. f tends to be a linear SVM, with \bar{C} being the best parameter selection.

In reference [10], the conclusion was drawn for the case in which C and σ^2 approach infinity. However there was no discussions on how C and σ^2 affect the generalization ability of SVMs when C and σ^2 are finite. In this study, we try to discuss this issue.

2.1. The Relation between Error Penalty Parameter C and Generalization Ability of SVMs

The function of the error penalty parameter C is to adjust the ratio of the learning machine's confidence interval and empirical risk, in order to give the learning machine the best generalization ability. Different data subspaces lead to different choices of C . When samples are completely separable, the value of C does not affect the generalization ability of an SVM; while the value of C has major impacts on the performance of an SVM when samples are not completely separable. In established data subspaces, small values of C indicate small penalty for empirical errors, and low complexity of a learning machine, while empirical risk is high. On the contrary, the complexity of a learning machine is high and empirical risk is low. The former is known as "underfitting", while the latter "overfitting". In each data subspace, there is at least one appropriate C which gives the best generalization ability of an SVM. A value of C either too big or too small can affect the generalization ability of an SVM. When C exceeds a certain value, the complexity of the SVM approaches the maximum value allowed by the data subspace. In this situation, the empirical risk and the generalization ability of the SVM cannot vary.

2.2. The Relation between the Kernel Parameter σ and the Generalization Ability of SVMs

In reference [11], the conclusion about the effect of Gaussian kernel parameter σ on the performance of SVMs was proved based on the KKT conditions in solving SVMs:

(1) Fixing C , when $\sigma \rightarrow 0$, all training samples are support vectors and are correctly classified. However the generalization ability of the SVM is 0, because the SVM is severe overfitting.

(2) Fixing C , when $\sigma \rightarrow \infty$, the learning ability of the SVM is almost 0, which means that it assigns all samples to the same class and it has no judgment on new samples. That is i severe

underfitting. It should be mentioned that the above conclusion is based on the precondition assuming that no completely identical samples exist.

In fact, when σ is small enough, the SVM will show overfitting performance; when σ is big enough, the SVM will show underfitting performance. With the increasing of σ , the empirical risk increases, indicating that the dimensionality of the data subspace is reducing. Furthermore, in the subspace, the VC dimensionality of the most complex SVM which can be constructed is reducing. The generalization ability is improving as well, and the prediction accuracy approaches the best. The optimal σ value is obtained at this moment. If σ is still increasing, the learning ability of the SVM will decrease and the empirical risk will increase. Consequently, the prediction ability of the SVM will decrease because of underfitting. Moreover, when σ approaches to a certain big level, the SVM almost loses its learning ability.

2.3. Rethinking of the Relation between (C , σ) and the Generalization Ability of SVMs

In fact, C and σ have a combined effect on the generalization ability of SVMs. It is difficult to construct an SVM with good generalization ability by fixing one of the parameters and adjusting the other one. Next, we try to show how C and σ affect the performances of SVMs based on the intrinsic features of Gaussian kernel and SVMs.

The essential of classification problems is a matter of similarity. The solution of a classification problem depends on the estimation of similarity. The similarity between two samples is described by the inner product of the two samples. The decision function of an SVM depends on the inner products of the input space, or the high dimensional Hilbert space which is transformed from the original space. These inner products are strongly dependent on the selection of mappings. Different choices of mappings mean different evaluation criterions of sample similarities. For SVMs, inner products in high dimensional spaces are calculated by kernel functions. Therefore, the selection of kernel functions and kernel parameters is just the selection of mappings and, thereby, the selection of similarity evaluation criterions.

