Parameter Selection of Gaussian Kernel SVM based on Local Density of Training Set

Jiawei Yang[1], Zeping Wu[1*], Ke Peng[1], Patrick N. Okolo[2,3], Weihua Zhang[1], Hailong Zhao[1], Jingbo Sun[1]

1. College of Aerospace Science and Engineering, National University of Defense Technology, Changsha, Hunan, P. R. China, 410073

2. Dept. of Mechanical and Manufacturing Engineering, Trinity College Dublin, University of Dublin, Dublin 2, Ireland

3. Buildings Fluid Dynamics Ltd, Dublin 2, Ireland.

Abstract: Support vector machine (SVM) is regarded as one of the most effective techniques for supervised learning, while the Gaussian kernel SVM is widely utilized due to its excellent performance capabilities. To ensure high performance of models, hyperparameters, i.e. kernel width and penalty factor must be determined appropriately. This paper studies the influence of hyperparameters on the Gaussian kernel SVM when such hyperparameters attain an extreme value (0 or ∞). In order to improve computing efficiency, a parameter optimization method based on the local density and accuracy of Leave-One-Out method are proposed. Kernel width of each sample is determined based on the local density needed to ensure a higher separability in feature space while the penalty parameter is determined by an improved grid search using the Leave-One-Out method. A comparison with grid method is conducted to verify validity of the proposed method. The classification accuracy of five real-life datasets from UCI database are 0.9733, 0.9933, 0.7270, 0.6101 and 0.8867, which is slightly superior to the grid method. The results also demonstrate that this proposed method is computationally cheaper by 1 order of magnitude when compared to the grid method.

Key words: Support vector machine； Gaussian kernel； Kernel width； Parameter selection； Penalty factor

1. Introduction

Support Vector Machine is a machine learning method proposed by Vapnik in the mid-1990s based on the statistical learning theory and the principle of structural risk minimization[35]. This machine learning method

* Corresponding author: zeping90315@126.com; zeping123@nudt.edu.cn

seeks the best compromise between empirical risk and confidence range through limited number of samples and ensures good generalization of the model. Due to its excellent performance, SVM is widely used in face recognition, text classification and other fields.

SVM is a binary classification model, which maps training data into a feature space by kernel function and finds the optimal classification hyperplane by solving the classifier with the largest interval within the feature space. Theoretically, as long as a function satisfies Mercer condition, it can be selected as a kernel function[27], however, different kernel functions will lead to completely different properties. The commonly utilized kernel functions are the linear kernel functions, polynomial kernel functions, Gaussian kernel functions and the Sigmoid kernel functions which are appropriate for different applications respectively[1, 3, 6, 7, 18, 34]. Among these functions, the Gaussian kernel is the frequently used function, where a penalty parameter C and kernel width σ are optimized[8].

Various researchers have presented different ideas for parameter optimization of SVM. Initially, the grid search was widely used to simultaneously optimize parameters σ and C. Grid search method firstly determines the optimal range of σ and C, followed by separating the range within M and N points respectively, thereby achieving an $M \times N$ combination of (C, σ). For each combination, SVM training model is utilized to evaluate the learning rate and a selection of the optimal parameter is achieved by picking the one possessing the highest accuracy. However，this method is computationally demanding, as it makes appropriate choice of the optimal range and separates the grid in an extremely small approach to get the optimal parameters, which makes it extremely time-consuming[12, 20].

More recently, large progress has been achieved in intelligent algorithms, and new methods are discovered to optimize the hyperparameters of a Gaussian kernel SVM. Research has been conducted to combine

intelligence algorithm with SVM. Genetic Algorithm (GA) is relatively complicated but can locate global optimum while Particle Swarm Optimization (PSO) is not complex and possesses high convergence rate. Both of these methods are continually applied to parameters optimization of Gaussian kernel SVM[4, 9, 25, 30, 33, 37]. Huang C L utilized Genetic Algorithm (GA) in both parameter selection and feature selection, and the optimization model showed superior performance than the grid search model[13] . Liu Y proposed an improved PSO-SVM (IPSO-SVM) model which obtained a higher precision than the PSO-SVM model[24]. Algorithms such as the Cuckoo Algorithm[29], Slap Swarm Algorithm (SSA)[28], Water Wave Optimization[17] have shown to be more effective when compared to models like the normal SVM or BP neural network. Compared to GA-SVM and PSO-SVM, Yang Dalian showed that the artificial bee colony algorithm could obtain a higher performance both in time consumption and recognition rates of gearbox faults[40]. Recently, Bayesian Optimization (BO) became a popular choice amongst researchers because it calculates a prior belief about behavior of the hyperparameters and then searches the parameter space by accomplishing and updating it based on its current measurement[2, 15]. In general, intelligence algorithm optimization can attain relatively good results, but cannot avoid multiple iterations in determination of parameters. When the iteration interval is too small, the model is computationally expensive. Therefore, studies on reducing iterations during optimization has become a hotspot of research interest amongst researchers.

