Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 5594899, 9 pages
https://doi.org/10.1155/2021/5594899

Research Article

Improvement of Support Vector Machine Algorithm in Big Data Background

Babacar Gaye, Dezheng Zhang, and Aziguli Wulamu

School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China

Correspondence should be addressed to Babacar Gaye; babacargaye92@gmail.com

Received 21 February 2021; Revised 30 April 2021; Accepted 3 June 2021; Published 17 June 2021

Academic Editor: Sang-Bing Tsai

With the rapid development of the Internet and the rapid development of big data analysis technology, data mining has played a positive role in promoting industry and academia. Classification is an important problem in data mining. This paper explores the background and theory of support vector machines (SVM) in data mining classification algorithms and analyzes and summarizes the research status of various improved methods of SVM. According to the scale and characteristics of the data, different solution spaces are selected, and the solution of the dual problem is transformed into the classification surface of the original space to improve the algorithm speed. *Research Process.* Incorporating fuzzy membership into multicore learning, it is found that the time complexity of the original problem is determined by the dimension, and the time complexity of the dual problem is determined by the quantity, and the dimension and quantity constitute the scale of the data, so it can be based on the scale of the data Features Choose different solution spaces. The algorithm speed can be improved by transforming the solution of the dual problem into the classification surface of the original space. *Conclusion.* By improving the calculation rate of traditional machine learning algorithms, it is concluded that the accuracy of the fitting prediction between the predicted data and the actual value is as high as 98%, which can make the traditional machine learning algorithm meet the requirements of the big data era. It can be widely used in the context of big data.

1. Introduction

The support vector machine (SVM) is a traditional machine learning method based on classification. It is derived from the idea of solving the dual form of large-dimensional problems, so that the classifier only relies on a small number of support vectors to achieve the principle of structural risk minimization. Statistical learning theory solved nonlinear and local minimum problems. The system call can use the short system frequency to convert the sequence into a call sequence with a certain length of vectors in a high-dimensional space. Therefore, anomaly detection can be performed based on support vector machines. In the context of the current big data era, it can implement multidomain applications in a big data environment. With the development of the current era, the scale of data is getting larger and larger, and the attributes of the data are also increasing. At the same time, the diversification of the values of data attributes makes it more difficult to classify such data. However, the massive amounts of data that appear now generally have such high dimensionality and diversity, which makes it difficult for some classification algorithms to use such data to build predictive models. These difficulties are manifested in the lack of scalability of the algorithm, the long time to build the model, or the problem of dimensional disasters. Support vector machines can be well applied to high-dimensional data, and there is no limit to the value of each attribute.

In the improvement of traditional algorithms under the big data platform, the most popular framework is the Spark Framework. The Spark Big Data Framework is an iterative distributed computing framework based on memory. According to the official description of the Apache Spark Open Source Organization, Spark performs iterative calculations based on disk and compares with other frameworks and finds that its speed is more than 10 times that of other frameworks; and iterative calculation based on memory is beyond a hundred times more. Therefore, it is very fast and efficient to build applications based on Spark's API. For example, classic traditional algorithms such as collaborative filtering algorithm and Bayesian recommendation algorithm have achieved large improvements and optimization based on this. Subsequently, finding a method to improve the support vector machine algorithm under the Spark big data framework is the problem to be solved [1, 2].

Guo proposed two multifault diagnosis methods based on the improved support vector machine (SVM), which are used for sensor fault detection and identification, respectively. First, use Online Sparse Least Squares Support Vector Machine (OSLSSVM) to detect and predict sensor failure. On this basis, Guo S. proposed a sensor fault feature extraction and online recognition method based on the combination of SVM and the Error Correction Output Code (ECOC). They used nonlinear transformation as input to the classifier to improve the separability of the initial features. Use ECOC-SVM to classify the fault status. After studying some typical faults, they found that ECOC-SVM has high recognition accuracy and can be realized in real time to meet the requirements of online fault recognition. This method can also be extended to solve other related problems [3]. Scholars such as Carrizosa found that in linear classifiers (such as support vector machines (SVM)), each feature has a score and assigns objects to classes based on a linear combination of score and feature value. Inspired by the Discrete Mental Scale (DILSVM), they proposed the Discrete Horizontal Support Vector Machine (DILSVM). The DILSVM classifier benefits from interpretability because it can be viewed as a set of Likert scales, each with one feature, where the level of consistency is scored with the positive class. In order to establish a DILSVM classifier, Carrizosa and other scholars proposed a mixed integer linear programming method and proposed a set of strategies to reduce the generation time. Our calculation experience shows that the 3-point and 5-point DILSVM classifiers have considerable accuracy and support vector machines and have a great improvement in interpretability and sparseness, thanks to the appropriate feature level selection [4].

