Research in Astronomy and Astrophysics, 23:065004 (9pp), 2023 June https://doi.org/10.1088/1674-4527/accbb0

Flare Forecast Model Based on DS-SMOTE and SVM with Optimized Regular Term

Jie Wan[1,2], Jun-Feng Fu[1,2], Ren-Qing Wen[1], Ke Han[3], Meng-Yao Yu[2,3], and Peng E[1,2]
[1] School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin 150001, China; epeng@hit.edu.cn
[2] Laboratory for Space Environment and Physical Sciences, Harbin Institute of Technology, Harbin 150001, China
[3] School of Computer and Information Engineering, Harbin University of Commerce, Harbin 150028, China
Received 2022 November 6; revised 2023 March 19; accepted 2023 March 20; published 2023 May 10

Abstract

The research of flare forecast based on the machine learning algorithm is an important content of space science. In order to improve the reliability of the data-driven model and weaken the impact of imbalanced data set on its forecast performance, we proposes a resampling method suitable for flare forecasting and a Particle Swarm Optimization (PSO)-based Support Vector Machine (SVM) regular term optimization method. Considering the problem of intra-class imbalance and inter-class imbalance in flare samples, we adopt the density clustering method combined with the Synthetic Minority Over-sampling Technique (SMOTE) oversampling method, and performs the interpolation operation based on Euclidean distance on the basis of analyzing the clustering space in the minority class. At the same time, for the problem that the objective function used for strong classification in SVM cannot adapt to the sample noise, In this research, on the basis of adding regularization parameters, the PSO algorithm is used to optimize the hyperparameters, which can maximize the performance of the classifier. Finally, through a comprehensive comparison test, it is proved that the method designed can be well applied to the flare forecast problem, and the effectiveness of the method is proved.

Key words: Sun: flares – Sun: magnetic fields – Sun: X-rays – gamma-rays – (Sun:) sunspots

1. Introduction

A solar flare is a transient outburst phenomenon that occurs in the solar atmosphere and is highly correlated with sunspot activity in the photospheric layer. Its powerful magnetic reconnection process affects all layers on the solar surface. Solar flares are characterized by a rise time of several minutes and a decay time of tens of minutes, and are observed as a creeping phenomenon of X-rays (especially soft X-rays). The Sun is of profound significance in the study of the relationship between the Sun and the Earth, and its radiation risk is a matter of discussion and major concern for manned missions to Mars, the moon or other planets. It will not only damage the power transmission system, affect the climate and atmospheric environment, but even endanger human survival. Therefore, in the study of the Sun-terrestrial relationship, the study of solar flare forecast is not only an important goal of human space exploration and space activities, but also an important link in understanding the mechanism of solar activity, which has important practical value and scientific significance.

There are currently machine learning-based forecasting models and non-machine learning-based forecasting models. The mainstream solar flare forecasting models are constructed through data-driven machine learning models. In terms of physics-oriented non-machine learning methods, the core of their research is: trying to analyze the relationship between the morphology of the sunspots and the intensity of the flares. At the end of the last century, McIntosh (1990) generalized a taxonomy that can describe the above relationships to some extent. This study is consistent with Miller's forecast system Miller (2005) were both groundbreaking at the time and provided an important foundation for subsequent research Ribeiro & André (2021). It is important to know, however, that these models, which are not machine learning methods, are highly dependent on expert knowledge and experience for their forecast power and are therefore subjective in their forecast results. In this century, Kusano et al. (2020) developed a κ-scheme approach, which is its underlying logic forecast large solar flares by the key condition of magnetohydrodynamic instability triggered by magnetic reconnection. However, since magnetohydrodynamics is slightly incomplete in describing the complex process of solar flares, in general, the current performance of the method still leaves much room for improvement. Stanislavsky presents a full solar flare forecast method based on a Hidden Markov model with two hidden states. Soft X-ray data are used and independent identically distributed (IID) random variables and autoregressive (AR) processes are considered. Finally, the AR method is shown to be significantly better than the IID. In addition, it can well detect higher states, which can lead to very strong energy release. Significant development of the model is necessary in

order to forecast solar flares throughout the solar cycle (Stanislavsky et al. 2020). In the field of machine learning, data-driven methods have been utilized to forecast solar flares. Early approaches, such as Fozzard et al. (1988) connectionist expert system, employed a simple neural network model for solar flare forecasting. Although their accuracy and speed were not ideal, these early attempts paved the way for incorporating machine learning techniques in solar research. In fact, one important branch of machine learning, deep learning, has been used to develop solar flare forecast models (Huang et al. 2018). Ali et al. (2021) used a deep learning approach to analyze and forecast sunspot activity, and developed an automated hybrid computer system that utilizes a convolutional neural network (CNN) training scheme based on the integration of a back-propagation algorithm and a mini-batch AdaGrad optimization method for near real-time solar data processing and solar flare forecast.

The Support Vector Machine (SVM) algorithm is the most typical method due to its elegant mathematical structure and strong performance, attracting attention from scholars at home and abroad since its inception. For example, Qahwaji & Colak (2007) established a hybrid flare short-term forecasting system by applying neural network and SVM algorithms. However, the unoptimized SVM has low adaptability to the problem of sample noise and overfitting. The commonly used method is to optimize the regular term of the SVM (Karaboga & Basturk 2007; Ruan et al. 2019; Fauzi et al. 2021), which essentially uses fuzzy theory to alleviate the overfitting and noise adaptability problems. In fact, flares are rare events in the entire solar cycle and they exhibit class imbalance in the sample, which will affect data-driven forecasting models Jie et al. (2021). In addition, the combination of different forecasting algorithms and data preprocessing methods also has different effects (Jie et al. 2022).