To reduce iterations and improve optimization efficiency, researchers equally pay attention to properties of the sample data and SVM itself. Keerthi S S studied the asymptotic behaviors of Gaussian kernel SVM, and proposed a heuristic optimization method called Bilinear method[16]. This method firstly solves the best C to linear SVM as C， and then searches for the best combination of (C, σ) which satisfies $log\sigma=logC-logC$ in the Gaussian kernel SVM. Based on Bilinear method, Haochen Shi combined Segmented Dichotomy (SD) with

Gird Searching (GS) method, and proposed a model with higher generalization ability on the basis of a composite parameter selection method named the SD-GS algorithm[32]. Han Y studied the property of sample distribution and proposed a method to judge if the sample data is in agreement with the Gauss distribution or not. If in agreement, the kernel width σ was set equal to shape parameter a and then the best σ was searched through simple iteration[11]. To improve the efficiency of grid search, Wang D put forward two heuristic search methods named as the two-point central vertical method and the multi-point barycenter method, which could also be used in GA, PSO, Ant Colony Algorithm in order to accelerate convergence rate[36]. Distributed Learning and Searching (DL&S), which could be combined with other intelligence algorithm like Bees Algorithm (BA), was also found to be an effective method to reduce computing time[38, 39]. Some researchers also aimed to construct evaluation function so as to ensure the largest nonlinear between-class separability and the smallest nonlinear within-class separability, which could also be used in feature selection[21]. Generally, researchers constructed evaluation function by combining two criterions $w(\sigma)$ and $b(\sigma)$ [19, 22]. A better σ should be determined such that $w(\sigma)$ and $b(\sigma)$ were close to 1 and 0 respectively. Hence, the optimal σ could be obtained by solving the optimization problem showing in formula (1.2). The advantage of this method is that, calling the SVM model during the optimization process is avoided and therefore computing resources are minimized.

$$\min_{\sigma>0} \ J(\sigma)=1\text{-}w(\sigma)+b(\sigma) \tag{1.2}$$

However, all methods mentioned so far have concentrated on optimizing a fixed σ, which might not be useful in certain applications, such as the transformer fault diagnosis, because the fixed parameter σ would reduce the function of useful feature in the transformer fault diagnosis[31]. Thus, research of variable parameter is necessary. In Section 2, this paper introduces the principle of Gaussian SVM and properties of the kernel parameters σ and C. Section 3 proposes a new optimization method based on local density of samples in order

to avoid iterations in determination of kernel parameters. Section 4 gives the experimental results of the proposed method on 5 real-life datasets and makes a comparison with grid method. In Section 5, the paper is summarized, and general conclusions are given.

2. Gaussian kernel SVM

2.1 Principle of SVM

The principles of SVM are clearly doc

umented [10], with SVM theory used to solve typical classification problems with two-class data. The purpose of this theory is to locate an optimal classification hyperplane, which not only separates the two types of data within the training set, but also maximizes the margin area on both sides of the hyperplane. When the training sample is a two-dimensional linear separable data, Figure 2.1 shows that SVM can theoretically make the optimal classification.

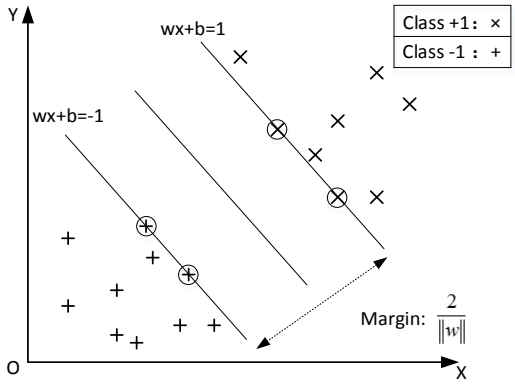

Figure 2.1 Optimal hyperplane of linear separable case

For a linear separable case，a set of training samples (x_i, y_i)，i=1，…，n，$y_i \in \{-1,+1\}$is given，among which x_i is the input sample data，y_i corresponds to the class label of sample i. A sample which can be classified correctly should satisfy formula (2.4)

$$\begin{aligned} w\cdot x_i + b \geq +1, y_i = +1 \\ w\cdot x_i + b \leq -1, y_i = -1 \end{aligned} \tag{2.4}$$

These two formulas could be combined into one equation as;

$$y_i(w\cdot x_i + b) - 1 \geq 0 \tag{2.5}$$

According to Figure 2.1, the margin between the two classes is given by $\frac{2}{\|w\|}$. To ensure the generalization ability of SVM, the optimal hyperplane should make the margin as large as possible. Accordingly, the origin problem turns out to be a convex quadratic programming problem, as presented in (2.6).