For Gaussian kernel functions, we consider them as the normalization results of similarity measurements. Assuming no completely identical samples, then $k(x, y) \in (0, 1)$. If $k(x, y)$ is closed to 1, the similarity degree of two samples is high; while if $k(x, y)$ is closed to 0, the similarity degree is low. The parameter σ is the adjustment factor of similarity range. The bigger the parameter σ is, the wider the similarity range is and the looser the similarity conditions are. The two samples can be similar in a relatively big range and the decision boundary is weakened in this situation. On the contrary, when the parameter σ goes smaller, the situation is just the opposite. When σ approaches infinity, the similarity range is unlimited and samples can be unconditionally similar, which makes, the decision boundary disappears and the number of support vectors to be 0. As a result, all samples are assigned to the same class, and the SVM has no learning ability or generalization ability. When σ approaches 0, the similarity range approaches 0, and the similarity conditions are tight, this makes all samples dissimilar. And all of them are considered as support vectors and are thus correctly classified. As a result, the SVM has no generalization ability because of overfitting.

It is clear that the conclusion drawn by evaluating the similarities between samples based on Gaussian kernel parameter is the same as in reference [10, 11], but it is more concise and intuitive. Furthermore, the function of adjusting the range of sample similarity is recognized for σ .

For two classes which are not completely separable, the value of C may affect the classification interval and thereby the misclassification rate. In other words, the value of C also adjusts the similarity between classes. The bigger the values of C are, the more the decision boundaries tighten; the stronger the similarity conditions are, the stronger the

discrimination between the classes is; on the contrary, the smaller the values of C are, the more the decision boundaries loosen; the weaker the similarity conditions are; the weaker the discrimination between the classes is.

Therefore, we believe that C globally adjusts the similarity between classes; while σ directly adjusts the similarities between individual samples. C and σ balance each other in a certain degree and they coordinately affect the generalization ability of an SVM.

2.4. The Optimal Ranges of (C, σ)

It can be known from the above that small values of C often result in underfitting; with the increasing of C , the generalization ability of SVMs increases. However when C exceeds a certain level, SVMs are easy to result in overfitting. At this stage, if σ^2 increases, the effect of increasing C will be compensated. If σ^2 still increases, however, the SVMs will be underfitting again. Therefore, the balance of (C, σ^2) forms a parameter alternative region with a sector shape which is referred as "sector regions" in this paper and shown in Figure 2.

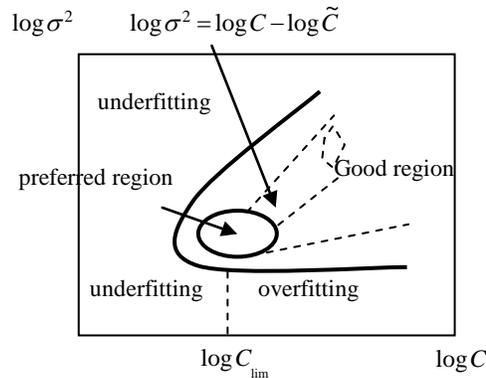


Figure 2. Reasonable Regions of (C, σ^2) for Selecting

It is difficult to accurately locate the optimal ranges of (C, σ^2) due to the differences between sample sizes, dimensionalities, spatial distributions, *etc.* In this study, we believe that a reasonable range of (C, σ^2) should be located within a region that makes testing accuracy high and it should have a continuous range excluding local "prominent regions". Moreover, both C and σ^2 have moderate values near the "sector regions", neither too high, nor too low, which avoids "underfitting", "overfitting" and unstable regions. Additionally, a reasonable pair of (C, σ^2) should near the "sector center", where both C and σ^2 are relatively small. The criterion of "reasonable regions" is not only based on prediction accuracy, but also based on stability, the difficulties of solution and search speed. It is also possible to find some "relatively good regions" far away from the "sector center", even a few "prominent regions" with prominent performances, due to the balance effect of (C, σ^2) . However, they are unstable, and the solution is difficult.

3. The Improvements of Parameter Selection Method for SVMs with Gaussian Kernel

There are many selection methods for (C, σ^2) . The commonly used methods are grid searching method and bilinear method. In grid searching method, N and M values are taken

for C and σ^2 , respectively. Then the optimal parameters are chosen from the $N \times M$ pairs of (C, σ^2) according to their generalization abilities. In the bilinear method, the optimal parameters are selected based on the different generalization abilities of the regions of (C, σ^2) , which is shown in Figure 1 [10]. The steps of the latter algorithm are as follows:

(1) Solving the optimal parameter \tilde{C} of the linear SVM, this results in the best generalization ability of the linear SVM.

