Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 4097660, 8 pages
https://doi.org/10.1155/2023/4097660

Research Article

Study on Gene Splicing Site Recognition Based on Particle Swarm Optimization Twin Support Vector Machine Algorithm for Smart Healthcare

Fuquan Zhang[1], **Yiou Wang**[2], **Peng Mei**[3], **Aibing Dai**[2], **Bo Wang**[4], **Laiyang Liu**[5] and **Yong Xia**[6,7]

[1]*Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350117, China*
[2]*Institute of Science and Technology Information, Beijing Academy of Science and Technology, Beijing 100089, China*
[3]*E-Government Research Center, Party School of the Central Committee of CP.C (National Academy of Governance), Beijing 100089, China*
[4]*Industrial Digital Finance Department, Huaxia Bank, Beijing 100020, China*
[5]*Digital Performance and Simulation Technology Lab, Beijing Institute of Technology, Beijing 100081, China*
[6]*Fuzhou Maternity and Child Health Care Hospital, Fuzhou 350005, China*
[7]*Fujian Medical University, Fuzhou 350122, China*

Correspondence should be addressed to Yiou Wang; wangyiou90@163.com and Peng Mei; meipeng1114@163.com

Received 2 August 2022; Revised 18 September 2022; Accepted 23 September 2022; Published 21 April 2023

Academic Editor: Chao-Yang Lee

Gene splicing site recognition is a very important research topic in smart healthcare. Gene splicing site recognition is of great significance, not only for the large-scale and high-quality computational annotation of genomes but also for the analysis and recognition of the gene sequences evolutionary process. It is urgent to study a reliable and effective algorithm for gene splice site recognition. Traditional Twin Support Vector Machine (TWSVM) algorithm has advantages in solving small-sample, nonlinear, and high-dimensional problems, but it cannot deal with parameter selection well. To avoid the blindness of parameter selection, particle swarm optimization algorithm was used to find the optimal parameters of twin support vector machine. Therefore, a Particle Swarm Optimization Twin Support Vector Machine (PSO-TWSVM) algorithm for gene splicing site recognition was proposed in this paper. The proposed algorithm was compared with traditional Support Vector Machine algorithm, TWSVM algorithm, and Least Squares Support Vector Machine algorithm. The comparison results show that the positive sample recognition rate, negative sample recognition rate, and correlation coefficient (CC) of the proposed algorithm are the best among the four different support vector machine algorithms. The proposed algorithm effectively improves the recognition rate and the accuracy of splice sites. The comparison experiments verify the feasibility of the proposed algorithm.

1. Introduction

With the development of big data, Internet of Things, cloud computing, artificial intelligence, and other information technologies, smart healthcare has emerged [1]. Among the many research fields of smart healthcare, gene splicing site recognition is a very important research field. In the promotion of the Human Genome Project (HGP), the volume of data in molecular biology has exploded [2]. Besides, high-throughput sequencing technologies are developed, and the cost of sequencing is reduced, which lead to the accumulation of massive genome data and the crazy growth rate of the data [3]. Now, biological genetic resources are greatly enriched, but some new challenges are also brought to people. Traditional recognition methods no longer meet the current needs of processing massive genome data. It has become a research trend to use computer as a tool to study genome data with the help of machine learning theory.

The application of machine learning technology in the field of genes has also been an important means to promote smart healthcare.

Genome research has entered the postgenome era. At this time, the focus of gene research has gradually shifted from gene sequencing to gene expression analysis [4]. Accurate recognition of splicing sites from gene sequences is very important for understanding gene expression. If the splicing sites in the gene sequences of eukaryotic organisms are accurately recognized, the expressed and nonexpressed regions of genes are distinguished. A splicing site is a short and highly consistent sequence of conservative features, located at the junction between exon and intron in the eukaryotic cell, as shown in Figure 1. Most splicing sites conform to the rule of "GT-AG" [5, 6]. At present, a difficult problem is how to recognize the correct splicing site from a large amount of gene data.

The recognition of gene splicing sites is a sequence recognition problem, which belongs to the intersection of biological information and sequence recognition. Some researchers have tried many algorithms of machine learning to recognize gene splicing sites and made great progress. Sharma et al. [7] proposed an acceptor site recognition method, which combined Adaptive Short Time Fourier Transform (ASTFT), period-3 measure, and with principal component analysis algorithm. Morfino et al. [8] dealt with the splicing site recognition problem in DNA sequences by using supervised machine learning algorithms included in the MLlib library of Apache Spark, a fast and general engine for big data processing. L. Wang et al. [9] proposed a recognition algorithm on slicing site based on improved SVM. The recognition sensitivity of SVM was optimized by kernel principal component analysis. R. Wang et al. [10] designed SpliceFinder based on the Convolutional Neural Network (CNN) to recognize splice sites. They used human genomic data to train a neural network. An iterative approach was adopted by them to reconstruct the dataset, which tackled the data unbalance problem and forces the model to learn more features of splice sites. Pashaei and Aydin [11] proposed the Markovian encoding models in human splice site recognition using SVM algorithm and developed the MMSVM algorithm that is a web tool to recognize splice sites in any human sequence. Javier et al. [12] presented a methodology for combining many sources of information to recognize any functional site using 'floating search,' a powerful heuristic applicable when the cost of evaluating each solution is high. Waseem et al. [13] developed a computational model for splicing sites identification. The three feature extraction methods were employed (i.e., DNC, TNC, and TetraNC) to extract features from DNA sequences and are then combined to develop a composite features space. Scalzitti et al. [14] developed the Spliceator to predict splice sites in a wide range of species, including model and nonmodel organisms. Spliceator used a convolutional neural network and is trained on carefully validated data from over 100 organisms. Zhang et al. [15] introduced the DeepPN, a deep parallel neural network that is constructed with a convolutional neural network (CNN) and graph convolutional network (GCN) for detecting RBPs binding sites. Ghazanfari et al. [16] used all the data and valuable information such as isoform sequences, expression profiles, and gene ontology graphs and proposes a comprehensive model based on deep neural networks. Shen et al. [17] proposed the CircCNN to predict pre-mRNA back-splicing sites. The convolution neural network and batch normalization were the main parts of CircCNN. However, although these algorithms realize the recognition of gene splicing sites to some extent, the computational cost and recognition accuracy of these methods are still not satisfactory.