There is no detailed demonstration of the SVM flare forecast model optimized using the regular term parameter. Therefore, it is valuable that we research in the flare forecast model based on SVM with the optimized regular term. In addition, we also use Synthetic Minority Over-sampling Technique (SMOTE) method based on density (DS-SMOTE) to resample the data, which is unusual in flare forecast community.

The sections of this article are primarily divided as follows: Section 1—Introduction will introduce research on solar flare forecast in recent years and attempt to analyze the "machine learning" and "non-machine learning" methods, pointing out areas that require additional attention in existing research. In Section 2—Data, we will introduce the data source and analyze issues of "intra-class imbalance" and "inter-class imbalance." In Section 3—Method, we will present a "Flare forecast model based on DS-SMOTE and SVM with optimized regular term" and gradually introduce this method. In Section 4—Result Analysis, we will compare the performance of various methods to demonstrate the effectiveness of the method proposed in this article. Finally, in Section 5—Conclusion, we will summarize the conclusions drawn from the testing and research conducted in this article.

2. The Problem of Class Imbalance in Data

2.1. Data Source

In this paper, the parameters of sunspots and the 10.7 cm solar radio flux are selected as forecast factors, the sunspot parameters include sunspot number, the precise positional information of sunspots on the solar surface, magnetic classification, and McIntosh classification. To obtain solar flare data, one can visit the National Geophysical Data Center (NGDC) website at (https://www.ngdc.noaa.gov/). Within the website, one can select Flare Index data from 1976 to 2021 at (https://www.ngdc.noaa.gov/stp/space-weather/solar-data/solar-features/solar-flares/index/flare-index/). Additionally, X-ray data from 1975 to 2017 can be obtained at (https://www.ngdc.noaa.gov/stp/space-weather/solar-data/solar-features/solar-flares/x-rays/goes/xrs/). Information on active region numbers and locations related to solar flares is also needed, it can be found in the Solar Region Summary (SRS) at (https://www.swpc.noaa.gov/products/solar-region-summary).

The data set we used is all the observation data from 2010 to 2017. There are a total of 10, 200 samples in the data set. Each sample consists of 26 attribute columns, of which the sunspot group area is one attribute column. The McIntosh classification of the sunspot group is based on the overall shape, the maximum sunspot morphology and the distribution between leading and trailing sunspots are divided into 17 attribute columns, one attribute column for the number of visible sunspots, four attribute columns for the magnetic classification of sunspot groups, and one attribute column for the 10.7 cm radio flux.

2.2. Inter-class and Intra-class Imbalance Problems

The solar flare data set is a typical class-imbalanced data set. When traditional machine learning methods are used for the classification of imbalanced data sets, there is often a large bias. In the problem of flare forecast, positive samples are more concerned. The Imbalance Rate (IR) of the data is defined as the ratio of the majority class samples to the minority class samples. The definition of imbalance ratio (IR) is as follows:

$$\mathrm{IR} = \frac{\max(\mathrm{nPos}, \mathrm{nNeg})}{\min(\mathrm{nPos}, \mathrm{nNeg})}. \quad (1)$$

The IR value of the data set is 16.51 (671 Positive samples and 11,079 Negative samples), which shows a high degree of imbalance.

In recent years, the problem of class imbalance has received more and more attention from scholars, and classification strategies to solve the problem of class imbalance have

emerged in an endless stream, which can be summarized into two categories: one is to reconstruct the data set from the data itself, in order to change the number and distribution of samples; the other category proposes a series of targeted improvement strategies from the aspects of classification algorithms and classification ideas, which tend to pay more attention to minority classes. As a data-level processing method, the resampling method is simple and can effectively retain sample information, which is favored by researchers. The SMOTE algorithm proposed by Alberto et al. (2018) analyzes the minority class and selects each minority class point and its k nearest neighbors to perform linear interpolation to synthesize new data.

But the SMOTE algorithm also has some flaws. First, the synthesis of new samples depends on the selection of root samples and auxiliary samples. If one of the two is a noise sample, the quality of the new samples obviously has problems; there are also dense areas in the area. After SMOTE sampling, the sparse areas are still sparse, and the dense areas are still dense. As shown in above figure, there is not only the imbalance between positive and negative samples in the flare data, but also the phenomenon of multiple clusters in a single category, which is manifested as intra-class imbalance. But in fact, samples from different distribution regions are of different importance to the classifier. In view of the shortcomings of the above SMOTE sampling method, we adopts DS-SMOTE to better describe the distribution of unbalanced data sets.

3. Flare Forecast Model Based on DS-SMOTE and SVM with Optimized Regular Term

3.1. Application of SVM Algorithm in Overfitting Problem

The SVM algorithm is to decompose a large-scale original problem into multiple small-scale sub-problems to solve, because most of the small problems are easy to solve, and the results of sequential solutions are exactly the same as the results of large-scale solutions. But the solution time of the Sequential Minimal Optimization (SMO) algorithm is much shorter than that of the other algorithms. We might as well focus on solving the optimization problem of SVM:

Given a set of training data in a feature space:

$$T = (x_1, y_1),\cdots,(x_l, y_l). \in (R^n \times \gamma)^l, \tag{2}$$

where $x_i \in R^n$, $y_i \in \gamma = (-1, 1)$, $i = 1, 2, \cdots, l$, x_i is the ith feature vector. We can call (x_i, y_i) is the ith sample point.