(2) Fix \tilde{C} for the SVM with Gaussian kernel, and train the SVM for (C, σ^2) which satisfy $\log(\sigma^2) = \log(C) - \log(\tilde{C})$. Select optimal (C, σ^2) according to the generalization abilities.

The grid searching method can obtain a higher accuracy, however the computational complexity is relatively high; while the bilinear method has lower computational complexity, and the prediction accuracy is also relatively low. Because the bilinear method depends much on optimal parameter \tilde{C} . In reference [12], the bilinear method was improved by setting the initial values to $0.5\tilde{C}$ and $2\tilde{C}$, and the bilinear grid searching method was proposed. The bilinear grid searching method can achieve high prediction accuracy and reduce the computation load at the same time. However there are still drawbacks, that is, (C, σ^2) can easily get trapped in "local optima". We will try to improve this method according to the previous analysis in the following passage.

For a given data set, a relatively big parameter range is needed. However a big range increases the uncertainty of parameter selection and the parameters are prone to fall into "local optima" which causes unstable performances. To avoid this situation, we propose to determine the reasonable range of (C, σ^2) firstly, then use bilinear grid searching method. The steps are:

(1) Set a relatively big range and a relatively big step for (C, σ^2) , and use grid searching method. Obtain the contour map of prediction accuracy by using 5-fold cross validation to evaluate the generalization abilities of SVMs. Then narrow the range of (C, σ^2) in order to encircle the "reasonable region", according to the conclusion of our study that the "reasonable region" is located at the "sector center".

(2) Use bilinear grid searching method in the refined ranges to determine the final optimal parameters (C, σ^2) .

To demonstrate the proposed method, we have carried out experiments based on LibSVM software toolkit [13]. Haberman, Glass, Iris and Lenses data sets in UCI database were chosen for our experiments. The 2nd and 3rd classes of Iris data set were selected, while multiple classes of other data sets were selected. 1-V-1 SVM was used for multiple-class cases. The values of (C, σ^2) were set to $\log_2 C \in [-5, -4, \dots, 14, 15]$, $\log_2 \sigma^2 \in [-10, -9, \dots, 9, 10]$. Five-fold cross validation was used for the evaluation of generalization ability. Experimental results showed that the characters of parameters selection processes for all data sets coincide with the conclusion in our study. That is, the preferred (C, σ^2) locate at "sector center", where the ranges of (C, σ^2) were more stable, and the prediction accuracies were higher. Detailed results are tabulated in Table. 1.

Table 1. The Comparison for (C, σ^2) Selected

Data sets	Bilinear Grid Searching	Our method
Iris 2 vs 3	$(2^4, 2^4) 0.97000$	$(2^{3.7}, 2^{3.7}) 0.97000$
Glass	$(2^{11.9}, 2^{4.5}) 0.74908$	$(2^3, 2^{-0.9}) 0.74475$
Haberman	$(2^8, 2^{9.5}) 0.75182$	$(2^{-1}, 2^4) 0.76150$
Lenses	$(2^5, 2^{-1}) 0.79000$	$(2^{2.3}, 2^{-0.5}) 0.79000$

Taking Haberman and Glass data sets as examples, the experiment process is as follows: First a relatively big range was set, and the search step was set to 1. The obtained contour maps of classification accuracy for Haberman and Glass data sets using SVM are shown in Figure 3.