This article introduces some basic concepts and principles of machine learning, provides necessary background support for the proposal of support vector machines, and then specifically introduces the basic ideas and specific theories and implementation algorithms of support vector machines. The solution space suitable for the scale of the data is selected, the solution of the dual space is converted into the classification surface of the original space, and it is divided into three groups for calculation. Research data shows that by improving the calculation speed of traditional machine learning algorithms, traditional machine learning algorithms can meet the requirements of the big data era. Research data shows that the improved algorithm greatly increases the time and space complexity of users. In addition, after three sets of experimental calculations, the prediction accuracy of the fitting between the predicted data and the actual value is as high as 98% (including two average tests, aptness and applicability).

2. Proposed Method

2.1. Support Vector Machine (SVM). SVM is a very classical two-classification model, and its working mechanism is to find a suitable hyperplane to segment the collected data samples. The principle of segmentation is to maximize the interval (including hard interval and soft interval), and finalize it into a special quadratic programming problem to solve. The main models are as follows: if the training sample is linearly time-sharing, use the linear separable support vector machine by maximizing the hard interval; if the training sample is approximately linearly time-sharing, use the linear support vector machine by maximizing the soft interval and selecting the appropriate kernel function; if the training sample is linearly non-time-sharing, make it possible to maximize the soft interval and select the appropriate kernel function, with a nonlinear support vector machine [5, 6]. The following is an overview of the main support vector machines.

2.1.1. Linear Separable Support Vector Machine. We first give a training sample set, the most basic idea of the so-called linear separable support vector machine is to find a suitable partition hyperplane in the sample space where the training sample set is M, separating the samples of different categories. If a linear function is able to separate samples, these data samples are called linearly separable. So specifically, what is a linear function? we generally think that a linear function is a straight line in a two-dimensional space, a plane in a three-dimensional space, and so on, if spatial dimensions are not considered; such a linear function is collectively called a hyperplane. In a two-dimensional space, for example, we look at a simple example of two-dimensional space. In the example, "O" represents positive classes and "X" refers to negative classes. Samples are linearly detachable; however, from a graphical point of view, it is clear that not only this straight line can separate samples, but also there are countless lines. The linear separable support vector machine corresponds to lines that can correctly divide the data and have the largest intervals.

Since the maximum interval is sought, it is imperative to calculate the interval in the sample space. In the sample space, we use the following linear equation to describe the division of the hyperplane:

$$W^T x + b = 0, \tag{1}$$

where W is a normal vector, which determines the direction of the hyperplane and b is a displacement, which determines the distance between the hyperplane and the origin. Assume that the hyperplane can correctly classify the training samples; that is, for the training samples, the following formula is satisfied:

$$w^t x_i + b \geq 1,\ y = 1, \tag{2}$$

$$w^t + b \leq -1,\ y = -1. \tag{3}$$

The above formula is called the maximum interval hypothesis. It indicates that the sample is a positive sample, expressed as a negative sample. In fact, the specified value of 1 or −1 here is only for the convenience of calculation, in principle can take any constant.

2.1.2. Nonlinear Support Vector Machines. For nonlinear support vector machine regression, the basic idea is to map the data to a high-dimensional feature space (Hilbert space) through a nonlinear mapping and perform linear regression in this space so that the nonlinear problem in the low-dimensional space corresponds to the high linear regression problem of dimensional feature control [7]. The specific algorithm is as follows:

$$f = \left[-\frac{1}{2}\sum_{c=1}^{i}\sum_{d=1}^{i}\left(a_c^* - a_c\right)\left(a_c - a_c^*\right)Q\left(x_c, x_d\right) - \vartheta\sum_{c=1}^{i}\left(a_c^* + a_c\right) + \sum_{c=1}^{i} b_c\left(a_c^* + a_d\right)\right]. \tag{4}$$

Among them,

$$\sum_{c=1}^{i} a_c^* = \sum_{c=1}^{i} ac \quad a_c^*,\ a_c \in [0, D]\,(c = 1, 2, 3, \ldots, m), \tag{5}$$

where D is a normal number; it is called the penalty factor. If the value of D is large, it means that the penalty for fitting deviation is large. At this time, the regression function can be expressed as shown below:

$$V = \sum_{c=1}^{i}\left(a_c^* - a_c\right)Q\left(x_c, x\right) + s. \tag{6}$$

The selection of regression model parameters mainly includes the selection of the type of kernel function, the parameters of the kernel function, the penalty factor *D,* and the insensitivity coefficient. The performance characteristics of the regression machine have a great relationship with these parameters.