When the samples are linearly separable, the optimal classification hyperplane problem is to maximize the distance between two types of samples. Suppose the classification hyperplane of the above samples is:

$$\omega \cdot x + b = 0, \tag{3}$$

where ω and b are the normal vector and the intercept, respectively, and the interval plane can be normalized by adjusting the two coefficients in equal proportions. In this way, all samples of the two classes satisfy $|\omega \cdot x + b| \geqslant 1$, so that the classification interval becomes $\frac{2}{||\omega||}$. Introducing slack variable ς_i and penalty factor C to allow partial misclassification of samples, the objective function is equivalent to the following convex optimization problem:

$$\min_{\omega,b} \frac{1}{2}||\omega^2|| + C\sum_{i=1}^{l} \varsigma_i \tag{4}$$

$$s.t: y_i(\omega \cdot x_i + b) \geqslant 1 - \varsigma_i, i = 1,\cdots,l \tag{5}$$

$$\varsigma_i \geqslant 0. \tag{6}$$

Obviously, turning this into a dual problem is:

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{l} \alpha_i \tag{7}$$

$$s.t: \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{8}$$

$$C \geqslant \alpha_i \geqslant 0, i = 1,\cdots,l, \tag{9}$$

where α_i is the ith sample's "Lagrange–Operator–Coefficient." By solving the above dual problem, the numerical solutions of w and b can be obtained, and the classifier can be obtained. However, linear division is not suitable for most problems in the real world, so we change Equations (7)–(9) into the following form:

$$\max(\alpha_1 + \alpha_2) - \frac{1}{2}\sum_{i=1}^{2} K_{ii}\alpha_i - y_1\alpha_1^2\sum_{i=3}^{l} y_i\alpha_i K_{i1} - y_2\alpha_2\sum_{i=3}^{l} y_i\alpha_i K_{i2} \tag{10}$$

$$y_1\alpha_1 + y_2\alpha_2 = -\sum_{i=3}^{l} y_i\alpha = \varsigma \tag{11}$$

$$0 \leqslant \alpha_i \leqslant C, i = 1, 2, \tag{12}$$

where K is the kernel function acting in Equation (3). Here we choose linear kernel function: $K(x, x_j) = x^t \cdot x_i$

For the classification problem of support vector machine, y_1 and y_2, takes the value of -1 or 1, under the constraint of Equation (11), the relationship of α_1 and α_2 can be expressed as a line segment limited in the [0, C] rectangle, as shown in Figure 1. Therefore, the bivariate optimization problem of the SMO algorithm can be transformed into a univariate optimization problem. Assuming that the solution obtained in the previous iteration are α_1^{old} and α_2^{old}, the solution obtained in new round are α_1^{new} and α_2^{new}, and the extension constraint direction is not pruned, the solution is $\alpha_2^{\text{new,unc}}$. Let $L = \max(0, \alpha_2^{\text{old}} - \alpha_1^{\text{old}})$ and $H = \min(C, C + \alpha_2^{\text{old}} - \alpha_1^{\text{old}})$, then we have $L \leqslant \alpha_2^{\text{new}} \leqslant H$, if $\alpha_2^{\text{new,unc}}$ is outside the

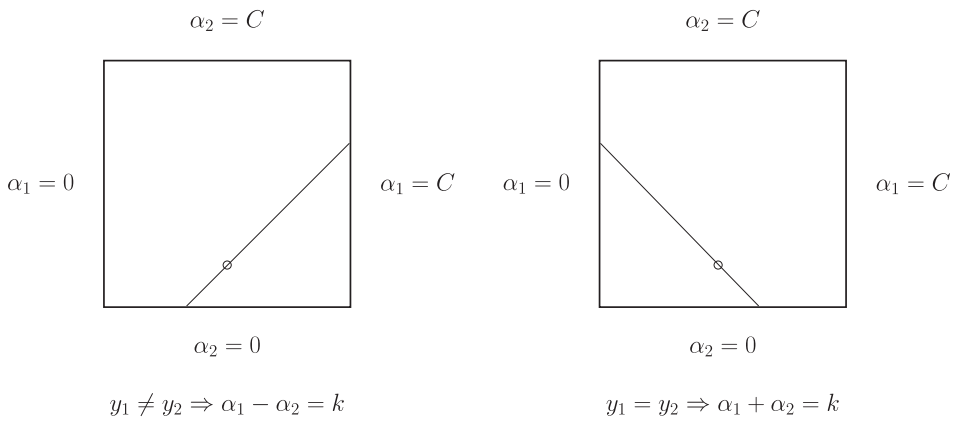

Figure 1. Square constraint map.

