RAA 2021 Vol. **21** No. **9**, 237(6pp) doi: 10.1088/1674-4527/21/9/237

http://www.raa-journal.org *http://iopscience.iop.org/raa*

***R**esearch in*
***A**stronomy and*
***A**strophysics*

Class imbalance problem in short-term solar flare prediction

Jie Wan[1,2], Jun-Feng Fu[2], Jin-Fu Liu[3], Jia-Kui Shi[2], Cheng-Gang Jin[1,2] and Huai-Peng Zhang[3]

[1] Laboratory for Space Environment and Physical Sciences, Harbin Institute of Technology, Harbin 150001, China; *jinfuliu@hit.edu.cn, fujunfeng1994@163.com*
[2] School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin 150001, China
[3] School of Energy Science and Engineering, Harbin Institute of Technology, Harbin 150001, China

Received 2020 December 16; accepted 2021 May 21

Abstract Using data-driven algorithms to accurately forecast solar flares requires reliable data sets. The solar flare dataset is composed of many non-flaring samples with a small percentage of flaring samples. This is called the class imbalance problem in data mining tasks. The prediction model is sensitive to most classes of the original data set during training. Therefore, the class imbalance problem for building up the flare prediction model from observational data should be systematically discussed. Aiming at the problem of class imbalance, three strategies are proposed corresponding to the data set, loss function, and training process: Type I resamples the training samples, including oversampling for the minority class, undersampling, or mixed sampling for the majority class. Type II usually changes the decision-making boundary, assigning the majority and minority categories of prediction loss to different weights. Type III assigns different weights to the training samples, the majority categories are assigned smaller weights, and the minority categories are assigned larger weights to improve the training process of the prediction model. The main work of this paper compares these imbalance processing methods when building a flare prediction model and tries to find the optimal strategy. Our results show that among these strategies, the performance of oversampling and sample weighting is better than other strategies in most parameters, and the generality of resampling and changing the decision boundary is better.

Key words: The Sun — Sun: X-rays, gamma rays — Sun: sunspots — Sun: magnetic fields — Sun: flares — methods: data analysis

1 INTRODUCTION

Solar flares strongly influence the space weather, especially the large flares. For solar flare prediction, the probability of large flare is small but of great interest. Roy et al. (2020) and Huang et al. (2018) believe that the predictor of solar flares is considered an important parameter in the field of solar research and is used to describe the laws of physics in the transient or steady-state process of the Sun's interior.

As shown in Table 1, many data mining methods can learn predictive models from the generated data sets, and these methods have been used to predict the outbreak of flares. In the flare forecast task, the small probability of "flare events" has caused a strong imbalance problem, that is, the absolute advantage of negative samples (non-flare events) in the data set. The unoptimized training process will aggravate the classification bias of the positive and negative samples in the flare prediction model. For example, the prediction model can simply forecast that no flare will happen with a high success rate. This kind of results is useless, and generally we much more focus on the eruption of flares.

The above-mentioned problem has caused widespread concern, and many international conferences have specifically discussed this problem. For example, AAAI'2000 seminar, ICML'2003 learning seminar and ACM SIGKDD Exploration 2004 learning topic. Three types of methods are used to deal with this problem normally. Type I resamples training samples. Type II changes decision discriminant boundary. Change the weight to smooth out the prediction loss between categories (Tulunay et al. 2004). This can change the discriminant boundary between flaring and non-flaring decision. Type III assigns different weights to training samples (Liu et al. 2008). The majority class is assigned a small weight or the minority class is assigned a large weight. This makes the learned model internally bias the minority class.

Zhou (2019) and Bobra & Couvidat (2015) generated a balanced dataset to train the prediction models. Yi et al.