Twin Support Vector Machine (TWSVM) is a machine learning algorithm based on statistical learning theory [18]. As a deformation algorithm of SVM, TWSVM algorithm inherits the excellent learning ability of SVM algorithm, moreover, the operating efficiency of TWSVM algorithm is improved by 4 times. However, there are still some shortcomings in TWSVM algorithm, one of which is that TWSVM algorithm cannot deal with parameter selection well. Parameter selection has a great influence on recognition results. Inappropriate parameters will seriously reduce the recognition accuracy of TWSVM algorithm for gene splicing sites. The Particle Swarm Optimization (PSO) algorithm has a good effect on parameter optimization. Therefore, the PSO algorithm was combined with the TWSVM algorithm in this paper to overcome the shortcomings of TWSVM algorithm. An improved gene splicing site recognition algorithm based on the Particle Swarm Optimization Twin Support Vector Machine (PSO-TWSVM) was proposed. The proposed algorithm not only solved the problem of parameter selection but also further improved the prediction accuracy of the traditional algorithm. Firstly, splicing site recognition was regarded as a machine learning problem of dichotomy. The recognition was completed by analyzing the sequence characteristics near the splicing sites. Then, aiming at the difficult problem of parameter setting of TWSVM algorithm, the PSO algorithm was used to optimize the parameters of TWSVM algorithm to further improve the recognition performance of splicing sites.

The rest of this paper was organized as follows. Some background information was introduced in Section 2, including the TWSVM algorithm and the PSO algorithm. To overcome the shortages of TWSVM, an improved algorithm of PSO-TWSVM was proposed and the recognition steps of the proposed PSO-TWSVM algorithm for gene splicing sites were shown in Section 3. Comparative experiments of the proposed PSO-TWSVM algorithm, traditional TWSVM algorithm and Least Squares Support Vector Machine (LSSVM) algorithm were performed, and the experimental results were analyzed in Section 4. Finally, a conclusion of this paper was made, and some future research work was given in Section 5.

2. Related Work

2.1. TWSVM Algorithm. On the basis of Proximal Support Vector Machine based on Generalized Eigenvalues (GEPSVM), the TWSVM algorithm was proposed by Jayadeva et al. in 2007 [19]. Different from traditional SVM algorithm, the TWSVM algorithm aims to find a pair of uneven

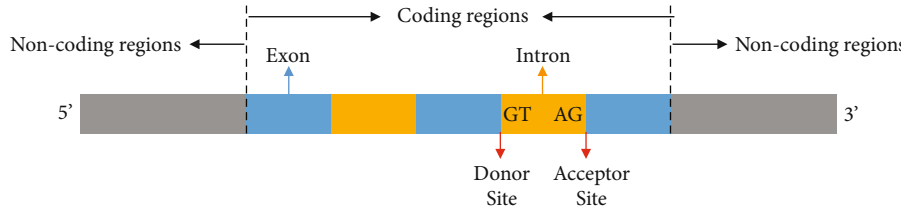

Figure 1: The gene structure of eukaryotes.

hyperplanes by solving two small quadratic programming problems [20, 21]. In real life, the recognition of gene samples is a very difficult problem. Linear TWSVM algorithm cannot accurately analyze the feature space of gene data. Kernel function was introduced into TWSVM algorithm to solve nonlinear analysis and recognition problem [22].