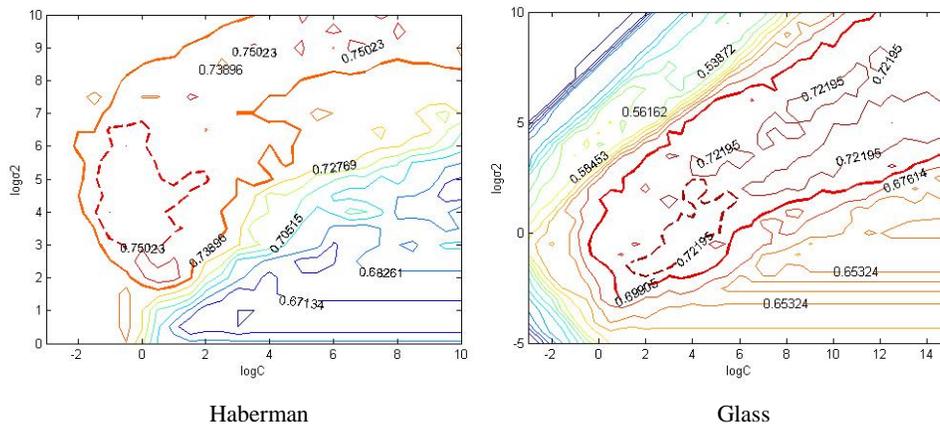


Figure 3. Contour Maps of Classification Accuracy for Haberman and Glass Data Sets

The "reasonable regions" are shown by thick dash lines. Then, within the "reasonable regions", the range of (C, σ^2) was refined to $\log_2 C \in [-2 \dots 3]$, $\log_2 \sigma^2 \in [2 \dots 7]$ for Haberman data set, and $\log_2 C \in [1 \dots 6]$, $\log_2 \sigma^2 \in [-3 \dots 3]$ for Glass data set, respectively. Then the bilinear grid searching method was used.

It can be seen in Table 1 that the results of the proposed method are comparable to that of the bilinear grid searching method. However, the parameters of the latter are too big, which increases the risk of underfitting and overfitting. Furthermore, because the reasonable regions are identified in advance, it speeds up the search procedure; the obtained parameters of the proposed method are relatively small, which results in stable performances.

4. The Relation between the Distribution Characters of (C, σ^2) and Sample Dimensionality

In order to quickly find a reasonable (C, σ^2) , it is necessary to find out the relation between the sector region and the sample dimensionality. Again, taking the 2nd and 3rd classes of Iris data set as an example, we increased the sample dimensionality step by

step, and compared the relation between the sample dimensionality and the optimal range of (C, σ^2) . The newly added dimensions to the samples are random values, which can be seen as noise. Their effect on the training results of SVMs is very little, and they have no major impact on the prediction accuracy of original samples. The original samples were 4-dimensional. In the experiments, the dimensionality was increased by a step of 2, with 6 total steps, which makes the dimensionalities of each experiment to be 4, 6, 8, 10, 12 and 14, respectively. The optimal values of (C, σ^2) after dimensionality increase are shown in Figure 4: 1-6.