$$\begin{aligned} &\min \frac{1}{2}\|w\|^2 \\ &s.t.\ \ y_i(w\cdot x_i + b) - 1 \geq 0 \end{aligned} \tag{2.6}$$

To solve this problem, the Lagrange formulation is used to translate the formula as follow:

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}(\alpha_i y_i(w\cdot x_i + b) - 1) \tag{2.7}$$

in which $\alpha_i > 0$, are the Lagrange multipliers. The partial derivatives on ω and b are respectively taken to obtain the minimum point

$$\begin{aligned} &\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0, \ \ \mathrm{w}=\sum_{i=1}^{n}\alpha_i y_i x_i \\ &\frac{\partial L}{\partial b} = \sum_{i=1}^{n}\alpha_i y_i = 0, \ \ \sum_{i=1}^{n}\alpha_i y_i = 0 \end{aligned} \tag{2.8}$$

The Karush Kuhn-Tucker (KKT) conditions are necessary and sufficient to solve the maximum of equation (2.7). The complete corresponding KKT conditions are

$$\begin{aligned} &\alpha_i \geq 0 \\ &y_i(w\cdot x_i + b) - 1 \geq 0 \\ &\alpha_i[y_i(w\cdot x_i + b) - 1] = 0 \end{aligned} \tag{2.9}$$

Substituting equation (2.8) into equation (2.7)，the convex quadratic programming problem becomes an extremum problem with regards to the Lagrange multipliers α_i which is represented by equation (2.10), and the final optimal hyperplane is determined by a decision function as shown in equation (2.11).

$$\max_{\alpha} L(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\left\langle x_i,x_j\right\rangle$$
$$s.t.\ \ \alpha_i\geq 0\ i=1,...,n\ and\ \sum_{i=1}^{n}\alpha_i y_i=0 \tag{2.10}$$

$$f(x,\alpha^*,b^*)=\sum_{i=1}^{n}y_i\alpha_i^*\left\langle x_i,x\right\rangle+b^* \tag{2.11}$$

From the KKT conditions, there exists two situations for each sample：(1) when α_i=0， sample *i* will not appear in equation (2.11), this indicates that the sample has no effect on the hyperplane selection; (2) when $\alpha_i > 0$, $y_i(w\bullet x_i+b)=1$, this means that the sample is on the boundary, which is the so-called support vector. Obviously, the hyperplane only relates to support vectors.

The SVM mentioned so far within this section holds only if the data are linear separable. For a linear inseparable case, the optimal hyperplane should reach a tradeoff between the maximization of margin and the minimization of allowable error. Therefore， a slack variable ξ_i and penalty parameter *C* are introduced into the problem. The resulting new equation is;

$$\min\frac{1}{2}\|w\|^2+\mathrm{C}\sum_{i=1}^{n}\xi_i$$
$$s.t.\ \ y_i(w\cdot x_i+b)\geq 1-\xi_i \tag{2.12}$$
$$\xi_i\geq 0$$

The penalty parameter *C* determines the penalty of misclassified samples. The larger the parameter *C*， the higher the penalty to errors of training set. Lagrange formulation are used to translate the equation (2.12) to equation (2.13). The penalty parameter *C* becomes the upper boundary of Lagrange multipliers α_i. Users can control the misclassified rate of training set through changing the *C* value.

$$\max_{\alpha} L(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\left\langle x_i,x_j\right\rangle$$
$$s.t.\ \ 0\leq\alpha_i\leq \mathrm{C},\ i=1,...,n\ and\ \sum_{i=1}^{n}\alpha_i y_i=0 \tag{2.13}$$

When training data are at a nonlinear situation， kernel methods are effective approaches to separate the

different training classes. SVM maps training data into a higher dimensional feature space through kernel functions and built optimal separation hyperplane in feature space. In this case, the inner product operation in equation (2.13) is replaced by kernel functions and the new equation becomes;

$$\begin{aligned} &\max_{\alpha} \ L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \\ &s.t. \ \ 0 \le \alpha_i \le \mathrm{C}, \ i = 1, ..., n \ \ and \ \ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \tag{2.14}$$

The decision function using kernel functions are

$$f(x, \alpha^*, b^*) = \sum_{i=1}^{n} y_i \alpha_i^* k(x_i, x) + b^* \tag{2.15}$$

Different kernel functions have completely different influence on SVM properties. In this paper, the Gaussian kernel is chosen due to its better performance, as represented by equation (2.16). The kernel width σ decides the feature space to which training data will be mapped. Thus, a proper σ greatly affects the SVM learning performance.