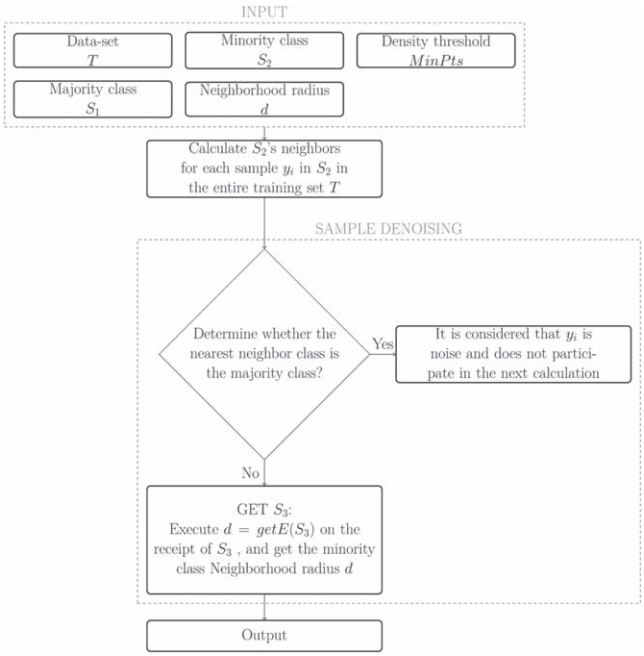

Figure 2. Samples denoising.

boundary, trim it to the adjacent boundary value, otherwise, $\alpha_2^{\text{new}} = \alpha_2^{\text{new,unc}}$.

Next, we need to solve$\alpha_2^{\text{new,unc}}$. Using the optimization constraints of Equation (11) to simplify the objective function to a formula with only α_2, and get:

$$W(\alpha_2) = \frac{1}{2}K_{11}(\zeta - y_2\alpha_2)^2 + \frac{1}{2}K_{22}\alpha_2^2 + (\zeta - y_2\alpha_2)v_1 + y_2\alpha_2 v_2 - (\alpha_1 + \alpha_2), \quad (13)$$

where $v_i = \sum_l^{i=3} y_i\alpha_i K(x, x_i)$. The partial derivative of the objective function containing only α_2 with respect to α_2 can be obtained:

$$\frac{\partial W}{\partial \alpha_2} = K_{11}\alpha_2 + K_{22}\alpha_2 - 2K_{12}\alpha_2 - K_{11}y_2\zeta + K_{12}y_2\zeta + y_2y_1 - y_2v_1 + y_2v_2 - 1 = 0 \quad (14)$$

the solution is :

$$\alpha_2^{\text{new,unc}} = \alpha_2^{\text{old}} + \frac{y_2(E_1 - E_2)}{K_{11} + K_{22} - 2K12}, \quad (15)$$

where $E_i = \sum_{j=1}^{l} \alpha_j y_j K(x_i, x_j) + b - y_i = g(x_i) - y_i$

α_2^{new} can be obtained by pruning $\alpha_2^{nwe,unc}$, and then α_1^{new} can be solved according to the linear relationship between α_1^{new} and α_2^{new}.

The selection of the first optimization variable is called the outer loop of the SMO algorithm, that is, it traverses all samples, selects α_i that violates the KKT condition as the first variable, and then selects the second sample according to the principle of making $|E_i - E_j|$ the largest. If no α_i that violates the KKT condition is found by traversing the entire sample set, exit the iteration.

KKT conditions refer to:

$$\begin{aligned} \alpha_i = 0 &\Rightarrow y_iE_i \geqslant 0 \\ \alpha_i = C &\Rightarrow y_iE_i \leqslant 0 \\ 0 < \alpha_i < C &\Rightarrow y_iE_i = 0. \end{aligned} \quad (16)$$

The original SMO algorithm is widely used in the construction of SVM models because of its detailed theory and simple implementation. However, the SVM model built with the original SMO algorithm has the shortcomings of poor stability and low forecast performance. So we will use an optimized SMO algorithm to solve the above problems.

3.2. DS-SMOTE Resampling Algorithm

Specific improvement of SMOTE for different problems Wang (2017), so that it has produced outstanding performance

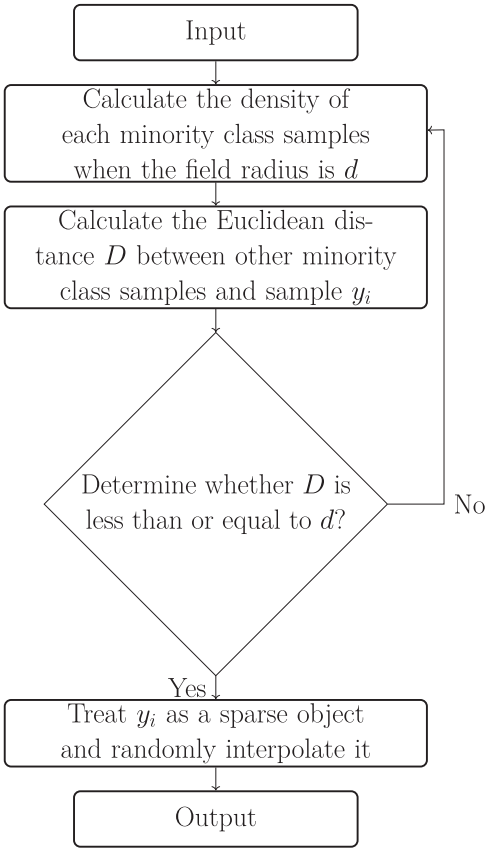

Figure 3. Determining sparse sets.

in rapidity and accuracy. Among them, Chinese scholars have proposed the SMOTE method based on density (DS-SMOTE) method. The core idea of the DS-SMOTE algorithm is: on the basis of the SMOTE algorithm, select sparse sets in the minority class and oversample them. The main steps of generating a sparse set are: removing the noise samples in the minority class samples; calculating the neighborhood radius of the minority class samples; calculating the minority class density threshold; obtaining the sparse set. The new samples synthesized by the DS-SMOTE algorithm are distributed among the sparse objects and their neighboring samples, and finally the minority class samples are roughly equal to the majority class samples.