2.1.3. Feature Structure of the Support Vector Machine Application. SVM has good generalization ability and strong theoretical support. Scholars at home and abroad have conducted much in-depth research on support vector machine algorithms and, based on this, have optimized the algorithm, so that the performance of support vector machines is continuously improved. Support vector machines are widely used in various fields, such as face recognition, image classification, note recognition, and voice recognition, in pattern recognition and many data analysis fields such as virus detection, spam filtering, and network intrusion detection [8].

(1) *Face Recognition*. The core idea of face recognition is to use knowledge or statistical methods to model the face [9]. It is more likely that the area to be inspected matches the face model in a complex background and judges whether a face exists and separates. At present, face recognition detection technology is relatively mature and applied to various fields. Osuna first proposed to use the SVM method in face recognition technology by training a nonlinear SVM classifier to detect and classify faces and nonfaces. A basic principal component analysis (PCA) was proposed. +LDA + SVM face recognition improvement framework, using particle swarm optimization algorithm to optimize the two important accommodation penalty parameters and kernel functions of SVM to obtain the optimal solution, is used to train the final classifier for face recognition and obtain more high recognition accuracy [10, 11].

(2) *Image Classification*. Images have become an important means of transmitting and obtaining information in people's life and work. Quickly positioning images and properly classifying images are very important to improve the accuracy of content-based image retrieval. Literature proposes simple image and complex image classification methods based on SVM. Literature effectively combines the idea of semisupervised learning with support vector machines and proposes a small graph classification method for label mean semisupervised SVM based on mean shift. The algorithm parameter value method is improved by the mean shift result so that the image classification result can obtain higher classification accuracy and time efficiency [12, 13].

(3) *Network Intrusion Detection*. Intrusion detection technology collects and analyzes information on key nodes in computer network systems and responds to security policy violations in a timely manner [14]. The data in network intrusion detection is very large and complex. It has the characteristics of high dimensions, small samples, and linear inseparability. SVM, as a method developed on the basis of small sample machine learning, uses the principle of risk minimization to solve problems such as small samples, nonlinearity, and high dimensions, yet is able to maintain a high level of lack of prior knowledge. The classification accuracy is very suitable for network intrusion detection systems [15, 16].

2.2. Integration of Support Vector Machines with Machine Learning

2.2.1. Kernel Function. In the previous discussion, it is believed that the samples that need to be trained satisfy the condition of linear separability in the feature space, but it is often difficult to determine the appropriate kernel function in the real task to make the training set linearly separable in the feature space [17]. For nonlinear problems, the linear separable support vector machine cannot be effectively solved, and it is necessary to use the nonlinear model in order to classify it well. Therefore, we need to introduce the concept of kernel function which refers to a symmetric function corresponding to the kernel matrix semipositive definite [18]. In other words, any kernel function implicitly defines a space called the "Regenerative Core Hilbert," the characteristic space [7]. Samples are linearly separable within the feature space, so the quality of the feature space is crucial to the performance of support vector machines. It is particularly important to note that when we do not know the form of feature maps, we do not know what kind of kernel function is appropriate, and the kernel function only implicitly defines this feature space. Thus, kernel function selection becomes the largest variable in support vector machines. If the kernel function selection is not appropriate, it means that the sample is mapped to an inappropriate feature space, which is likely to lead to poor performance and cannot get the results we want [19].

2.2.2. Soft Spacer Support Vector Machine. Although the introduction of kernel functions plays a crucial role in dividing samples of different classes, even if such kernel functions are found to make samples linearly separable in feature space, it is difficult to judge whether it is caused by overfitting. So in order to alleviate this problem, we allow SVM to have some fault tolerance on the sample; that is, the support vector machine of "hard interval" which we want to propose is different from that of "soft interval," which allows some samples not to meet the following constraints:

$$y_i\left(w^t x_i + b\right) \geq 1. \tag{7}$$

Naturally, this does not mean that our "error" samples are arbitrary and not limited by the number. In terms of requirements, we want as few samples as possible that do not satisfy the constraints, so we rewrite the optimization goal to

$$\min \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{m} l_0\left(y_i\left(w^t x_i + b\right) - 1\right). \tag{8}$$

Polynomial kernel function is as follows:

$$k\left(x_1, x_2\right) = \left(< x_1, x_2 > + R\right)^d. \tag{9}$$

Gaussian kernel function is as follows:

$$k\left(x_1, x_2\right) = \mathrm{Exp}\left(-\frac{\left\|x_1, x_2\right\|^2}{2\sigma^2}\right). \tag{10}$$

Linear kernel function is as follows:

$$k\left(x_1, x_2\right) = < x_1, x_2 >. \tag{11}$$

Support vector machine algorithm flow chart is shown in Figure 1.

In Figure 1, *Ht* is represented as an error sample, *Pc* is a sample that does not meet the constraints, and *ST* is an optimization goal. After the process in Figure 1, the fault tolerance of the support vector machine can be obtained.