$$k(x_i, x_j) = \exp\left(-\frac{\left\| x_i - x_j \right\|^2}{2\sigma^2}\right) \tag{2.16}$$

2.2 Algorithm of multiclass SVM

Multiclass SVM are aimed at solving multi-class classification problems in which the dataset are over two dimensions. There are generally two approaches to this. One approach is the direct method, performed through improving the objective function in order to create a multi-class model, but the objective function becomes complicated with low accuracy. The second approach is the combination method， performed through combining several binary models to construct a multi-class learning machine. Usually, there are One Against One (OAO) and One Against All (OAA) methods[14]. Melgani F evaluated the validity of four multi-class classification methods including OAO and OAA in classifying hyperspectral dataset. The results illustrated that

OAO and OAA algorithms showed better performance while the computational time was longer[26]. Practically, OAA algorithm is the earliest and most widely used algorithm[5]. If the data has k classes, then OAA constructs k two-class SVM model. Taking the i^{st} two-class SVM as an example, the i^{st}-class data is made as one class and the rest as another class so as to separate the i^{st}-class data. The algorithm is as shown in Figure 2.2.

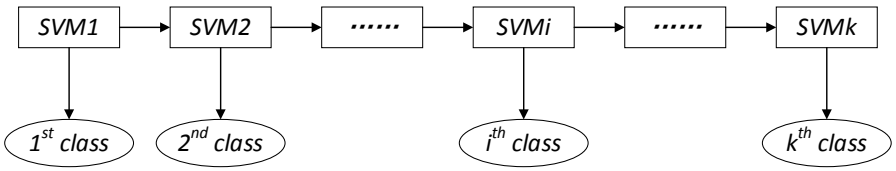

Figure 2.2 OAA algorithm

2.3 Kernel width σ and penalty parameter C

Gaussian kernel SVM is widely applied due to its excellent learning performance. The hyperparameter C and σ greatly influences the learning machine. The penalty parameter controls the compromise of model complexity and allowable error. The higher the C value，the higher the requirement for classification accuracy of training data，and the lower the generalization capability of the machine. When $C\rightarrow 0$，the model takes little punishment towards error samples，the machine complexity is low and classification accuracy is not satisfactory；when $C\rightarrow \infty$，all the training samples will be classified correctly，and the training model is extremely complex. Lin et al.'s research also shows that when the C value exceeds a certain value C^*, the machine is over-fitting, which is equal to a hard-margin SVM[23]. An easy method to attain the value is to solve equation (2.14) with $C=\infty$，then set $C^*=\max\alpha_i$[16]. This can also be explained by studying equation (2.13). For the same set of training data, if $C > \max\alpha_i$，this means the choice of C value will not affect the selection of α_i which determines the selection of the support vectors. Hence a change of C value has nothing to do with the optimization of a

hyperplane.

Kernel width σ determines the feature space which the samples will be mapped unto, and largely influences the classification accuracy. As $\sigma \to 0$， all training samples can be classified correctly， but generalization performance of the learning machine is poor and SVM can not classify new samples; as $\sigma \to \infty$， the whole training sample set is classified as one class.

This property can be explained through a mathematical approach. Function *φ(x)* is used to map the samples unto feature space. The Hilbert space distance square in high dimensional feature space of any two samples are represented as

$$\left\|\phi(x_i)-\phi(x_j)\right\|^2 = k(x_i,x_i)-2k(x_i,x_j)+k(x_j,x_j) \tag{2.17}$$

When $\sigma \to 0$，it is easy to find from equation (2.16) that

$$\begin{aligned} &k(x_i,x_i)=k(x_j,x_j)=1 \\ &k(x_i,x_j)=0 \end{aligned} \tag{2.18}$$

And equation (2.17) becomes

$$\left\|\phi(x_i)-\phi(x_j)\right\|^2 = 2 \tag{2.19}$$

Equation (2.19) indicates that，when $\sigma \to 0$，the distance between any two samples in feature space is $\sqrt{2}$. Samples from the same class won't gather，and each sample will be classified as one single class，so all the training data can be correctly classified. However, the machine is over-fitting, and can't classify new samples.

When $\sigma \to \infty$, equation (2.16) becomes

$$\begin{aligned} &k(x_i,x_i)=k(x_j,x_j)=1 \\ &k(x_i,x_j)=1 \end{aligned} \tag{2.20}$$

And equation (2.17) can be simplified as

$$\left\|\phi(x_i)-\phi(x_j)\right\|^2 = 0 \tag{2.21}$$

Equation (2.21) indicates that，when $\sigma\to\infty$，samples which have been mapped to feature space become the same point and the distance of any two samples is zero. Thus, all samples will be classified as one class and the machine will not be able to distinguish amongst the training data.