Assume that (x_j^i, y_j) represents the training sample set in the $n-$dimension real space R^n, where $i = 1, 2$ and $j = 1, 2, \cdots, m$. m is the total number of samples, and $m = m_1 + m_2$, where m_1 is the number of positive samples and m_2 is the number of negative samples. Then, the nonlinear hyperplane of TWSVM algorithm is calculated as follows [23].

$$\begin{cases} K\left(x^T, C^T\right)u_1 + b_1 = 0, \\ K\left(x^T, C^T\right)u_2 + b_2 = 0, \end{cases} \tag{1}$$

where u is the normal vector of the hyperplane, b is the offset sample, and K is the kernel function. Now, the GAUSSIAN radial basis function is used as the kernel function of the TWSVM algorithm. $C^T = [A\ B]^T$, where A is the $m_1 \times m$ positive sample matrix, and B is the $m_2 \times m$ negative sample matrix. Then, the plane dividing positive samples and negative samples is obtained by solving two quadratic programming problems [24].

$$\begin{aligned} &\min_{u_1 b_1 \xi} \frac{1}{2}\left\|K\left(A, C^T\right)u_1 + e_1 b_1\right\|^2 + c_1 e_2^T \xi, \\ &s.t. -\left(K\left(B, C^T\right)u_1 + e_2 b_1\right) + \xi \geq e_2, \xi \geq 0, \end{aligned} \tag{2}$$

$$\begin{aligned} &\min_{u_2 b_2 \xi} \frac{1}{2}\left\|K\left(B, C^T\right)u_2 + e_2 b_2\right\|^2 + c_2 e_1^T \xi, \\ &s.t. -\left(K\left(A, C^T\right)u_2 + e_1 b_2\right) + \xi \geq e_1, \xi \geq 0, \end{aligned} \tag{3}$$

where $e_1 = (1, \cdots, 1)^T \in R^{m_1}$, $e_2 = (1, \cdots, 1)^T \in R^{m_2}$, c_1, and c_2 are the penalty factors which are used to control the degree of punishment for missampling. To simplify Formulas (2) and (3), the dual transformation is carried out.

$$\begin{aligned} &\max_{\alpha} e_2^T \alpha - \frac{1}{2}\alpha^T R\left(S^T S\right)^{-1} R^T \alpha, \\ &s.t. 0 \leq \alpha \leq c_1 e_2, \end{aligned} \tag{4}$$

$$\begin{aligned} &\max_{\gamma} e_1^T \gamma - \frac{1}{2}\gamma^T S\left(R^T R\right)^{-1} S^T \gamma, \\ &s.t. 0 \leq \gamma \leq c_2 e_1, \end{aligned} \tag{5}$$

where $R = [K(B, C^T)e_2]$ and $S = [K(A, C^T)e_1]$. The following results are obtained by solving Formulas (4) and (5).

$$\left(u_1^T, b_1\right)^T = -\left(S^T S\right)^{-1} R^T \alpha, \tag{6}$$

$$\left(u_2^T, b_2\right)^T = -\left(R^T R\right)^{-1} S^T \gamma. \tag{7}$$

According to Formulas (6) and (7), the hyperplane is obtained, and its decision function is shown in Formula (8).

$$\text{classlabel} = \arg \min_{k=+1,-1} \left|K\left(x^T, C^T\right)u_k + b_k\right|. \tag{8}$$

Compared with the SVM algorithm, TWSVM algorithm has better recognition ability. When the data is unbalanced, that is, the numbers of positive samples and negative samples are much different, SVM often fails to achieve the expected results. However, TWSVM still performs well in this case, mainly because the TWSVM has two penalty factors in the hyperplane, which can adjust the penalty degree of the hyperplane at the same time [25]. The recognition results of the TWSVM algorithm are closer to the true values of the sequence. In addition, as TWSVM algorithm solves two SVM problems, its computing efficiency is higher. The time complexity of SVM algorithm is $O(m^3)$, while that of TWSVM algorithm is $O(2 \times (m/2)^3)$, where m is the number of samples. The computing overhead of TWSVM algorithm is about 1/4 of that of the SVM algorithm.

2.2. PSO Algorithm. The Particle Swarm Optimization (PSO) algorithm was proposed by Eberhart, an American electrical engineer, and Kennedy, a social psychologist, in 1995 [26]. PSO algorithm is a process of continuous optimization by simulating the foraging behavior of birds, starting from random solutions, and iteratively updating the speed and position of particles [27]. Particles change their flight speed and position by learning their own experience (personal optimization) and social experience (global optimization) [28].

Assume N represents the population size of the $D-$dimension search space. Then, the position of the ith particle is expressed as $X_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,D})$ and the velocity of the ith particle is expressed as $V_i = (v_{i,1}, v_{i,2}, \cdots, v_{i,D})$, respectively. And the updating formula of the ith particle are shown below.

$$\begin{aligned} v_{i,d}(t+1) = \text{wv}_{i,d}(t) &+ c_1 r_1\left(\text{pbest}_{i,d}(t) - x_{i,d}(t)\right) \\ &+ c_2 r_2\left(\text{gbest}_d(t) - x_{i,d}(t)\right), \end{aligned} \tag{9}$$

$$x_{i,d}(t+1) = x_{i,d}(t) + v_{i,d}(t+1), \tag{10}$$

where $v_{i,d}(t+1)$ and $x_{i,d}(t+1)$ are the velocity and the position of the ith particle of the $d-$dimension in $t+1$ iteration, respectively, $\text{pbest}_{i,d}(t)$ and $\text{gbest}_d(t)$ are the positions of personal optimal particle and population optimal particle in t iteration, respectively. W is the inertia weight, c_1 and c_2 are learning coefficients, and r_1 and r_2 are random numbers that uniformly distributed between $[0, 1]$. c_1 and r_1 are used to regulate the maximum stride length of particles flying to the personal optimization, while c_2 and r_2 are used to regulate the maximum stride length of particles flying to the global optimization.