2.3. Space-Time Complexity of Support Vector Machines. The research shows that when using support vector machines for classification, there are actually two main processes of training and classification. Therefore, the complexity of the discussion cannot be unified, and the space-time complexity of support vector machines is simplified, that is, the complexity of solving this quadratic programming problem: analytical solution, numerical solution [20].

2.3.1. Analytical Solution. An analytical solution is a theoretical solution. That is to say, as long as there is a solution to a problem, its analytical solution must exist. Of course, existence is one thing, and it can be solved, or it can be solved within a tolerable time frame, which is another matter. For SVM, the time complexity of finding the analytic solution is the worst, which is the number of support vectors. Although there is no fixed ratio, the number of support vectors is also related to the size of the training set.

2.3.2. Numerical Solution. The numerical solution is a solution that can be used, but in actual situations, it is often an approximate solution [21]. The process of finding a numerical solution is similar to the exhaustive method, and of course there must be certain rules to follow. Different algorithms have different ways of finding breakpoints, and even the stopping conditions are different. The accuracy of the resulting solutions is also different. It can be seen that the discussion of the complexity of the numerical solution cannot be separated from the specific algorithm analysis [22].

2.4. The Relationship and Development Trend of Big Data Platforms and Machine Learning. The core of big data is to use the value of data, and machine learning is just the key technology to better use the value of data. For big data platforms, machine learning is indispensable. Correspondingly, for machine learning, the greater the amount of data information, the higher the precise value for the data model [23]. Coincidentally, complex machine learning algorithms are trapped in the complexity of time and space and urgently need key technologies such as distributed computing and memory computing, and this key technology is the core of big data. Therefore, from the perspective of dialectics, big data and machine learning are mutually reinforcing and interdependent.

Although, nowadays, machine learning is closely connected with big data, it must be made clear that big data is not equivalent to machine learning. Similarly, machine learning is not equivalent to big data. This means that machine learning is a part of the big data analysis, and it is not the only analysis method under big data. There is no doubt that the close integration of machine learning and big data has produced huge benefits for the current social situation. Based on the development of machine learning

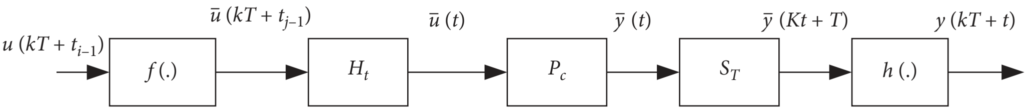

Figure 1: Boxplot of petal length.

technology, massive amounts of data information can be nearly "reasonably predicted." As far as human society is concerned, especially at the stage where the Internet is popular and developing rapidly, the richer the accumulated experience, and the broader the experience, equates to more accurate projections for the future. This is relevant to the theory in the machine learning community: the more data the machine learning model has, the better the efficiency of machine learning prediction is [24].

3. Experiments

3.1. Experimental Background. Data is the raw material for machine learning models, and the current boom in machine learning is inseparable from the support of big data. In the field of machine learning, there are a large number of public datasets available, ranging from hundreds of samples to hundreds of thousands of samples. Some datasets are used for teaching, and some are used as standards for performance testing of machine learning models. These high-quality public datasets provide great convenience for us to learn and study machine learning algorithms, similar to the value of model organisms for biological experiments. In order to test the improvement and application of different algorithms on the support vector machine, this experiment selects the most classic Iris dataset as the test object.

3.2. Experimental Setup. In this paper, we find that the time complexity of the original problem is determined by the dimension of the feature space, while the time complexity of the dual problem is determined by the number of training samples. Therefore, this paper selects a suitable solution space for the data scale and converts the solution of the dual space into the classification surface of the original space and divides them into three groups for calculation.

In order to verify the above conclusions, we use MATLAB to conduct Experiment 1, write training algorithms according to the principle of SVM, customize different dimensions and different numbers of Iris datasets for training experiments, and compare training time. The featured space dimension is defined as D and the number of sample points as C, half of which are positive samples, assigned as real numbers between 0 and 100, and half as negative. Samples were assigned as real numbers between 200 and 300.

From the results of Experiment 1, it can be seen that when the dimension/number is much less than 1, the time consumed to solve the original problem is significantly less than the time consumed to solve the dual problem. Conversely, when the dimension/number is much greater than 1, the time consumed to solve the dual problem is obviously less than the time consumed to solve the original problem.