3. Local density-based parameter selection of Gaussian kernel SVM

3.1 Determination of kernel width σ

This paper presents a novel method which is different from previous documented research based on sample distribution. Nonlinear SVM maps data which cannot be distinguished in origin space into feature space. If a region of dense distribution exists, then after mapping data in this region with the same kernel function, a dense region will appear in feature space which makes it difficult to classify these sample data. From previous section of this paper it is clearly shown that when $\sigma\to 0$，all samples would be classified correctly, yet the learning machine is over-fitting. This also inspires the notion that different kernel parameter σ for each sample could be selected based on the sample distribution in origin space. The principle of selection is; σ value of samples in dense region takes relatively small values and σ value of samples in sparse region takes relatively large values. In this manner, the method can theoretically ensure better separability after the samples are mapped to feature space. The specific steps of parameter selection are summarized as:

STEP 1：Normalize the samples to n-dimensional unit cube by following the equation;

$$x_k = \frac{X_k - X_k^{\,L}}{X_k^{\,U} - X_k^{\,L}} \quad k = 1,2,...,n \tag{3.23}$$

Where，X_k^U and X_k^L are upper and lower bounds on the k^{th} dimension design variables，X_k is the value of the k^{th} design variable before mapping while x_k is the corresponding value after mapping. n represents the sample dimensions.

STEP 2：Calculate the local density of each sample according to sample distribution using the following

equation;

$$\rho(x_i)=\sum_{j=1}^{N} e^{-\frac{\|x_i-x_j\|^2}{c^2}}=\sum_{j=1}^{N} e^{-\frac{(x_i-x_j)^T(x_i-x_j)}{c^2}} \tag{3.24}$$

N is the total number of samples. Setting $c=1/\sqrt[n]{N}$, then equation (3.24) becomes;

$$\rho(x_i)=\sum_{j=1}^{N} e^{-N^{2/n}(x_i-x_j)^T(x_i-x_j)} \tag{3.25}$$

STEP 3：Calculate minimum distance $d_{z,min}$ of the sample with minimum local density to other samples

$$d_{z,\min}=\min_{j=1}^{N}[(x_s-x_j)^T(x_s-x_j)] \tag{3.26}$$

STEP4：Calculate the kernel width σ

$$\sigma_i=\sqrt[n]{\rho(x_s)/\rho(x_i)}d_{z,\min} \tag{3.27}$$

Following the method steps as outlined, the optimal model will not only satisfy the selection principle, but will also avoid negative effects to the SVM performance due to oversized deviation of σ value of each sample. Consequently, there is no need to iterate σ or call SVM training model. The procedure is also shown in Figure 3.1.

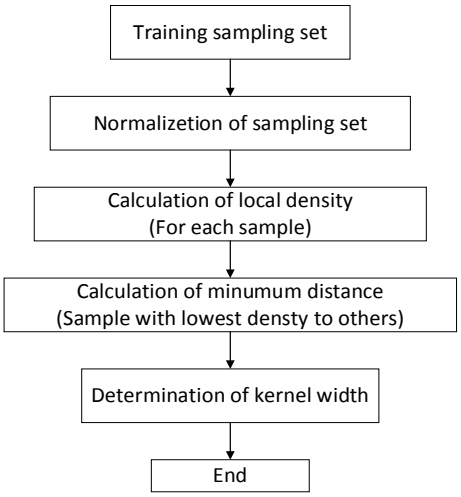

Figure 3.1 Determination of kernel width

3.2 Optimization of penalty parameter *C*

Grid search or intelligence algorithms are frequently applied towards optimizing σ and *C*. In this paper, parameter *C* is determined by grid search using Leave-One-Out (LOO) method. The theory of LOO method is as follows; select one sample as the test set each time and the rest becomes the training set. After all samples are tested, take an average of all results to evaluate the precision. LOO method can make best use of samples. Normal grid search method demands searching the design space as careful as possible in order to get a satisfactory result, which is apparently expensive in computational effort. Thus, a simple improvement is applied here. The design space is firstly searched with a coarser grid. When the better region on the grid is identified utilizing the LOO method, a finer grid search can then be executed. The *C* value with highest accuracy of LOO method is picked as the optimal *C* value. The procedural steps are:

STEP 1: Set $C=2^{\theta}$，the range of θ is between -4 and 10，a coarser grid is applied with $C = 2^{-4}, 2^{-2}, 2^{0}, 2^{2}, \dots, 2^{8}, 2^{10}$.;

STEP 2: Utilize LOO method to evaluate the performance of SVM with different *C* value. The best *C* value 2^{p} is picked and therefore a better region is identified as $(2^{p-2}, 2^{p+2})$;

STEP 3: A finer grid search with LOO method is executed on the region obtained in STEP 2 with $C = 2^{p-2}, 2^{p-1.5}, 2^{p-1}, 2^{p-0.5}, \dots, 2^{p+1.5}, 2^{p+2}$.