To effectively adjust the global and local search capabilities of the algorithm, w linear reduction is widely used.

$$w = w_{\max} - \frac{w_{\max} - w_{\min}}{T} t, \tag{11}$$

where t is the current number of iterations, T is the total number of iterations, and $w_{\max}$ and $w_{\min}$ are the maximum weight and the minimum weight, respectively.

3. PSO-TWSVM Algorithm for Gene Splicing Site Recognition

3.1. PSO-TWSVM Algorithm. The TWSVM algorithm has advantages in solving small-sample, nonlinear, and high-dimensional problems. However, the parameter selection of the TWSVM algorithm is a difficult problem. According to the calculation process of TWSVM algorithm, there are three parameters to be determined, including penalty factor c_1, penalty factor c_2, and kernel parameter of GAUSSIAN radial basis function σ. The calculation formula of GAUSSIAN radial basis kernel function $K(x, x')$ is shown as follows.

$$K\left(x, x'\right) = \exp\left(\frac{-\left\|x - x'\right\|^2}{\sigma^2}\right). \tag{12}$$

When ε is determined, finding appropriate parameter values of c_1, c_2, and σ^2 is very important to the performance of TWSVM algorithm, which directly affects the recognition accuracy of gene splicing sites. The effects of parameters to TWSVM algorithm are analyzed as below.

3.1.1. Penalty Factors c_1 and c_2. The function of penalty factors c_1 and c_2 is to achieve a certain compromise by adjusting the balance between confidence risk and empirical risk of TWSVM algorithm, so that the generalization ability of TWSVM algorithm reaches the best [29]. When the values of c_1 and c_2 are small, it means that the penalty degree to the empirical error is small. At this point, the complexity of TWSVM algorithm is low, and the data fitting degree of TWSVM algorithm is weak. On the contrary, when the values of c_1 and c_2 are large, the complexity of TWSVM algorithm is high, and the data fitting degree of TWSVM algorithm is strong. Nonetheless, overfitting is easy to occur, and the generalization ability of TWSVM algorithm is weak.

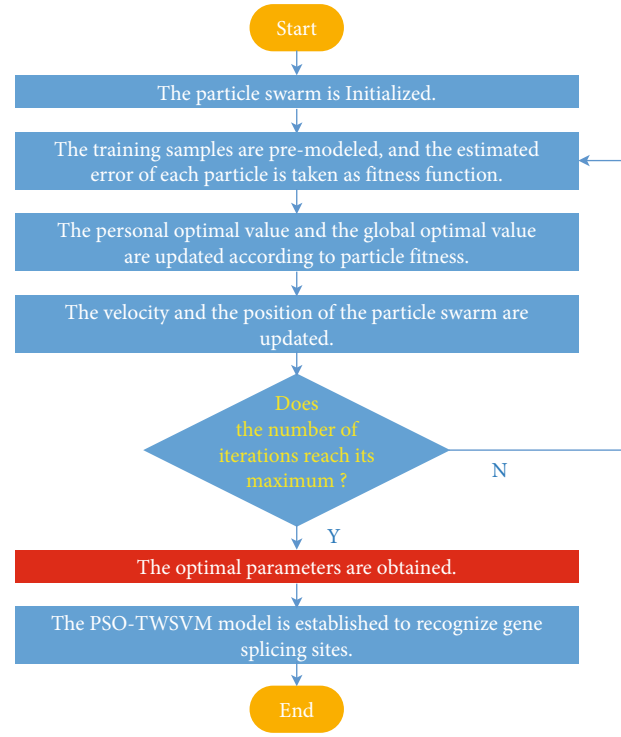

Figure 2: The flow chart of PSO-TWSVM algorithm.

3.1.2. The Kernel Function Parameter σ. The kernel function parameter σ determines the learning degree of TWSVM algorithm and affects the distribution complexity of sample data in high-dimensional space [30]. When the value of σ^2 is small, the output interval corresponding to the sample is small, and the empirical risk corresponding to the optimal hyperplane is small. At this point, the structural risk of TWSVM algorithm is large and TWSVM algorithm is prone to overfitting. On the contrary, when the value of σ^2 is large, the model of TWSVM algorithm is complex and does not generalize well.

The parameters of TWSVM algorithm play a key role in the stability and generalization of the recognition model. Cross-validation is usually adopted for parameter selection in TWSVM algorithm, but the effect is not ideal [31]. The parameter selection of TWSVM algorithm is a dynamic optimization process. Fortunately, PSO algorithm can find the global optimal solution with high probability in the process of multiobjective optimization. Besides, the structure of PSO is simple. The calculation efficiency, the solution accuracy and the robustness of PSO algorithm are good. Therefore, the Particle Swarm Optimization Twin Support Vector Machine (PSO-TWSVM) algorithm is proposed to use PSO algorithm optimize the parameters of TWSVM and to further improve the recognition accuracy of splicing sites.