The density D_i of minority class sample x_i is defined as the number of samples whose distance from sample x_i is less than or equal to the average distance ε between samples in the minority class. Record the density threshold DT as the mean of the density D_i of all minority class samples x_i. Use DT to judge whether the minority class sample is a sparse point, that is, if $D_i > DT$, the sample is a "dense point," otherwise it is a "sparse point." The "two-step" process of the DS-SMOTE algorithm is shown in Figures 2 and 3:

DS-SMOTE is an optimization algorithm of the SMOTE algorithm. By selecting a sparse set, it selectively discriminates minority samples, thereby obtaining a new sample set with uniform distribution. The main steps of generating a sparse set are: removing the noise samples in the minority class samples; calculating the average distance between samples within the minority class samples; calculating the density threshold; and obtaining the sparse set. The new samples synthesized by the DS-SMOTE algorithm are distributed between the samples in the sparse area and their neighbors. By setting the sampling ratio, the minority class samples that are roughly equal to the majority class samples are finally obtained.

3.3. SVM Regular Term Optimization Method

The penalty factor C is used to control the complexity and regression accuracy of the model. The kernel function we used is the Radial Basis Function (RBF), which is:

$$k(x_i, x_j) = \exp(-\gamma \cdot |x_i - x_j|^2), \gamma > 0. \quad (17)$$

In formula 17, γ is the width parameter of the radial basis function, which controls the radial range of the function, $|x_i - x_j|$ represents the modulus of the difference vector. After the kernel function is determined, it is necessary to find the two optimal parameters, namely the penalty factor C and the kernel parameter γ.

Both the penalty factor and the kernel parameter affect the performance of the model, in order to quickly and reasonably determine C and γ of SVM, we introduces the PSO Abdulhamit (2013) algorithm to optimize these two parameters, get a Particle Swarm Optimization-Support Vector Machine (PSO-SVM) algorithm. Particle Swarm Optimization (PSO) algorithm by simulating the foraging behavior of flock of birds, by simulating the foraging behavior of flocks, iterative search is continuously performed to obtain information sharing among individuals and finally obtain the global optimal solution. The PSO algorithm updates the particle's position and velocity by the following formulas:

$$v_{id}^{k+1} = wv_{id}^{k} + C_1\gamma_1(p_{id}^{k} - x_{id}^{k}) + C_2\gamma_2(p_{gd}^{k} - x_{id}^{k}) \quad (18)$$

$$x_{id}^{k+1} = x_{id}^{k} + v_{id}^{k}, \quad (19)$$

where x_{id}^{k} and v_{id}^{k} are respectively the position and velocity of the i particle in the d−dimensional space in the k iteration; p_{id}^{k} and p_{gd}^{k} are respectively the real−time optimal position of the i particle and the optimal position searched by the entire particle swarm so far; w is the inertia weight factor, and the value of ω will affect the ability of optimization; C_1 and C_2 are learning factors, γ_1 and γ_2 are independent random numbers between [0, 1].

In this paper, we first use the PSO algorithm to find the optimal hyperparameter, and then SVM is used to train the data class, the specific operations are as Figure 4:

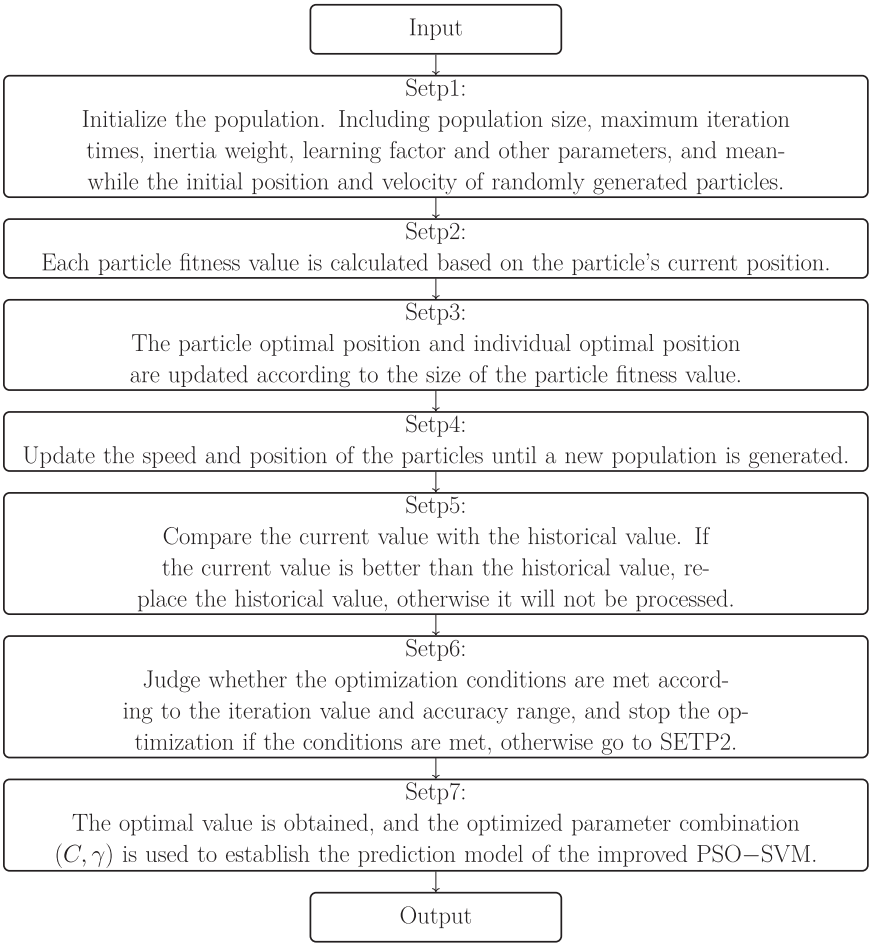

Figure 4. Particle Swarm Optimization (PSO) method process.

Table 1
Confusion Matrix

	Forecast for Flares	Forecast for No Flares
Actual flare	TP	FN
Actually no flare	FP	TN

Table 2
Schematic Representation of Datasets

Methods of Resampling	Total Samples	Positive/Negative
Original	10,200	600/9600
SMOTE	192,000	9600/9600
DS-SMOTE	19,176	9576/9600

4. Test Scheme Design and Forecast Result Analysis

4.1. Performance Index

We investigates data-based forecasting strategies for the solar flare event. The research includes the selection of a suitable combination of parameters and data extraction based on the physical mechanism of solar flare eruption. Learning algorithms are used for solar flare forecasting methods, and multiple sets of comparative examples are designed. The evaluation index used in this study is the confusion matrix of the two-class problem, which is a commonly used index and its derivative index to measure the performance of the model (as shown in Table 1).

We choice various indices are typically derived from the confusion matrix to evaluate model performance. The key indices and their formulas are:

Accuracy: Measures the overall correctness of the model's forecasts. Even if positive samples (flare outbreaks) are misclassified, the accuracy rate can still be high when negative samples dominate.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \tag{20}$$

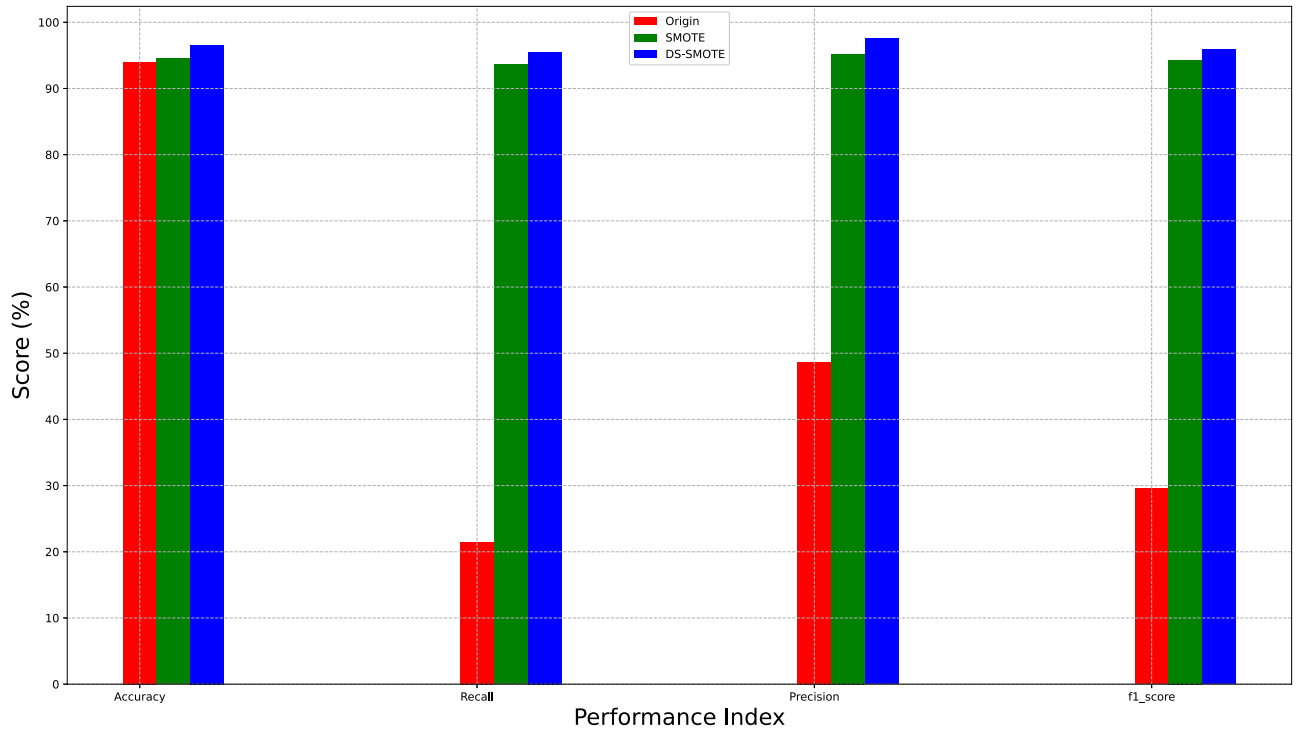

Figure 5. Class imbalance processing results comparison.