STEP 4: Pick the C value with the highest accuracy of LOO method as the optimal C.

4. Numerical analysis and discussion

4.1 Setting of the parameters

To evaluate the validity of the optimal method proposed within this paper, five datasets are chosen from

UCI database to test：Iris，Wine，Glass Identification，Heart Disease and Ionosphere. They are all collected from real life cases and their specific properties are as shown in Table 4.1.

Table 4.1 Experimental data

Number	Dataset	Numbers	Attributes	Classes
①	Iris	150	4	3
②	Wine	178	13	3
③	Glass	214	10	6
④	Heart Disease	303	13	5
⑤	Ionosphere	351	34	2

Table 4.1 presents the number of samples, properties and classes of each dataset which are repeatedly used in literatures. Taking Iris dataset as an example, there are a total of three classes: Iris Setosa、Iris Versicolour and Iris Virginica, 150 samples, and each sample contains four properties: sepal length, sepal width, petal length, petal width. Iris is perhaps the best-known database to be found in the pattern recognition literature. The Wine dataset is usually used to test a new model; however, it is not very challenging. Glass Identification, Heart Disease, Ionosphere dataset have noise and therefore need an improved SVM model.

To compare the performance of different combinations of (C, σ), a SVM model evaluation method is needed. The most famous method is the *k-fold* cross validation method. This method devises the sample to k subsets that do not intersect each other, $S_1 \ldots S_k$. When it comes to the i^{th} validation, take S_i as test set， the rest are the training set, and solve the classification accuracy e_i, after $e_1,\ldots,e_k$ have been solved，take $\sum_{i=1}^{k} e_i/k$ as the evaluation standard of the SVM performance. When the k value is equal to the number of samples, the evaluation method is called LOO method. LOO is the most accurate method to evaluate the SVM performance. However, it takes much time to accomplish when the sample dataset is large. LOO is used to evaluate the performance of this proposed method because the dataset utilized in this paper are relatively small. Results are compared with the grid search method in order to verify and validate the proposed method within this paper. In

order to eliminate the influence of dimension between sample features, all samples utilized are normalized.

4.2 Results and discussion

Using OAA algorithm to handle multi-class problem, the results are as shown in table 4.2.

Table 4.2 Experiment results

Data	Local density(LD)		Grid search(GS)		Range	SVM calling (LD/GS)
	Local density	C	Grid	(C, σ)		
Iris	0.9733	0.5	0.9733	(1,1.4142)	$2^{(-4,10)}$	17/841
Wine	0.9933	0.707107	0.9867	(1,1)	$2^{(-4,10)}$	17/841
Glass	0.7270	4	0.7383	(8,22.6274)	$2^{(-4,10)}$	17/841
Heart Disease	0.6101	2.828427	0.6007	(1.4142,2.8284)	$2^{(-4,10)}$	17/841
Ionosphere	0.8867	22.62742	0.8800	(2,1)	$2^{(-4,10)}$	17/841

When proceeding with the experiment, take 0.5 as the iteration interval. The results illustrate that the proposed method in this paper achieve approximately the same accuracy with grid search method for the dataset with good separability such as Iris and Wine. For datasets of the Glass Identification, Heart Disease and Ionosphere, the proposed method shows a much better performance. Furthermore, compared to the grid search method, parameter selection based on local density of training set greatly reduces the model calling times and improves the computational efficiency.

Huang C L utilized the IRIS dataset to evaluate the performance of SVM optimized by GA and the classification accuracy was 0.9756, which is almost the same as this proposed method. The accuracy to Ionosphere dataset of GA method is 0.9661. Considering the GA method was also constant iterations of two hyperparameters, and the authors conducted feature selection method in order to improve the machine performance, a better result compared to our proposed method was reasonable and acceptable. It is necessary to state that for the Heart Disease dataset, the optimal method and grid method within this paper did not achieve satisfying result, while accuracy of the GA-based method was up to over 0.80. This can be explained due to this paper using a different dataset. The Heart Disease dataset within this paper is from the UCI Cleveland database

which consists of five classes, but the dataset in Wang D's research is taken from Statlog Project which contains only two classes. As a result, the optimal method proposed in this paper has a better performance than the grid search in terms of accuracy and computational efficiency. When compared to intelligence method, our proposed method only needs to iterate parameter C, therefore this improves its computing efficiency.