The main idea of the PSO-TWSVM algorithm is to select appropriate values of penalty factor c_1, penalty factor c_2, and the kernel function parameter σ of TWSVM by PSO optimization, and then establish PSO-TWSVM model through the optimal parameters and the training sample. The PSO-TWSVM algorithm takes the penalty factors and the kernel function parameter as the initial positions of particles. After

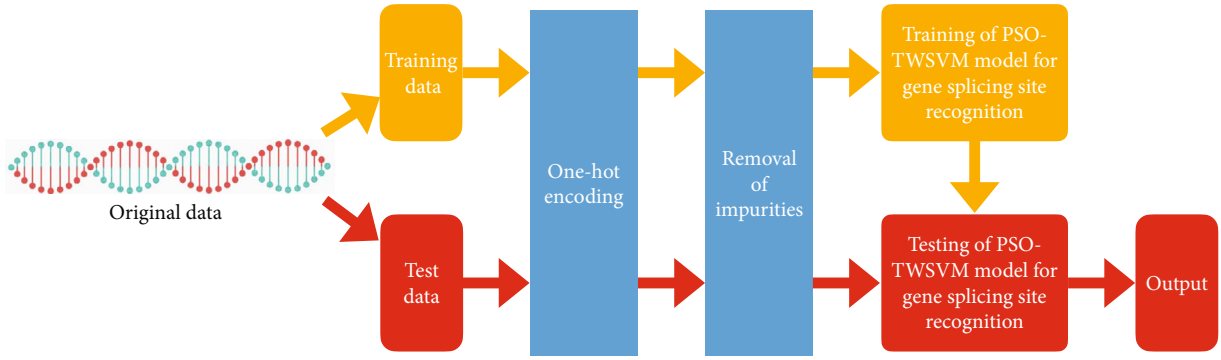

Figure 3: Gene splicing site recognition model based on the PSO-TWSVM algorithm.

Table 1: Splicing site sample dataset.

Samples	True samples	False samples
Training samples of the acceptor sites	2000	1200
Test samples of the acceptor sites	2000	1200
Training samples of the donor sites	2000	1100
Test samples of the donor sites	2000	1100

Table 2: Recognition results of different support vector machines on acceptor sites.

Four different algorithms	Positive sample recognition rate (%)	Negative sample recognition rate (%)	*CC* (%)
SVM	89.33	85.97	77.01
TWSVM	91.42	88.02	79.12
LSSVM	90.02	86.53	76.31
PSO-TWSVM	92.26	89.13	80.65

Table 3: Recognition results of different support vector machines on donor sites.

Four different algorithms	Positive sample recognition rate (%)	Negative sample recognition rate (%)	*CC* (%)
SVM	94.28	90.29	85.41
TWSVM	95.13	91.33	87.60
LSSVM	94.40	88.68	85.21
PSO-TWSVM	96.27	92.05	88.93

the optimal parameters are found by PSO algorithm, the mathematical model of TWSVM for gene splicing site recognition is established. The flow chart of PSO-TWSVM algorithm is shown in Figure 2. The calculation steps of PSO-TWSVM algorithm are shown as follows.

Step 1. A particle swarm is initialized. The position of the ith particle is expressed as a three-dimensional vector $X_i = (x_{i,1}, x_{i,2}, x_{i,3})$, where $x_{i,1}$ and $x_{i,2}$ represent two penalty factors, respectively, and $x_{i,3}$ represents the kernel function parameter σ.

Step 2. The training samples are premodeled, and the estimated error of each particle is taken as fitness function.

Step 3. According to the particle fitness, the personal optimal value and the global optimal value are updated.

Step 4. The speed and the position of the particle swarm are updated.

Step 5. Determine whether the termination condition is met, that is, whether the maximum number of iterations is reached: If it is met, the optimal parameter value of TWSVM algorithm is obtained; otherwise, return to Step 2.

Step 6. The PSO-TWSVM model is established by using the optimal parameters and training samples.

3.2. Application of PSO-TWSVM Algorithm for Gene Splicing Site Recognition. The proposed PSO-TWSVM algorithm is applied for gene splicing site recognition. First, the PSO-TWSVM model is trained using the training samples of gene splicing sites. Then, the trained PSO-TWSVM model is used to recognize the test samples of unknown gene splicing sites. Gene splicing site recognition model based on PSO-TWSVM algorithm is shown in Figure 3. The recognition steps of PSO-TWSVM algorithm for gene splicing sites are shown as follows.

Step 1. The gene data are divided into training samples and test samples.