3.3. Experimental Procedure.

(1) Prepare experimental data: import the Iris flower dataset

(2) Data feature analysis: multiangle feature analysis of the dataset

(3) Visual analysis: analyze the correlation between different features in the data set, which is the core of algorithm reformation based on support vector machine

(4) Use support vector machines for machine learning: use the correlation analysis generated in the previous experimental step to compare and test different algorithms

(5) Analyze the experimental results

4. Discussion

4.1. Experimental Data Analysis

(1) The Iris dataset contains 150 records in 3 categories, 50 data in each category, and each record has 4 features: calyx length, calyx width, petal length, and petal width, which can be predicted by these 4 features. Which species does the iris flower belong to? The basic discrimination basis for the three types of irises is that the seeds have 4 dimensions in the entire Iris dataset, namely, petal length, petal width, calyx length, and calyx width. After integrating the above algorithm improvements and performance optimizations, this paper draws the decision surfaces of the support vector machine classification with different kernels in these four dimensions. The decision boundaries of the two linear support vector machines are straight lines and nonlinear. The decision boundary of the kernel support vector machine (polynomial kernel and Gaussian radial basis kernel) is a nonlinear curve boundary, and 80% of users of the improved algorithm have been greatly improved in time and space complexity. It can be seen that in the era of big data, both the quantity and the dimensions of the data may be huge, and we should select the appropriate algorithm according to the scale and characteristics of the data. When the dimension is greater than the number, the dual problem is chosen to be solved; when the number is greater than the dimension, the original problem is

chosen to be solved. This will undoubtedly greatly improve the processing speed of the algorithm. Based on the big data platform, the dataset processed in this article will use more nonlinear support vector machines, and the summary analysis of the Iris dataset is shown in Table 1 and Figure 2.

(2) Unbalanced data classification algorithms can only solve the accuracy problem. When the data size increases, the training time of the algorithm is very long; although some distributed classification algorithms can be trained for a short time, these classification algorithms are not unbalanced data classification algorithm. Describing the distribution of data, including upper and lower bounds, upper and lower quartiles, and median, you can simply view the distribution of data. Combining with the above summary analysis, if the upper and lower quartiles are far apart, they can generally be easily divided into the following categories. As shown in Figure 3 and Table 2, the following exhibits the summary statistics of each feature column of the entire dataset.

Table 1: Comparison of experimental training time.

Parameter setting	Dimensions/ quantity	Time spent in solving the original problem (s)	Time spent in solving dual problems (s)
$D = 50$ $C = 300$	0.15	0.4562	14.3513
$D = 100$ $C = 300$	0.26	1.1367	15.5391
$D = 200$ $C = 300$	0.53	3.2782	15.6821
$D = 300$ $C = 200$	1.40	5.1125	1.3429
$D = 300$ $C = 100$	3.00	3.2381	0.2713
$D = 300$ $C = 50$	5.00	2.1357	0.2731

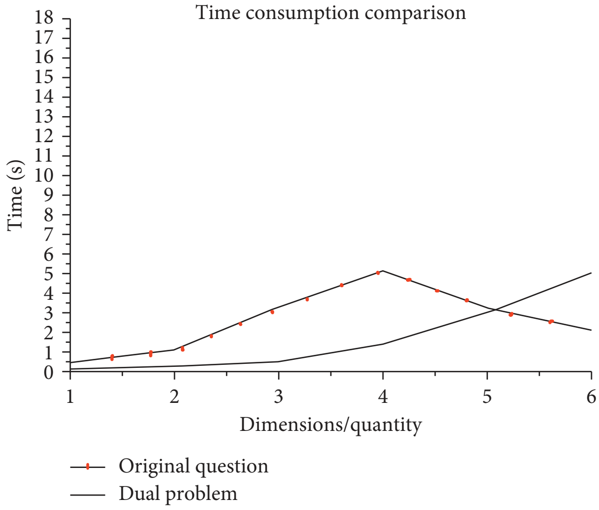

Figure 2: Comparison of experimental training time.

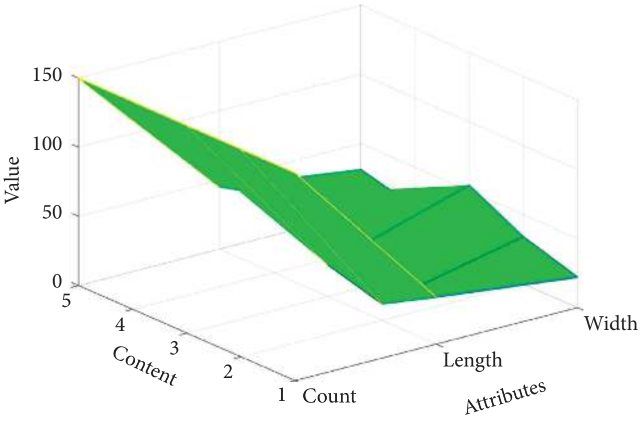

Figure 3: Summary information.

Table 2: Summary information.