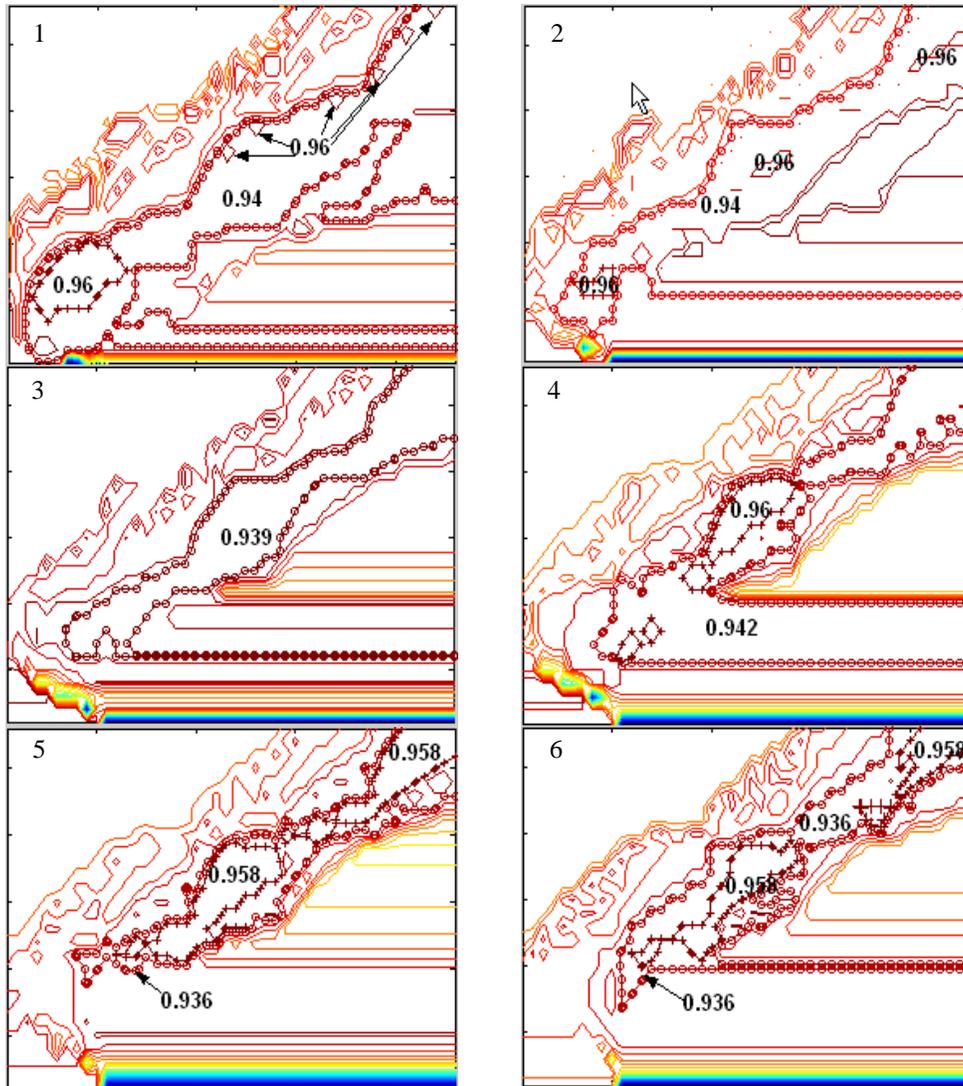


Figure 4. The Relation between the Selected (C, σ^2) and the Dimensions of the Samples

In Figure 4, for comparability, we try to use contours with the same prediction accuracy to mark the sector region. Here the boundaries of sector regions are labeled by contours with prediction accuracy of 0.94. These contours are marked by dots and

circles. Note that in the experiments, the contours might not exactly take the value of 0.94, so the boundaries of the sector regions in the figure may have some subtle differences.

Ignoring the differences of prediction accuracies of the sector boundaries, it can be seen that with the increase of sample dimensionality, the sector boundaries continuously move towards the direction in which both C and σ^2 increase. The shapes are getting more and more irregular. The reasonable regions change according to good regions, but they are generally near the sector center. Meantime, in good regions, the effect of σ^2 on the generalization ability is bigger than that of C , which is demonstrated by the fact that the prediction accuracy contours closed to C -coordinate are mainly smooth and straight, even completely horizontal.

To better classify samples, it is necessary for users to select certain attributes of the samples. However, the increase of sample dimensionality will lead to the increase of the VC dimensionality of the learning machine, which reduces the generalization ability of the SVM. When the sample dimensionality increases, $\|x-y\|^2$ will also increase, while $k(x,y)$ decreases. That is to say, the increase of dimensionality causes the sample complexity to increase and the similarity between samples to decrease, and brings uncertainties to the classification. In order to improve the performance of an SVM, it is necessary to increase σ^2 when sample dimensionality increases. Only in this way can the similarity range be expanded and the effect brought by the increase of $\|x-y\|^2$ be compensated. Thereby the prediction ability of the SVM can be preserved. With the increasing of σ^2 , it is also necessary to increase C to guarantee the performance of the SVM. The result of such adjustments is the movement of the sector region towards top-right, as shown in Figure 4., which indicates that when sample dimensionality is big, C and σ^2 should also be big.