Step 2. Both the training samples and the test samples are coded using One-Hot Encoding to convert the gene data into downstream numerical representations that can be processed by machine learning. The basic principle of One-Hot Encoding is to convert the variable to binary representation containing only 0 and 1, which is widely used in biological sequence processing. Since each DNA sequence is composed of four deoxynucleotides of adenine (A), guanine (G), cytosine (C), and thymine (T), the four-digit unique One-Hot Encoding is adopted. The deoxynucleotides

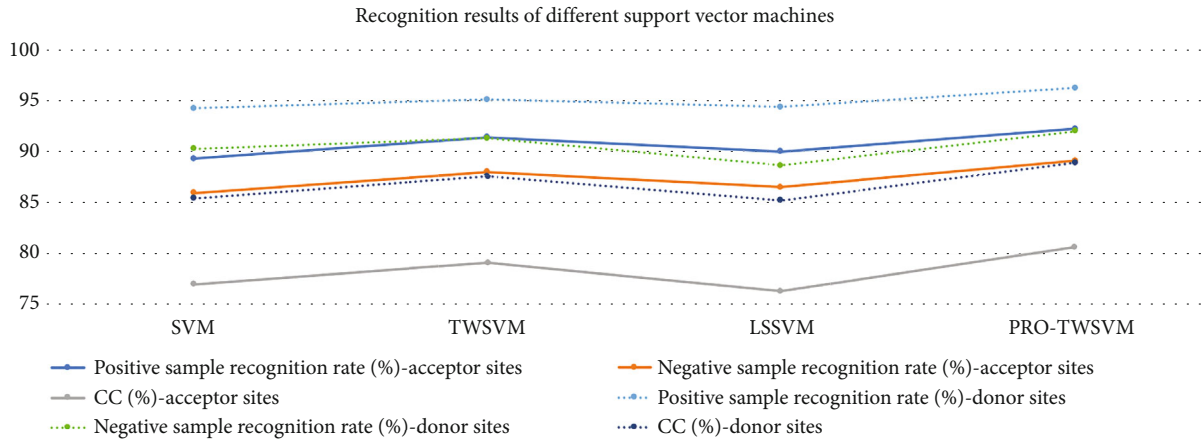

Figure 4: Recognition results of different support vector machines.

of "A," "G," "C," and "T" are encoded as "1000," "0100," "0010," and "0001," respectively.

Step 3. Impurity items are removed from both the training samples and the test samples. Very few impurity items other than "A," "G," "C," and "T," such as "N," "Q," and "W," may be mixed in the gene data. If these impurity items are deleted, they may lead to the unauthenticity of the genetic data and increase the meaningless workload. Therefore, the impurity items are represented by the One-Hot Encoding "0000," which not only removes the impurity items but also ensures the consistency of the data structure.

Step 4. Training samples are used to train the PSO-TWSVM model, so as to obtain the correct PSO-TWSVM model suitable for this set of gene data.

Step 5. Test samples are used to test the recognition performance of PSO-TWSVM model for such gene data.

4. Comparative Experiments of Gene Splicing Site Recognition

4.1. Description of the Dataset. The experiments were performed on the well-known benchmark Homo Sapiens Splice Site Dataset (HS^3D), which is available at URL: http://www.sci.unisannio.it/docenti/ to assess the recognition performance of the proposed PSO-TWSVM algorithm. HS^3D is the sequence dataset of gene splicing sites extracted from GeneBank DNA sequence database by Pollastro from Italy [32]. The length of each DNA sequence in the dataset is 140 nucleotides. The DNA sequences follow the GT-AG rules: GT are located at sites 71 to 72, and AG are located at sites 69 and 70. The dataset is divided into four parts: true exon to intron, false exon to intron, true intron to exon, and false intron to exon.

In the experiment, 12,600 DNA fragments that met the GT-AG rules were selected from the HS^3D database, including 3,200 splicing acceptor sites for training, 3,200 splicing acceptor sites for testing, 3,100 splicing donor sites for training, and 3,100 splicing donor sites for testing. The experimental sample data parameters are shown in Table 1. The samples were coded using One-Hot Encoding in order of 140 nucleotides.

4.2. Analysis of Experimental Results. Experimental parameters were set as population size was 40, acceleration coefficient was 2.0, and maximum number of iterations was 200. The optimal parameters (i.e., penalty factor c_1, penalty factor c_2, and kernel parameter of GAUSSIAN radial basis function σ) were calculated by PSO algorithm, and then these optimal parameters and training samples of gene splicing sites were input for testing. The proposed PSO-TWSVM algorithm was compared with traditional SVM algorithm, TWSVM algorithm, and Least Squares Support Vector Machine (LSSVM) algorithm.

In the experiment, positive sample recognition rate, negative sample recognition rate, and correlation coefficient (CC) were used to evaluate the recognition effects of these four splicing sites recognition models. The calculation formula of CC is shown as follows [33].

$$CC = \frac{TF - \bar{T}\bar{F}}{\sqrt{(T+\bar{T})(T+\bar{F})(F+\bar{F})(F+\bar{T})}}, \quad (13)$$

where T is the positive samples that are correctly recognized, F is negative samples that are correctly recognized, $\bar{T}$ is the positive samples that are wrongly recognized, and $\bar{F}$ is the negative samples that are wrongly recognized.

The recognition results of different support vector machines on acceptor sites are shown in Table 2. The recognition results of different support vector machines on donor sites are shown in Table 3.