Content	Count	Length	Width
Sepal length (cm)	150	34	22
Sepal width (cm)	150	12	34
Petal length (cm)	150	23	54
Petal width (cm)	150	56	21
Species	150	45	31

4.2. Analysis of Data Characteristics

(1) When analyzing the relationship between features and varieties through data distribution, in order to have a deeper understanding of the dataset, this article must explore the relationship between various variables, specifically in the Iris dataset, and observe the characteristics and varieties. The relationship is shown in Figure 4. In this research, by analyzing the relationship between each feature and variety (since this is a binary variable), the linear relationship is used to analyze the relationship between the feature and the variety; then for the linear relationship, the slope relationship becomes its investigation, the core. As shown in Figure 4, the relationship between each feature and variable is compared by slope, and the relationship between each feature is analyzed. Through the above analysis, the experiment needs to perform feature analysis on all variables in the data set in order to facilitate subsequent experiments get on. The specific data is shown in Table 3 of Figure 4.

(2) Use Andrews Curves to convert each multivariate observation to a curve and express the coefficients of the Fourier series, which is useful for detecting outliers in time series data. Andrews Curves is a method to visualize multidimensional data by mapping each observation to a function. After conducting a general analysis based on the linear regression visualization of the calyx and petals, this paper analyzes the main variables. After the above analysis, this paper finds the correlation between different features in the dataset. A high positive or negative value indicates that the features are highly correlated. This is shown in Table 4. Through the analysis of the experimental results in this paper, it is

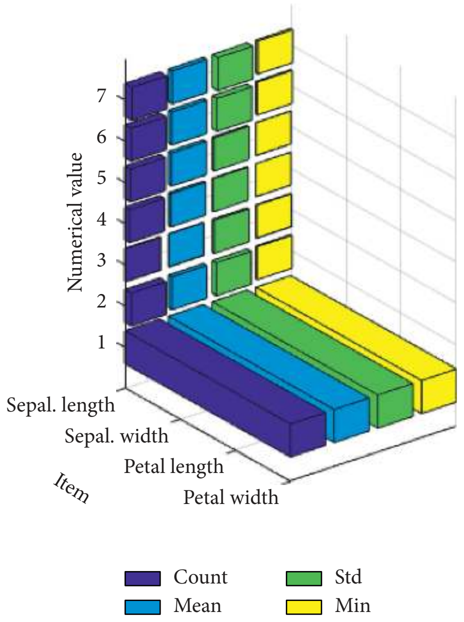

Figure 4: Boxplot of petal length.

Table 3: Feature column summary statistics.

Item	Sepal length	Sepal width	Petal length	Petal width
Count	150.000000	150.000000	150.000000	150.000000
Mean	5.843333	3.057333	3.758000	1.199333
Std	0.828066	0.435866	1.765298	0.762238
Min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
Max	7.900000	4.400000	6.900000	2.500000

Table 4: Data distribution analysis of the relationship between characteristics and varieties.

Vector 1	Vector 2	Vector 3
1.2	2.2	1.9
1.3	2.3	2.2
2.6	1.6	1.8
1.8	2.7	1.7
2.4	2.8	2.9

found that the length and width of the flowers are not related, and the length and length of the petals have a strong correlation. Through the above analysis, it can be found that, after three sets of calculations, the accuracy of the prediction between the predicted data and the actual value is 98% (including two average tests), indicating that the improvement of the algorithm here is for the support vector machine The rewriting is very effective. In addition to the above decision tree analysis addition, the classification method of support vector machine needs to be further improved, then as a classic traditional clustering method KMeans cluster analysis is a very good idea analysis, this paper will experiment with the algorithm improvement results. Based on the above two-classification methods, K-means clustering and decision tree analysis, it is found through experimental results that the two can be more effectively combined to achieve algorithm optimization and process improvement in support vector machines. Such optimization results are on big data platforms which is also helpful. Figures 5 and 6 show the analysis of making a linear regression visualization.

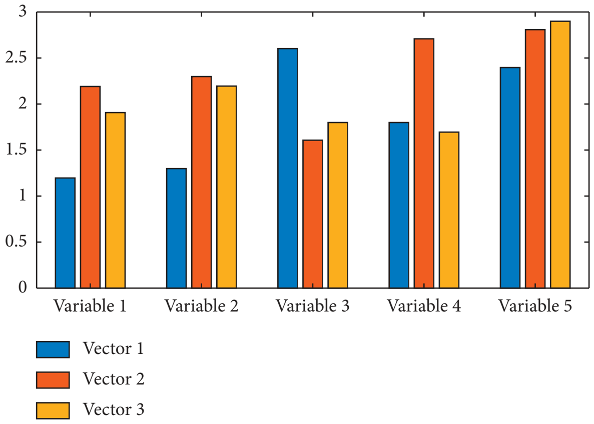

Figure 5: Data distribution analysis of the relationship between characteristics and varieties.