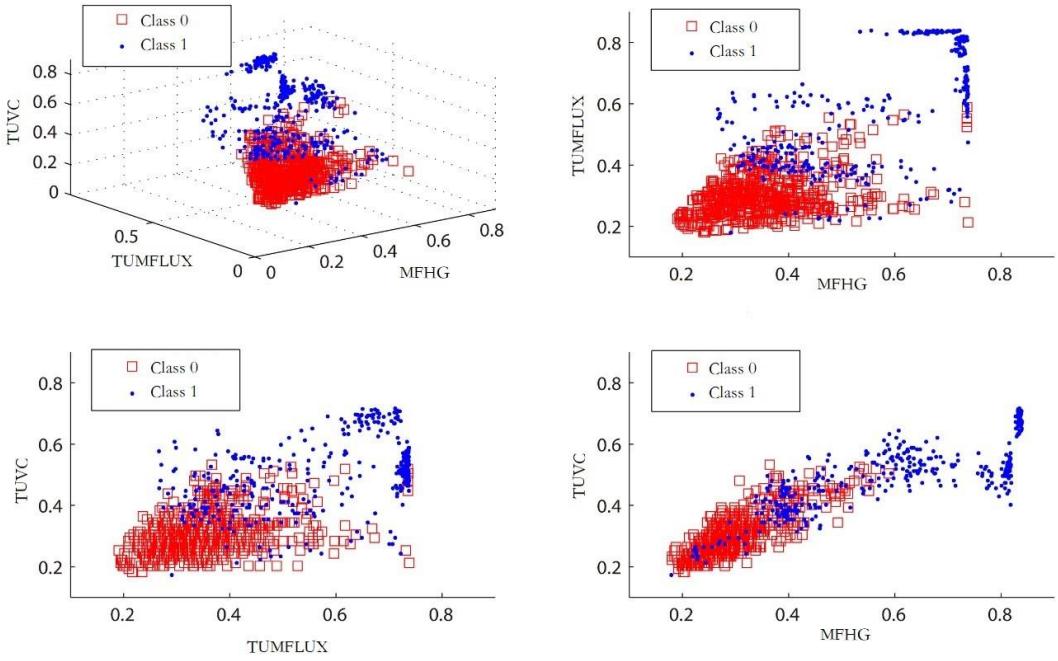

Fig. 1 Class imbalance in predictors.

(2020) used the X-ray flux profile of solar flares without any preprocessing to directly extract feature data. This method can be used to quickly generate forecast models. Park et al. (2020) believed that the core technology of flare prediction separates positive samples from negative samples, which is often referred to as a labeling process in the machine learning community. They tried a series of methods and compared their details and results. In some articles, 1457 unrelated sunspot groups in the NGDC catalog were processed to complete the source data set. Then, the correlation function between flares and sunspot groups was established by using sunspot typing and special designed software. Li et al. (2007) randomly selected a subset of negative samples to match the positive sample set to train the learning vector quantization (LVQ) network, and proposed a fusion of support vector machine (SVM) and K-nearest neighbor (KNN) technology to build a prediction model. Yu et al. (2009), Yu et al. (2010) and Huang et al. (2010) selected the flaring samples and the same number of non-flaring samples to form the dataset. Florios et al. (2018) pointed out the probability of a hit and the probability of a false alarm depend on the response bias, which determines the activation level used as a threshold for a yes/no response. Leka & Barnes (2007) and Barnes et al. (2007) were based on the assumption that the prior probability increases with the increase of the sample size, and believed that the boundary of the discriminant calculation does not always pass through the class center.

Table 1 Major Methods for Treating the Class Imbalance Problem in Solar Flare Prediction and the Corresponding Evaluations

References	Algorithms	Strategies
Zhou (2019)	BPNN	Type I
Bobra & Couvidat (2015)	SVM,CCNN,RBF	Type I
Yu et al. (2010)	BN, DT, LVQ	Type I
Huang et al. (2010)	DT	Type I
Florios et al. (2018)	SVM,MLP,RF	Type II
Leka & Barnes (2007)	DA	Type II
Barnes et al. (2007)	PDA	Type II
Al-Ghraibah et al. (2015)	FH	Type III
Zheng et al. (2019)	CNN	Type III

It can be seen that the problem of class imbalance has received widespread attention, although it has not been studied systematically. We will focus on introducing sample weighting strategies to deal with class imbalances in short-term forecasts, and support them by comprehensively comparing the performance of various strategies.

The paper is organized as follows: Section 2 introduces data characteristics. Section 3 introduces and discusses several types of imbalance handling methods. Section 4 presents the evaluation results between methods. Section 5 conducts test experiments and performance comparison. Finally, Section 6 gives the conclusion of this article.