It can be seen from Tables 2 and 3 that the CC values of the four different support vector machines for gene splicing site recognition reached more than 76%, which verified the excellent performance of support vector machines in the field of sequence recognition. The positive sample recognition rates, negative sample recognition rates, and the CC values of the proposed PSO-TWSVM algorithm and

TWSVM algorithm are higher than those of SVM algorithm and the LSSVM algorithm. This shows that the TWSVM has better recognition performance than SVM and LSSVM when solving nonlinear problems. A twin support vector machine is more suitable for gene splicing site recognition than the other two support vector machines. Moreover, the PSO algorithm is conducive to the more accurate parameter selection of a twin support vector machine, instead of blindly looking for parameters. PSO-TWSVM algorithm finds the optimal parameters of a twin support vector machine, which makes it achieve the optimal recognition ability. Among the four different SVM algorithms, the proposed PSO-TWSVM algorithm obtained the highest positive sample recognition rates, the highest negative sample recognition rates, and the best *CC* values on both acceptor sites and donor sites (as shown in Figure 4). To sum up, the PSO-TWSVM algorithm has a good recognition performance for gene splicing sites.

5. Conclusions

A Particle Swarm Optimization Twin Support Vector Machine (PSO-TWSVM) algorithm was proposed to recognize gene splicing sites, which is conducive to the development of smart healthcare. The Particle Swarm Optimization (PSO) algorithm optimizes the parameter of twin support vector machine, making the selection of parameters more accurate instead of blindly looking for parameters. Experimental results show that compared with SVM algorithm, TWSVM algorithm, and LSSVM algorithm, the PSO-TWSVM algorithm has the highest recognition rates on both acceptor sites and donor sites for splicing site recognition. However, although PSO-TWSVM algorithm shows excellent performance on gene splicing site recognition, it is sensitive to noise. Next, further research will be carried out on how to reduce noise interference. On the basis of the research of this method, we will combine the proposed algorithm with deep learning algorithms to further improve the recognition results of gene splicing sites.

Abbreviations

SVM:	Support vector machine
TWSVM:	Twin support vector machine
LSSVM:	Least squares support vector machine
PSO:	Particle swarm optimization
PSO-TWSVM:	Particle swarm optimization twin support vector machine
HGP:	Human genome project
CC:	Correlation coefficient.

Data Availability

The data can be available upon request to the corresponding author.

Disclosure

An earlier version of this work has been presented as a preprint according to the following link https://www.researchsquare.com/article/rs-417904/v1 [34], which is not published. The earlier version was revised to form the current version. This work is subject to the latest version, copyrighted by the Wireless Communications and Mobile Computing.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

This work is partially supported by the Digital Media Art, Key Laboratory of Sichuan Province, Sichuan Conservatory of Music, Project No.: 21DMAKL01; supported by the First Batch of Industry-University Cooperation Collaborative Education Project funded by the Ministry of Education of the People's Republic of China, 2021, Project No.: 202101071001; supported by the Reproductive Health (Maternal and Infant Fertility/Development Health) Innovation Platform funded by the Science and Technology Projects of Fuzhou Health Commission, Project No.: 2019-s-wp; and supported by the Clinical Study of Gangliosides for Early Warning of Offspring Cognitive Development in Thyroid Disease during Pregnancy funded by the Fujian Natural Science Foundation of China, Project No.: 2022J01521.

References

[1] C. Wang, *Research on the Protection Mechanism of Citizen Health Data under the Background of Intelligent Medical, M.S. Thesis*, Jilin University, Jilin, China, 2021.

[2] F. Tang, D. Li, W. Liu et al., "Evolutionary tendency of clearhead icefish Protosalanx hyalocranius inferring mitochondrial DNA variation analyses in Amur (Heilongjiang) river catchment, China," *International Journal of Agriculture and Biology*, vol. 20, no. 10, pp. 2329–2334, 2018.

[3] J. Wang, L. Guo, J. Wu, L. Tang, and D. Hu, "Status of bioinformatics research in big data," *Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition)*, vol. 37, pp. 62–67, 2017.

[4] J. Zhao, B. Wei, M. Chen, and X. Zhang, "A splice site prediction algorithm based on SNP and neural network," *Computer Enginering & Science*, vol. 38, pp. 885–890, 2016.

[5] L. Rui, *Using Deep Learning to Identify Gene Splicing Sites of Crops, M.S. Thesis*, School of Information Science and Engineering, Shandong Agricultural Univ, Shandong, China, 2019.

[6] L. A. Skurikhina, A. G. Oleinik, A. D. Kukhlevsky, N. E. Kovpak, S. V. Frolov, and D. S. Sendek, "Phylogeography and demographic history of the Pacific smelt Osmerus dentex inferred from mitochondrial DNA variation," *Polar Biology*, vol. 41, no. 5, pp. 877–896, 2018.

[7] S. D. Sharma, S. N. Sharma, and R. Saxena, "Model independent method for acceptor splice sites prediction in DNA sequences," in *2019 IEEE Conference on Information and Communication Technology*, Allahabad, India, 2019.

[8] V. Morfino, S. Rampone, and E. Weitschek, "A comparison of Apache Spark supervised machine learning algorithms for DNA splicing site prediction," *Smart Innovation, Systems and Technologies*, vol. 151, pp. 133–143, 2020.

[9] L. Wang, D. Zheng, and C. Zheng, "Prediction on slice site based on improved SVM model," *Journal of Yibin University*, vol. 14, pp. 93–98, 2014.

[10] R. Wang, Z. Wang, J. Wang, and S. Li, "SpliceFinder: ab initio prediction of splice sites using convolutional neural network," *BMC Bioinformatics*, vol. 20, no. S23, p. 652, 2019.