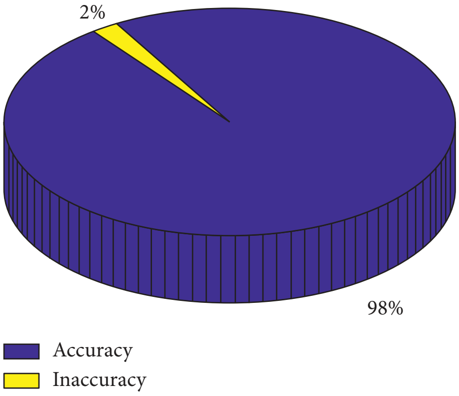

Figure 6: Visual analysis of linear regression based on calyx.

5. Conclusion

(1) As a very efficient classification model in machine learning, support vector machine has many advantages such as good generalization, few parameters, and the ability to generate global optimal solutions. It is a very good choice for people to process data, analyze data, and predict data. In the context of today's big data, support vector machines, as a traditional classification method, are still applicable due to the superiority of their architecture and algorithms. However, people must improve their algorithm to get rid of the process of processing large sample data sets. This algorithm will occupy a long training time and occupy a high space-time complexity, which leads to the problem of low efficiency. Judging from the final experimental results, the problems we raised have been effectively resolved.

(2) This article aimed to study the improvement of support vector machine algorithms in the context of big data, and the research discovered problems. This article intended to solve the problem of traditional SVM sensitive to noise points and outliers and could not solve the problem of large sample feature scale, heterogeneous information, and feature space. We then have the problem of uneven distribution. During the research process, incorporating fuzzy membership into multicore learning, it is discovered that the time complexity of the original problem is determined by the dimension, and the time complexity of the dual problem is determined by the quantity of the data. The dimension and quantity constitute the scale of the data, so it can be based on the features we chose for different solution spaces. The algorithm speed can be improved by transforming the solution of the dual problem into the classification surface of the original space. To conclude, by improving the calculation rate of traditional machine learning algorithms, it is established that the accuracy of the fitting prediction between the predicted data and the actual value is as high as 98%, which can enable traditional machine learning algorithms to meet the requirements of the big data era in the future. It can be widely used in the context of big data.

(3) This article used statistical learning and optimization theory to analyze the working process of support vector machines and the needs of big data platforms for support vector machines in principle, while updating and improving the kernel algorithm of support vector machines. In the experimental stage, the classic optimization dataset Iris flower dataset is used to further test the improved optimization algorithm of support vector machine. The final experimental results are a good evidence of the good fit and adaptability of researchers to improve and optimize the algorithm. In the process of improvement, it was found that many different variant models can be derived when small sample classification is performed. And the support vector machine itself also has a significant advantage as it can handle various complex operations of vector inner product in high-dimensional space through kernel functions. Therefore, it is necessary to choose a suitable solution space and a suitable kernel function for the data scale and convert the solution of the dual space into the classification surface of the original space. The final experimental results show that by further improving the calculation rate of traditional machine learning algorithms, the traditional support vector machine algorithm can meet the requirements of the era of big data. The time complexity of the original problem is determined by the dimension of the feature space, while the time complexity of the dual problem is determined by the number of training samples. Therefore, people should choose a suitable solution space for the data scale and convert the solution of the dual space into the classification surface of the original space. The final experimental results show that by improving the calculation rate of traditional machine learning algorithms, the traditional machine learning algorithms can meet the requirements

of the era of big data. Experimental data shows that 80% of users of the improved algorithm have greatly improved the complexity of time and space, showing a good fit and applicability.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare no potential conflicts of interest in our paper.

Authors' Contributions

All authors have seen and approved the final version of the manuscript.

References

[1] P. Borah and D. Gupta, "Unconstrained convex minimization based implicit Lagrangian twin extreme learning machine for classification (ULTELMC)," *Applied Intelligence*, vol. 50, no. 4, pp. 1327–1344, 2020.

[2] P. Borah and D. Gupta, "Functional iterative approaches for solving support vector classification problems based on generalized Huber loss," *Neural Computing and Applications*, vol. 32, no. 1, pp. 1135–1139, 2020.

[3] S. Balasundaram and D. Gupta, "Knowledge-based extreme learning machines," *Neural Computing and Applications*, vol. 27, no. 6, pp. 1629–1641, 2016.

[4] E. Carrizosa, A. Nogales-Gómez, and D. Romero Morales, "Strongly agree or strongly disagree?: rating features in support vector machines," *Information Sciences*, vol. 329, no. C, pp. 256–273, 2016.