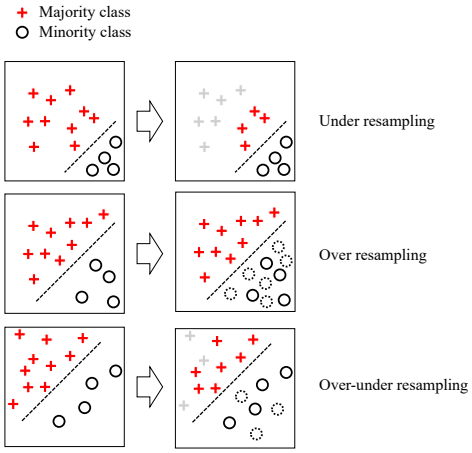

Fig. 2 Schematic diagram of resampling.

2 DATA CHARACTERISTICS

Flare data are from http://www.ngdc.noaa.gov/stp/SOLAR/ftpsolarflares.html#xray. Solar flares are generally graded as C, M or X. The 185 active regions from May 1996 to November 1999 were selected based on the criteria used. Each observable sample in these active regions was labeled as either positive or negative samples depending on the threshold of $I_{\rm tot}$, resulting in a dataset containing 1472 positive and 8528 negative samples. Considering that the difference in the number of positive and negative samples is around five times, this is known as the class imbalance problem and may bias the non-combustion predictions if the model is learned directly.

Taking the three predictors of total unsigned magnetic flux (TUMFLUX), magnetic field horizontal gradient (MFHG) and total unsigned vertical current (TUVC) as an example, the sample distribution of flare predictors is shown in Figure 1. The positive and negative samples are asymmetric in the samples space. After resampling processing such as oversampling processing, the sample will show a balanced trend.

3 COMPARISON OF CLASS IMBALANCE TREATMENT SOLUTIONS

3.1 Type I: Resampling

Resampling techniques reduce the problem of uneven distribution among samples and are commonly used to solve class imbalance problems. Once the training set is resampled, all algorithms can be adapted to the sampled dataset without further modification. The dataset is usually divided into two parts during training: the training set and the test set. The learning algorithm learns the prediction model from the training set and evaluates it with the test set.

Table 2 Different Outcomes of Two-class Prediction

	Predicted positive class	Predicted negative class
Actual positive class	True Positive (TP)	False Negative (FN)
Actual negative class	False Positive (FP)	True Negative (TN)

The basic resampling techniques include under-sampling (shown in Fig. 2) and over-sampling methods. The samples from the majority class are eliminated and the samples from the minority class are duplicated for the under-sampling and over-sampling methods, respectively. The over-under sampling is to eliminate majority class and duplicate minority class simultaneously, which modifies samples equal to the average of the minority and majority class numbers. Furthermore, some more intelligent resampling methods have been proposed to obtain the appropriate distribution.

3.2 Type II: Changing Decision Discriminant Boundary

Unlike the resampling technique, this method does not change the distribution of the training set. Generally, weights of prediction errors of each class are equal, and the discriminant boundary is determined by minimizing the overall error rate. The discriminant boundary is the vertical line passed through the midpoint between two populations.

To compensate for the effects of class imbalance, prediction errors for majority and minority classes are given different weights (Tulunay et al. 2004). Prediction errors of the majority class are assigned with a small weight and/or prediction errors of the minority classes are assigned with a large weight. This makes the decision bias to the minority class. Usually when determining the weight value of the decision boundary, the "average value" is used as the weight of the majority class and the minority class. Specifically, the influence of the two categories on the decision boundary is obtained by counting the numerical average of the majority class and the minority class, and then dividing it by the number of the majority class and the minority class, respectively, to obtain the weight ratio. Figure 3 shows the change of decision discriminant boundary, and the discriminant boundary moves from boundary 1 to boundary 2.

3.3 Type III: Sample Weighting

This method assigns different weights to training samples Liu et al. (2008). Most samples are given smaller weights and/or a few samples are given larger weights, which are then added to the learning algorithm. Sample weighting techniques have been widely used in class imbalance learning to improve the performance of some standard machine learning methods, such as decision tree and SVM (Brefeld et al. 2003).