5. Conclusion

SVM is widely applied in pattern recognition and fault diagnosis recently. Hyperparameters σ and C are of great importance to SVM performance. This paper proposed a parameter selection method in order to enhance the performance of SVM classifier and reduce iterations during hyperparameters optimization. The kernel width σ of each sample is determined respectively based on local density to ensure the separability of samples in feature space and avoid complicated iterations. Penalty parameter C is firstly searched with a coarser grid based on LOO method, then a finer grid search is conducted on the identified region with better classification accuracy to locate the optimal parameter C. To evaluate the efficiency of proposed method, 5 real-life datasets for classification from UCI database are tested and compared to the Grid search method. Results indicate that times of SVM calling by proposed method is only 17, which is much cheaper compared to 841 times of SVM calling by Grid search method. Besides, the proposed method obtains relative better accuracy than grid method. Consequently, the proposed parameter selection method based on local density outperforms Grid search in almost all test datasets and is computational superior to Grid search and other methods based on iterations.

Acknowledgement

This research was sponsored by the Research Project of National University of Defense Technology (Project No.:ZK19-11).

References

[1] I.S. Al-Mejibli, D.H. Abd, J.K. Alwan, A.J. Rabash, Performance Evaluation of Kernels in Support Vector Machine, in: 2018 1st Annual International Conference on Information and Sciences (AiCIS), 2018, pp. 96-101.

[2] S. Anis, M.H. El-Mahlawy, M.E.A. Gadallah, E.A. El-Samahy, Parametric Fault Detection of Analogue Circuits, International Journal of Computer Applications, 96 (9) (2014) 14-23.

[3] C. Antonio Alves Kaestner, Support Vector Machines and Kernel Functions for Text Processing, Revista de Informática Teórica e Aplicada, 20 (2013) 130.

[4] V. Azimirad, M. Hajibabzadeh, P. Shahabi, A new brain-robot interface system based on SVM-PSO classifier, (2017).

[5] L. Bottou, Comparison of classifier methods : a case study in handwritten digit recognition, Proc. 12th ICPR, 1994, (1994) 77-82.

[6] X. Dai, N. Wang, W. Wang, Application of machine learning in BGP anomaly detection, Journal of Physics: Conference Series, 1176 (2019) 032015.

[7] S. Fadel, S. Ghoniemy, M. Abdallah, H.A. Sorra, A. Ashour, A. Ansary, Investigating the Effect of Different Kernel Functions on the Performance of SVM for Recognizing Arabic Characters, International Journal of Advanced Computer Science & Applications, 7 (1) (2016).

[8] J. Fan, F. Jing, Z. Fang, M. Tan, Automatic recognition system of welding seam type based on SVM method, The International Journal of Advanced Manufacturing Technology, 92 (1) (2017) 989-999.

[9] Y. Gao, J. Leng, S. Li, Residual life prediction method for remanufacturing sucker rods based on magnetic memory testing and a support vector machine model, Insight, 61 (1) (2019) 44-50.

[10] P.V. Gehler, B. Schölkopf, An Introduction to Kernel Learning Algorithms, 2001.

[11] Y. Han, J. Li, J.Z. Li, H.W. Xing, A.M. Yang, Y.H. Pan, Demonstration of SVM Classification Based on Improved Gauss Kernel Function, (2016).

[12] C.W. Hsu, C.J. Lin, A Simple Decomposition Method for Support Vector Machines, Machine Learning, 46 (1-3) (2002) 291-314.

[13] C.L. Huang, C.J. Wang, A GA-based feature selection and parameters optimizationfor support vector machines, Expert Systems with Applications, 31 (2) (2006) 231-240.

[14] Y. Huang, C.Y. Zheng, Z.H. Song, Multi-class Support Vector Machines algorithm summarization, Computing Technology & Automation, (2005).

[15] M. Injadat, F. Salo, A.B. Nassif, A. Essex, A. Shami, Bayesian Optimization with Machine Learning Algorithms Towards Anomaly Detection, in: 2018 IEEE Global Communications Conference (GLOBECOM), 2018, pp. 1-6.

[16] S.S. Keerthi, C.-J. Lin, Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel, Neural Computation, 15 (7) (2003) 1667-1689.

[17] M. Kilany, E.H. Houssein, A.E. Hassanien, A. Badr, Hybrid water wave optimization and support vector machine to improve EMG signal classification for neurogenic disorders, in: 2017 12th International Conference on Computer Engineering and Systems (ICCES), 2017, pp. 686-691.

[18] D. Kumar, R. Kumar, Spam Filtering using SVM with different Kernel Functions, International Journal of Computer Applications, 136 (5) (2016) 16-23.

[19] B. Kuo, H. Ho, C. Li, C. Hung, J. Taur, A Kernel-Based Feature Selection Method for SVM With RBF Kernel for Hyperspectral Image Classification, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7 (1) (2014) 317-326.

[20] S.M. Lavalle, M.S. Branicky, On the Relationship between Classical Grid Search and Probabilistic Roadmaps, International Journal of Robotics Research, 23 (23) (2003) 673-692.