[5] G. Taherzadeh, Y. Zhou, A. W.-C. Liew, and Y. Yang, "Sequence-based prediction of protein-carbohydrate binding sites using support vector machines," *Journal of Chemical Information and Modeling*, vol. 56, no. 10, pp. 2115–2122, 2016.

[6] M. Tanveer, M. A. Khan, and S.-S. Ho, "Robust energy-based least squares twin support vector machines," *Applied Intelligence*, vol. 45, no. 1, pp. 174–186, 2016.

[7] W. Gu, W.-P. Chen, and C.-H. Ko, "Two smooth support vector machines for ε-insensitive regression," *Computational Optimization & Applications*, vol. 70, no. 1, pp. 1–29, 2018.

[8] T. Tanino, R. Kawachi, and M. Akao, "Performance evaluation of multiobjective multiclass support vector machines maximizing geometric margins," *Numerical Algebra Control & Optimization*, vol. 1, no. 1, pp. 151–169, 2017.

[9] M. Malvoni, M. G. De Giorgi, and P. M. Congedo, "Data on support vector machines (SVM) model to forecast photovoltaic power," *Data in Brief*, vol. 9, no. C, pp. 13–16, 2016.

[10] R. Darnag, B. Minaoui, and M. Fakir, "QSAR models for prediction study of HIV protease inhibitors using support vector machines, neural networks and multiple linear regression," *Arabian Journal of Chemistry*, vol. 10, no. S1, pp. S600–S608, 2017.

[11] J.-Y. Gotoh and S. Uryasev, "Support vector machines based on convex risk functions and general norms," *Annals of Operations Research*, vol. 249, no. 1-2, pp. 1–28, 2017.

[12] T. Singh, F. Di Troia, and C. Aaron Visaggio, "Support vector machines and malware detection," *Journal of Computer Virology & Hacking Techniques*, vol. 41, no. 10, pp. 1–10, 2016.

[13] J. Li, Y. Cao, and Y. Wang, "Online learning algorithms for double-weighted least squares twin bounded support vector machines," *Neural Processing Letters*, vol. 45, no. 1, pp. 1–21, 2016.

[14] C. Ehrentraut, M. Ekholm, H. Tanushi, J. Tiedemann, and H. Dalianis, "Detecting hospital-acquired infections: a document classification approach using support vector machines and gradient tree boosting," *Health Informatics Journal*, vol. 24, no. 1, pp. 24–42, 2016.

[15] X. Zhang, Y. Li, and X. Peng, "Brain wave recognition of word imagination based on support vector machines," *Chinese Journal of Aerospace Medicine*, vol. 14, no. 3, pp. 277–281, 2016.

[16] J. Nalepa and M. Kawulok, "Selecting training sets for support vector machines: a review," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 857–900, 2019.

[17] A. Gangopadhyay, O. Chatterjee, and S. Chakrabartty, "Extended polynomial growth transforms for design and training of generalized support vector machines," *IEEE Transactions on Neural Networks & Learning Systems*, vol. 29, no. 5, pp. 1–14, 2018.

[18] Y. Bai and X. Yan, "Conic relaxations for semi-supervised support vector machines," *Journal of Optimization Theory and Applications*, vol. 169, no. 1, pp. 299–313, 2016.

[19] L. Zhang, X. Lu, and C. Lu, "National matriculation test prediction based on support vector machines," *Journal of University of Science & Technology of China*, vol. 47, no. 1, pp. 1–9, 2017.

[20] M. Ahmer, A. Shah, S. M. Zafi S. Shah et al., "Using non-linear support vector machines for detection of activities of daily living," *Indian Journal of Science and Technology*, vol. 10, no. 36, pp. 1–8, 2017.

[21] K. H. Yoo, Y. D. Koo, H. B. Ju, and M. G. Na, "Identification of LOCA and estimation of its break size by multiconnected support vector machines," *IEEE Transactions on Nuclear Science*, vol. 64, no. 10, p. 1, 2017.

[22] Y. Lou, Y. Liu, J. K. Kaakinen, and X. Li, "Using support vector machines to identify literacy skills: evidence from eye movements," *Behavior Research Methods*, vol. 49, no. 3, pp. 887–895, 2017.

[23] A. U. Mageswari and R. Vinodha, "Engine knock detection based on wavelet packet transform and sparse fuzzy least squares support vector machines (SFLS-SVM)," *IIOAB Journal*, vol. 7, no. 11, pp. 194–199, 2016.

[24] M. Erdem, F. E. Boran, and D. Akay, "Classification of risks of occupational low back disorders with support vector machines," *Human Factors and Ergonomics in Manufacturing & Service Industries*, vol. 26, no. 5, pp. 550–558, 2016.