5. Conclusion

1. We analyzed the optimal range of (C, σ^2) in details based on the viewpoint of similarity. We pointed out reasonable parameter selection regions and proved the existence of such regions by experiments, which serves as a clear direction of parameter selection for SVMs with Gaussian kernels.

2. By identifying the reasonable regions in advance, we improved the bilinear grid searching method, which not only achieved the comparable accuracy with that of bilinear grid searching, but also avoided the situations in which (C, σ^2) get trapped into "local optima".

3. The identification of reasonable regions will be a base for the parameter selection methods of all kinds of evolutionary algorithms, which not only avoided getting trapped into "local optima", but also saved searching time, as well as improved searching efficiency.

Acknowledgements

This work was supported by a grant from Natural Scientific Fund of China 41271445 and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

- [1] V. Vapnik, "An Overview of Statistical Learning Theory", IEEE Trans Neural Networks, vol. 10, no. 5, (1999), pp. 988-999.
- [2] I. S. S. Keerthi, "Efficient tuning of SVM hyper parameters using radius/margin bound and iterative algorithms", IEEE Trans on Neural Networks, vol. 13, no. 5, (2002), pp. 1225-1229.
- [3] O. Chappelle, V. Vapnik and O. Bousquet, "Choosing multiple parameters for support vector machines", Machine Learning, vol. 46, no. 1, (2002), pp. 131-160.
- [4] N. E. Ayat, M. Cheriet and C. Y. Suen, "Automatic model selection for the optimization of SVM kernels", Pattern Recognition, vol. 38, no. 10, (2005), pp. 1733-1745.
- [5] T. Glasmachers and C. Igel, "Gradient-based adaptation of general Gaussian kernels", Neural computation, vol. 17, no. 10, (2005), pp. 2099-2105.
- [6] C. Gold and P. Sollich, "Model selection for support vector classification", Neurocomputing, vol. 55, no. 1-2, (2003), pp. 221-249.
- [7] L.-H. Chen and H.-D. Hsiao, "Feature selection to diagnose a business crisis by using a real GA-based support vector machine: An empirical study", Expert Systems with Applications, vol. 35, no. 3, (2008), pp. 1145-1155.
- [8] P. I. Pingfeng and H. Weichiang, "Support vector machines with simulated annealing algorithms in electricity load forecasting", Energy Conversion and Management, vol. 46, no. 17, (2005), pp. 2669-2688.
- [9] X. C. Guo, J. H. Yang, C. G. Wu, C. Y. Wang and Y. C. Liang, "A novel LS-SVMs hyper-parameter selection based on particle swarm optimization", Neurocomputing, vol. 71, no. 16-18, (2008), pp. 3211-3215.
- [10] S. S. Keerthi and C. J. Lin, "Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel", Neural Computation, vol. 15, no. 7, (2003), pp. 1667-1689.
- [11] Z. Xiaoyun and L. Yuncai, "Performance Analysis of Support Vector Machines with Gauss Kernel", Computer Engineering, (in Chinese), vol. 29, no. 8, (2003), pp. 22-25.
- [12] W. Peng and Z. Xiaoyan, "Model Selection of SVM with RBF Kernel and its Application", Computer engineering & application, (in Chinese), vol. 39, no. 24, (2003), pp. 72-73.
- [13] <http://www.csie.ntu.edu.tw/~cjlin/>.

Authors

Zhigang Yan received B.Sc. degree from China University of Mining and Technology in 1997, and Ph.D. degree from China University of Mining and Technology in 2007. He is currently a associate professor at faculty of China University of Mining and Technology, China. His field of interest is spatio-temporal data mining and knowledge discovering.

Yuanxuan Yang, Master student, received B.Sc. degree in Geographic Information System in 2013 from China University of Mining and Technology. Now he study in China University of Mining and Technology, supervised by Zhigang Yan.

Yunjing Ding received M.Sc. degree in Cartography and Geographic Information Engineering from China University of Mining and Technology in 2011.