Table 3
Hyperparameter Optimization Results

Parameters	Original	Hybrid	SMOTE
C	20.2	76.13	65.2
γ	16.97	107.5	23.7

Recall: Measures how well the model detects positive or negative samples. For a flare forecasting model, a high true positive (TP) rate is desirable to maximize detection of all flare events and prevent space weather disasters.

$$\begin{cases} \mathrm{TP}_{\mathrm{rate}} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ \mathrm{TN}_{\mathrm{rate}} = \dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \end{cases} \tag{21}$$

Precision: Measures the confidence level of the model's positive or negative forecasts. In practice, precision is often a trade-off with recall.

$$\begin{cases} P^{+} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \\ P^{-} = \dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \end{cases} \tag{22}$$

F_1 score: A comprehensive index that combines recall and precision using their harmonic mean. A high F_1 score is necessary to ensure high accuracy rate when a high recall rate is desired for forecasting.

$$\begin{cases} F_1^{+} = \dfrac{2\mathrm{TP}_{\mathrm{rate}}P^{+}}{\mathrm{TP}_{\mathrm{rate}}+P^{+}} \\ F_1^{-} = \dfrac{2\mathrm{TN}_{\mathrm{rate}}P^{-}}{\mathrm{TN}_{\mathrm{rate}}+P^{-}} \end{cases} \tag{23}$$

$$\mathrm{HSS} = \frac{2(\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN})}{(\mathrm{TP}+\mathrm{FN})\times(\mathrm{FN}+\mathrm{TN})+(\mathrm{TP}+\mathrm{FP})\times(\mathrm{FP}+\mathrm{TN})} \tag{24}$$

4.2. Test Scheme Design

The data set has a total of 10,200 samples, of which the number of positive samples is 600 and the number of negative samples is 9600. On the original data set, the SMOTE oversampling algorithm and the DS−SMOTE oversampling algorithm are used to obtain a balanced data set. The data set is shown in the following Table 2:

For the verification of the experimental data sets, the ten-fold cross-validation method is adopted, that is, the data is randomly divided into ten parts, 9 of them are used as the training set in turn, and the remaining 1 is used as the test set. The tests are carried out 10 times in turn to ensure that the whole data set is tested. The forecast results are represented by a confusion matrix. In the binary classification problem, the confusion matrix is shown in the following table.

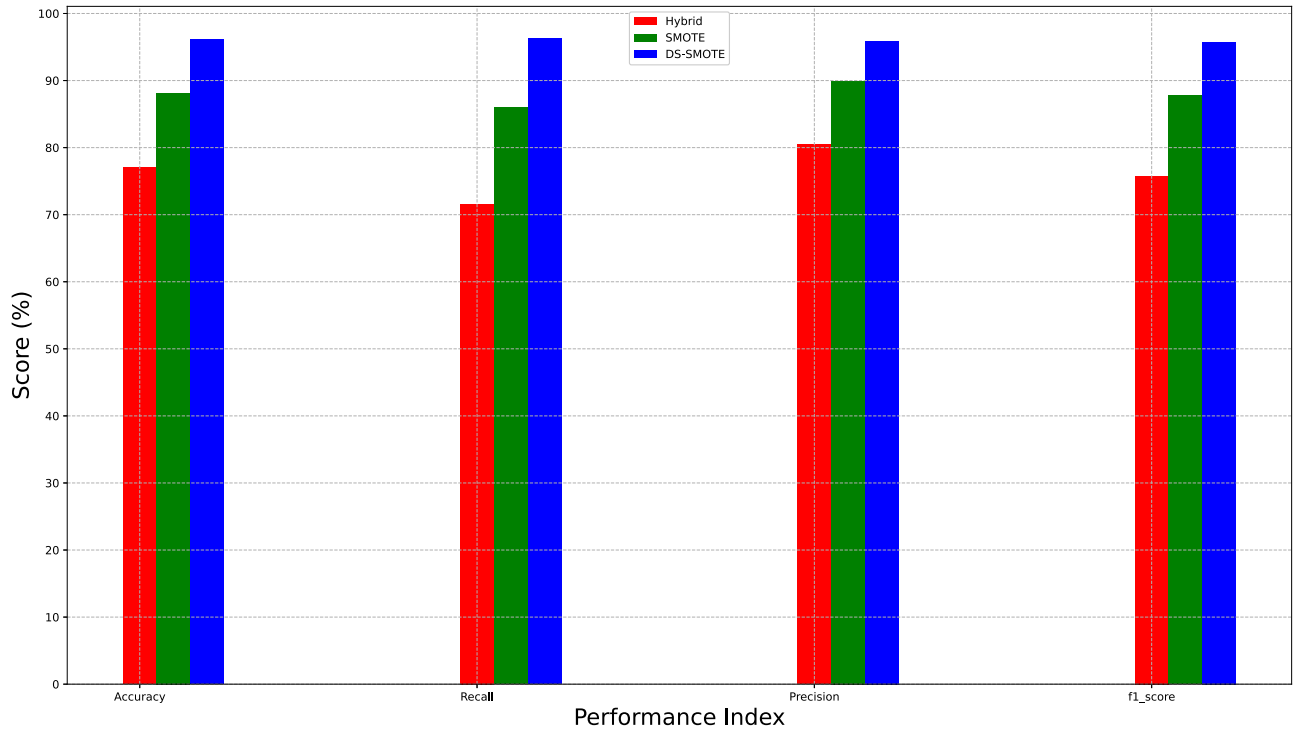

Figure 6. Class imbalance processing results comparison (with hyperparameters).