[21] C. Li, P. Hsieh, B. Kuo, Multiple SVMS based on random subspaces from kernel feature importance for hyperspectral image classification, in: 2017 IEEE International Geoscience and Remote Sensing Symposium

(IGARSS), 2017, pp. 574-577.

[22] C.H. Li, H.H. Ho, Y.L. Liu, C.T. Lin, B.C. Kuo, J.S. Taur, An Automatic Method for Selecting the Parameter of the Normalized Kernel Function to Support Vector Machines, in: International Conference on Technologies & Applications of Artificial Intelligence, 2010.

[23] C.J. Lin, Formulations of Support Vector Machines: A Note from an Optimization Point of View, Neural Computation, 13 (2) (2001) 307-317.

[24] Y. Liu, Y.K. Shi, M.W. Xu, L.L. Zhang, Y.L. Ding, A further improved support vector machine model along with particle swarm optimization for face orientations recognition based on eigeneyes by using hybrid kernel, in: IEEE International Conference on Industrial Engineering & Engineering Management, 2018.

[25] J. Manurung, H. Mawengkang, E. Zamzami, Optimizing Support Vector Machine Parameters with Genetic Algorithm for Credit Risk Assessment, in, 2017, pp. 012026.

[26] F. Melgani, L. Bruzzone, Classification of hyperspectral remote sensing images with support vector machines, IEEE Transactions on Geoscience and Remote Sensing, 42 (8) (2004) 1778-1790.

[27] J. Mercer, Functions of Positive and Negative Type, and Their Connection with the Theory of Integral Equations, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character (1905-1934), 83 (559) (1909).

[28] M.B. Nejad, M.E.S. Ahmadabadi, A Novel Image Categorization Strategy Based on Salp Swarm Algorithm to Enhance Efficiency of MRI Images, Cmes-Computer Modeling in Engineering & Sciences, 119 (1) (2019) 185-205.

[29] D. Niu, W. Zhao, L. Si, R. Chen, Cost Forecasting of Substation Projects Based on Cuckoo Search Algorithm and Support Vector Machines, Sustainability, 10 (1) (2018) 118.

[30] G.-c. Niu, Y. Wang, Z. Hu, Q. Zhao, D.-m. Hu, Application of AHP and EIE in Reliability Analysis of Complex Production Lines Systems, Mathematical Problems in Engineering, 2019 (2019) 10.

[31] L. Qu, H. Zhou, C. Liu, Z. Lu, Study on Multi-RBF-SVM for Transformer Fault Diagnosis, in: 2018 17th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), 2018, pp. 188-191.

[32] H. Shi, H. Xiao, J. Zhou, N. Li, H. Zhou, Radial Basis Function Kernel Parameter Optimization Algorithm in Support Vector Machine Based on Segmented Dichotomy, in: 2018 5th International Conference on Systems and Informatics (ICSAI), 2018, pp. 383-388.

[33] A. Subasi, Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders, Computers in Biology & Medicine, 43 (5) (2013) 576-586.

[34] S. Sutao, Z. Zhichao, L. Zhiying, Z. Jiacai, Y. Li, Comparative study of SVM methods combined with voxel selection for object category classification on fMRI data, Plos One, 6 (2) (2011) e17191.

[35] V.N. Vapnik, An overview of statistical learning theory, IEEE transactions on neural networks, 10 (5) (1999) 988-999.

[36] D. Wang, W.U. Xiang-Bin, D.M. Lin, TWO HEURISTIC STRATEGIES FOR SEARCHING OPTIMAL HYPER PARAMETERS OF C-SVM, in: International Conference on Machine Learning & Cybernetics, 2009.

[37] Z. Xiaomu, W. Tao, H. Jianjun, L. Chunfang, Z. Zhixian, Study on Tariff Risk Early Warning of Electric Power Users Based on PSO-SVM Algorithm, in: 2018 International Conference on Big Data and Artificial Intelligence (BDAI), 2018.

[38] Y. Xie, C. Bian, Y. Lu Murphey, D. Kochhar, An SVM parameter learning algorithm scalable on large data size for driver fatigue detection, 2017.

[39] Y.Q. Xie, Y.L. Murphey, D.S. Kochhar, SVM Parameter Optimization Using Swarm Intelligence for

Learning from Big Data, in: N.T. Nguyen, E. Pimenidis, Z. Khan, B. Trawinski (Eds.) Computational Collective Intelligence, Iccci 2018, Pt I, 2018, pp. 469-478.

[40] D. Yang, Y. Liu, S. Li, X. Li, L. Ma, Gear fault diagnosis based on support vector machine optimized by artificial bee colony algorithm, Mechanism & Machine Theory, 90 (2015) 219-229.