Table 3 Performance of Different Strategies for Dealing with Class Imbalance Problem

index	raw	resample			decision weight	sample weight
		over	under	over and under		
Minority	0.11±0.04	0.70±0.05	0.60±0.06	0.70±0.04	0.63±0.04	0.72±0.04
Majority	0.99±0.01	0.73±0.02	0.77±0.04	0.72±0.03	0.76±0.02	0.71±0.02
AUC	0.55±0.02	0.71±0.03	0.68±0.02	0.71±0.02	0.69±0.02	0.71±0.02
HSS	0.17±0.06	0.28±0.03	0.26±0.04	0.27±0.03	0.27±0.03	0.27±0.03
Time	80.96±9.03	898.30±324.75	6.61±0.65	178.74±45.36	79.64±6.89	174.42±48.67

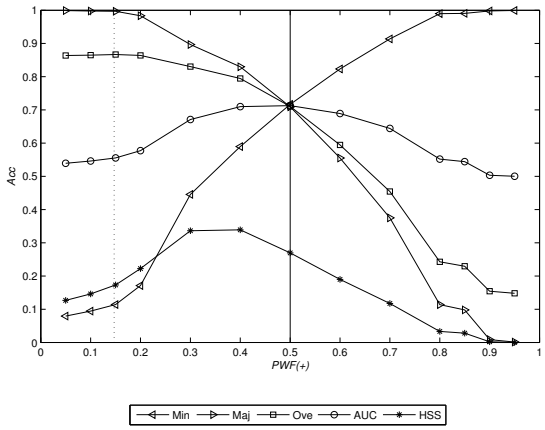

Fig. 3 Schematic diagram of changing decision discriminant boundary.

Both the method of type II and type III involve weights. Compared with type II, in terms of changing the decision boundary in the iterative process, type III usually requires modification of conventional machine learning methods to process sample weights. The fastest and most effective method is to calculate the average of the majority and minority samples, assign the weights of the two categories in the model training, and try to solve the class imbalance.

The weight of the sample represents the importance of the sample in the data mining process. Under the default weight (that is, without artificially changing the weight), the learned model is internally biased towards minority groups. "Sample weighting technology" is characterized by: it does not significantly change the execution time of the machine learning method, because it does not change the number of training samples; it has the ability to consider the prior knowledge of each sample separately; it can assign samples different weights to consider the difference in importance of different samples (such as redundant samples, boundary samples and noise samples). Liu et al. (2008) tries to combine the intelligent identification algorithm into the sample weighting technology.

4 PERFORMANCE EVALUATION

Learning from unbalanced data is still one of the difficulties of classifiers (Nnamoko & Korkontzelos 2020). Compared with the untreated samples directly used, if the effective class imbalance treatment is carried out, the prediction accuracy will inevitably be improved. This article used SVM as a prediction algorithm to compare different types of imbalances. SVM is a very simple and practical classifier, which has been applied maturely in all walks of life. The basic principle is to try to pass a hyperplane to correctly divide the positive and negative samples.

Because samples from minority classes have less impact on the overall accuracy, errors occur when the overall classification accuracy metric is used to evaluate model performance. In class imbalance problems, some additional performance metrics have been proposed. Such as estimating the performance of minority and majority classes (TP and TN), ROC (Receiver operating characteristic) curve and AUC (Area under the ROC curve), HSS (Heidke Skill Score), etc., these parameters are often used to quantify the performance of forecasting models. The TP rate and TN rate are considered here to evaluate the performance of the sample. The reason is that, theoretically if the classifier can perform the correct classification, then the TP rate and TN rate will be. There is a noticeable improvement. Therefore, TP rate and TN rate are the most direct and effective means. The above performance evaluation indexes are calculated based on the statistical results in Table 2.

The overall classification accuracy is computed as:

$$\mathrm{PC} = \frac{\mathrm{TP} + \mathrm{TN}}{N}, \tag{1}$$

where $N = \mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}$. For the class imbalance problem, the PC (Proportion Correct) is not suitable to measure the performance of the prediction model. A refined evaluation is to distinguish the performance of each class. TP rate and TN rate are used to estimate the performance of positive and negative classes, respectively.

$$\mathrm{TPrate} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{2}$$

$$\mathrm{TNrate} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}. \tag{3}$$

In order to get more satisfactory single performance evaluation indexes, AUC and HSS is proposed based on the signal detection theory.