4.3. Forecast Result Analysis

The columns of the confusion matrix represent the predicted results of the class, and the rows represent the actual class of the class. Among them, True Negative (TN) represents the number of correctly divided samples in the negative class, that is, the true negative class; True Positive (TP) represents the number of correctly divided samples in the positive class, that is, the true positive class; False Negative (FN) represents the number of samples in the positive class that are wrongly classified, that is, the false negative class; False Positive (FP) represents the number of wrongly classified samples in the negative class, that is, the true and false positive classes.

In addition, in order to further describe the performance of the model, we uses the following indicators to further describe: Accuracy rate, Recall rate, Precision rate, and f1−score criterion. On this basis, three forecast models are established for the original data set, the SMOTE oversampling data set, and the DS-SMOTE oversampling data set. The comparison results are shown in Figure 5:

As shown in the above figure, the recall of the model trained from the original data set is very low, which indicates that the model tends to identify the samples as negative samples, and what we really care about is the samples when the flare erupted, such a model has no application value. It can be seen from the models B and C that the Recall, Precision, and f1−score of the original data set increased significantly after the balancing process, the FAR decreased significantly, and the model performance was greatly improved. At the same time, when compared with the forecast models in the literature, it is found that the performance of the SVM model has been greatly improved by applying the optimal hyperparameters found from the PSO algorithm, which fully demonstrates the importance of appropriate hyperparameters for machine learning models. Based on the above experimental results, it is fully proved that the SVM solar flare forecast model combined with resampling and particle swarm optimization is an efficient solar flare forecast model. In order to further compare the advantages of DS-SMOTE, we show the forecast results in Table 3 and Figure 6:

5. Conclusion

In order to further improve the performance of data−driven solar flare forecast models, we study the flare forecast model based on DS−SMOTE and SVM with the optimized regular term. Conclusions are as below:

(1) Using the collected flare sample data, we analyze the imbalance distribution characteristics within and between classes, and point out the impact of this class imbalance on forecasting models.

(2) Using the improved DS−SMOTE algorithm to process samples, the results show that: due to the existence of intra −class imbalance problem, making DS−SMOTE better than

the classical SMOTE algorithm. On the other hand, when there is no intra−class imbalance in the data, the advantage of DS −SMOTE will not be obvious.

(3) The comprehensive comparison test results of different machine learning algorithms show that the performance of the SVM series models optimized by regularization parameters is significantly better than other models. Moreover, the hyper-parameters obtained by the PSO algorithm are more reliable than those obtained by other algorithms, which proves the necessity of optimizing the parameters of the regular term.

We provides a further solution to the problem of sample class imbalance in the data−driven modeling of flare forecasting. In the future, we will make more attempts on data−driven modeling methods combined with the physical mechanism of flare eruptions.

Acknowledgments

The authors would like to acknowledge the support of the National Key Research and Development Program of China (No. 2022YFF0503601) and the National Natural Science Foundation of China (No. 11975086). We would like to acknowledge the use of solar flare data provided by the Solar Region Summary (SRS) document, as well as the data access interface provided by the National Geophysical Data Center (NGDC).

ORCID iDs

Jun-feng Fu https://orcid.org/0000-0003-0242-2218

References

Abdulhamit, S. 2013, Comput. Biol. Med., 43, 576
Alberto., F., Salvador., G., Francisco., H., & Nitesh, V. C. 2018, JAIR, 61, 863
Ali, K. A., Rami, Q., & Ahmed, A. 2021, AdSpR, 67, 2544
Fauzi, I. R., Rustam, Z., & Wibowo, A. 2021, JPhCS, 1725, 012012
Fozzard, R., Gary, B., & Louis, C. 1988, Advances in Neural Information Processing Systems, 1, 248
Huang, X., Huaning, W., Long, X., et al. 2018, ApJ, 7, 856.1
Jie, W., Jun-Feng, F., Dai-Min, T., et al. 2022, RAA, 22, 085020
Jie, W., Jun-Feng, F., Jin-Fu, L., & Jia-Kui, S. 2021, RAA, 21, 237
Junhong, W., & Qian, D. 2017, J. Intell. Syst. (in Chinese), 12, 865
Karaboga, D., & Basturk, B. 2007, J Glob Optim, 39, 459
Kusano, K., Tomoya, I., Yumi, B., & Satoshi, I. 2020, Sci, 369, 587
McIntosh, P. S. 1990, SoPh, 125, 251
Miller, Richard W. 2005, in Knowledge-Based Systems in Astronomy, ed. A. Heck & F. Murtagh (Berlin: Springer), 107
Qahwaji, R., & Colak, T. 2007, SoPh, 241, 195
Ribeiro, F., & André, L. S. G. 2021, A&C, 35, 100468
Ruan, J., Jiang, H., Li, X., et al. 2019, IEEE Trans. Ind. Inf., 15, 6510
Stanislavsky, A., Nitka, W., Małek, M., et al. 2020, JASTP, 208, 105407