[11] E. Pashaei and N. Aydin, "Markovian encoding models in human splice site recognition using SVM," *Computational Biology and Chemistry*, vol. 73, pp. 159–170, 2018.

[12] P. R. Javier, D. G. Aida, and G. P. Nicolas, "Floating search methodology for combining classification models for site recognition in DNA sequences," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 6, pp. 2471–2482, 2021.

[13] U. Waseem, M. Khan, U. Ijaz, U. Amin, U. Khattak, and S. M. Saeed, "Splicing sites psrediction of human genome using machine learning techniques," *Multimedia Tools and Applications*, vol. 80, pp. 30439–30460, 2021.

[14] N. Scalzitti, A. Kress, R. Orhand et al., "Spliceator: multi-species splice site prediction using convolutional neural networks," *BMC Bioinformatics*, vol. 22, no. 1, 2021.

[15] J. Zhang, B. Liu, J. Wang, L. Klaus, and G. Mark, "DeepPN: a deep parallel neural network based on convolutional neural network and graph convolutional network for predicting RNA-protein binding sites," *BMC Bioinformatics*, vol. 23, no. 1, p. 257, 2022.

[16] S. Ghazanfari, A. Rasteh, S. A. Motahari, and M. S. Baghshah, "Isoform function prediction using deep neural network," *arXiv preprint*, no. article 2208.03325, 2022.

[17] Z. Shen, Y. L. Shao, W. Liu, Q. Zhang, and L. Yuan, "Prediction of back-splicing sites for circRNA formation based on convolutional neural networks," *BMC Genomics*, vol. 23, no. 1, pp. 581–581, 2022.

[18] L. Cao, C. Qin, and C. Hong, "Model selection of twin support vector machines based on genetic algorithm," *Modern electronic technology*, vol. 17, pp. 113–116, 2017.

[19] S. Ding, X. Zhao, J. Zhang, X. Zhang, and Y. Xue, "A review on multi-class TWSVM," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 775–801, 2019.

[20] A. R. Mello, M. R. Stemmer, and A. L. Koerich, "Incremental and decremental fuzzy bounded twin support vector machine," *Information Sciences*, vol. 526, pp. 20–38, 2020.

[21] L. Liu, M. Chu, Y. Yang, and R. Gong, "Twin support vector machine based on adjustable large margin distribution for pattern classification," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 10, pp. 2371–2389, 2020.

[22] J. Liao, F. Bai, and L. Ma, "V-twin bounded support vector machines based on L1-norm," *Journal of Chongqing Normal University (Natural Science)*, vol. 37, pp. 1–11, 2020.

[23] C. Wang, H. Chen, and Y. Wang, "Analysis and optimization of management ability of small and medium-sized enterprises based on twin support vector machine," *Journal of Beijing Institute of Graphic Communication*, vol. 28, pp. 100–113, 2020.

[24] J. 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. 301–328, 2017.

[25] Y. Gao, *Twin Support Vector Machine Based on Artificial Fish Swarm Algorithm and Its Application, M.S. Thesis*, Jiangnan University, Jiangsu, China, 2020.

[26] H. Zhang, Z. Zhong, and H. Huang, "Optimization on hot rolling load distribution by particle swarm optimization algorithm with constriction factor," *Forging & stamping technology*, vol. 45, pp. 194–199, 2020.

[27] X. He, "Research on an evaluation model of the impact of government WeChat based on PSO-BP neural network," *Scientific Information Research*, vol. 4, no. 3, pp. 60–72, 2022.

[28] W. Liao, S. Cheng, D. Shang, and Z. Wei, "Particle swarm optimization algorithm integrated with multiple-strategies," *Computer Engineering and Applications*, vol. 1, no. 57, pp. 69–76, 2021.

[29] L. Cao, C. Qin, and C. Hong, "Genetic algorithm based model selection of TWSVM," *Modern Electronics Technique*, vol. 40, pp. 105–108, 2017.

[30] X. Wang, J. Wen, F. Luo, W. Zhou, and H. Ren, *Research on Personalized Recommendation Based on Optimized Support Vector Machine*, College of Computer Science, Chongqing Univ, Chongqing, China, 2015, Ph.D. dissertation,.

[31] Y. Huang, Q. Xu, N. Xie, and Q. Tan, "Electricfield optimization in longitudinal modulated optical voltage sensor on SVM-PSO algorithm," *Journal of Fuzhou University (Natural Science Edition)*, vol. 48, pp. 464–470, 2020.

[32] P. Pollastro and S. Rampone, "HS3D, a dataset of homo sapiens splice regions, and its extraction procedure from a major public database," *International Journal of Modern Physics C*, vol. 13, pp. 1105–1117, 2002.

[33] X. Zhou, "Splice sites identification based on multi-scale component features and adjacent positions relationship features," *Computer Engineering and Applications*, vol. 50, no. 10, pp. 120–123, 2014.

[34] F. Zhang, "An advanced twin support vector machine algorithm for gene splicing sites prediction," *preprint*https://www.researchsquare.com/article/rs-417904/v1.