Table 4 Comparisons of Class Imbalance Learning Methods

Index	Resampling			Decision weighting	Sample weighting
	Over	Under	Over and Under		
Universality	Excellent	Excellent	Excellent	Average	Not good
Performance	Excellent	Average	Excellent	Average	Excellent
Time Consuming	Large	Small	Average	Average	Average

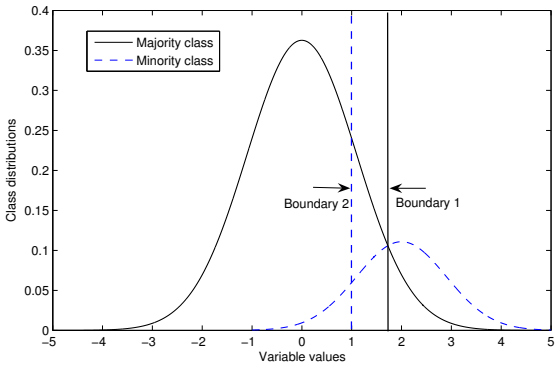

Fig. 4 Performance indexes versus PWF(+).

ROC, a curve of hit rate (TP rate) against false alarm rate (FP rate) for different decision thresholds, is a tool to distinguish two alternative possibilities in the signal detection theory. Given a certain threshold, a single point on this curve can be obtained. The area under the ROC curve (AUC) is used to calculate the ROC-based measure of skill (Liu et al. 2008). The AUC for a certain decision threshold is computed as:

$$\mathrm{AUC} = \frac{1 + \mathrm{TPrate} - \mathrm{FPrate}}{2}. \tag{4}$$

Equation (5) is used to quantify the performance of the prediction model (Yu et al. 2010).

$$\mathrm{HSS} = \frac{\mathrm{PC} - E}{1 - E}, \tag{5}$$

where $E = \frac{(\mathrm{TP+FN})(\mathrm{TP+FP})}{N^2} + \frac{(\mathrm{TN+FP})(\mathrm{TN+FN})}{N^2}$. When the forecasts and observations are independent, the prediction model is in absence of skill. As shown by the definition of HSS, this parameter shows the predictive power of the model. In general, HSS is -1 in number if the prediction is wrong, and 1 if all predictions are correct. its intermediate state of 0 represents the randomness of the prediction.

5 RESULTS

Cadence of the solar magnetogram data samples was 96-minute sets. Span of time is 3 years from 1996 for 185 active regions. The ratio of positive to negative samples is 1472:8528. We used a 10-fold cross-validation to verify the performance of the prediction model: the mean of the results of 10 folds of the prediction model was selected to characterize the model accuracy, and its standard deviation was used to estimate the uncertainty.

5.1 Performance Results of Different Strategies

We compare the results of using different strategies in solar flare prediction for dealing with class imbalance problems, and their performance is shown in Table 3. After adjusting the strategies for handling the class imbalance problem, the prediction performance is balanced between positive and negative samples. Analyzing the performance of different strategies, we found that (1) the performance of over-sampling is better than that of under-sampling, however, the over-sampling method is time-consuming. The strategy composed by over-sampling and under-sampling is a compromise between them. (2) The sample weighting strategy has the best performance. However, a machine learning algorithm should be redesigned to treat the weights of samples. (3) The performance of changing decision discriminant boundary is comparable with that of the resampling strategy.

5.2 Performances of Sample Weighting Strategy with Different Weights

In this experiment, we assign an inverse class sample weight to each sample before performing data learning. Then, in order to obtain the optimal value of the inverse sample weight, we performed an analysis of the impact of this value on the learning performance. Its probability weighting function is defined as follows:

$$\mathrm{PWF}(+) = \frac{n_1 \cdot w_1}{n_1 \cdot w_1 + n_2 \cdot w_2}, \tag{6}$$

where n_1 and n_2 denote the number of positive and negative samples, respectively. The corresponding weights are w_1 and w_2.

PWF(+) is positively correlated with a few sample weights. In the special case i.e., PWF(+) $=$ 0.5, the class imbalance problem has been eliminated for the characterized dataset. In the comparison experiments, by changing the weights w_1 and w_2, the variation of performance metrics with PWF(+) is obtained as shown in Figure 4. As shown in this figure, the use of changing weights has a positive impression for solving sample problems with class imbalance; on the other hand, by comparing a large number of datasets it is clear that the accuracy of the model varies with PWF(+). Exceptionally, the AUC is usually optimal or suboptimal for PWF(+) $=$ 0.5. The above results reinforce that the method (i.e.,

inverse class probability weighting) is a useful approach to solve the class imbalance problem in the data sample.

6 CONCLUSIONS

The prediction of solar flares usually has the problem of imbalance, which will reduce the relevant performance of the model. For most data mining algorithms used to build the model of flare prediction, the problem of class imbalance should be solved before constructing a prediction model based on observation data.

The main work of this paper summarizes the characteristics of a series of methods that are typically used to solve the problem of class imbalance. We found that there is not a perfect algorithm that performs best in all aspects. We believe that in practical applications, we need to choose an appropriate strategy according to different computing time requirements, accuracy requirements, and other goals. Refer to Table 4 for specific recommended content. For example, in solar flare forecasting problem, if only the algorithm with the fastest calculation speed is needed and other requirements are abandoned, we recommend Type I under-sampling method.

To better illustrate the sensitivity of class imbalance processing methods to different prediction algorithms, more detailed work is needed, such as comparing the performance of class imbalance algorithms combined with SVM and LS-SVM. Due to space limitations, this article does not give a key explanation. Obviously, this is one of the tasks we have to do next.

Acknowledgements We thank the SOHO/MDI consortium for the data. SOHO is a project of international cooperation between ESA and NASA.

References

Al-Ghraibah, A., Boucheron, L. E., & McAteer, R. T. J. 2015, in 2015 IEEE International Conference on Data Mining Workshop (ICDMW) (IEEE)

Barnes, G., Leka, K. D., Schumer, E. A., & Della-Rose, D. J. 2007, Space Weather, 5, S09002

Bobra, M. G., & Couvidat, S. 2015, ApJ, 798, 135

Brefeld, U., Geibel, P., & Wysotzki, F. 2003, in Machine Learning: ECML 2003, eds. N. Lavrač, D. Gamberger, H. Blockeel, & L. Todorovski (Berlin, Heidelberg: Springer Berlin Heidelberg), 23

Florios, K., Kontogiannis, I., Park, S.-H., et al. 2018, Sol. Phys., 293, 28

Huang, X., Wang, H., Xu, L., et al. 2018, ApJ, 856, 7

Huang, X., Yu, D., Hu, Q., Wang, H., & Cui, Y. 2010, Sol. Phys., 263, 175

Leka, K. D., & Barnes, G. 2007, ApJ, 656, 1173

Li, R., Wang, H.-N., He, H., Cui, Y.-M., & Zhan-LeDu. 2007, ChJAA (Chin. J. Astron. Astrophys.), 7, 441

Liu, J., Hu, Q., & Yu, D. 2008, Information Sciences, 178, 1235

Nnamoko, N., & Korkontzelos, I. 2020, Artificial Intelligence in Medicine, 104, 101815

Park, S.-H., Leka, K. D., Kusano, K., et al. 2020, ApJ, 890, 124

Roy, S., Prasad, A., Ghosh, K., Panja, S. C., & Patra, S. N. 2020, RAA (Research in Astronomy and Astrophysics), 20, 110

Tulunay, Y., Tulunay, E., & Senalp, E. T. 2004, Advances in Space Research, 33, 988

Yi, K., Moon, Y.-J., Shin, G., & Lim, D. 2020, ApJL, 890, L5

Yu, D., Huang, X., Wang, H., & Cui, Y. 2009, Sol. Phys., 255, 91

Yu, D., Huang, X., Wang, H., et al. 2010, ApJ, 710, 869

Zheng, Y., Li, X., & Wang, X. 2019, ApJ, 885, 73

Zhou, G. 2019, in Journal of Physics Conference Series, 1302, 042027