Sérgio Luisir Díscola Junior, José Roberto Cecatto,
Márcio Merino Fernandes, and Marcela Xavier Ribeiro

29.1 Introduction

Solar Flares (SF) are sudden releases of large amounts of energy from the solar atmosphere [1]. They are categorized into 5 classes, namely, *A, B, C, M and X*, respectively in order of their strength, where SF of class *A* are the least harmful, while *X* flares are the most powerful and dangerous ones. They are categorized according to the level of X-ray emitted by Sun during the event. These phenomena impact satellite communications [2], Global Positioning System (GPS) and may also produce electricity power blackouts. So, it is imperative to develop robust solar flare forecasting systems.

The features that influence the forecasting process are not known. We can find papers that use features derived from magnetogram vector [3–5], sunspot area [6], radio flux or X-ray flux [7] and [8] the X-rays time series.

Solar Flare datasets are extremely imbalanced. Most work in literature using traditional classifiers that deal with imbalanced datasets have the drawback of producing biased results [9–11]. An alternative to handle the poor results of the learning in imbalanced data is the usage of an Ensemble of Classifiers (EC) [12–14]. The EC main goal is to improve "weak" classification methods by applying many of weak classifiers (also called base inducers), so that the final classification may produce more accurate results. In this sense,

we propose an EC tuned-up for the domain of the Solar Flare forecasting.

Most of the previous works of Solar Flare forecasting perform binary forecasting, classifying solar flare only as "Positive" or "Negative". Few works predict individual classes. In the latter case, they usually use purely statistical methods in the forecasting process. Some methods consider "Positive" results for classes greater than or equal to "C" [15], others consider "Positive" for forecasts greater than or equal to a class M [3–5, 7]. Furthermore, a recurrent practice in the previous literature works is the forecasting of just the maximum Solar Flare of a given day.

Accordingly, we propose a method called ECID (Ensemble of classifiers for imbalanced datasets) that tackles some important open issues:

(1) Perform individual class forecasting producing a multi-class result for a given day, so that the method provides to the astrophysicist a tool that shows possible Solar Flare categories that may happen in a given day;
(2) Treat the imbalanced dataset issue using Ensemble with a stratified random sampling for the training of the inducers;
(3) Perform also a multi-label solution, giving the possibility to the astrophysicist to decide (when adjacent classes are indiscernible).

29.2 Method Description

In this section, we describe ECID (Ensemble of Classifiers for Imbalanced Datasets) and its pre-processing steps. Figure 29.1 presents its overview.

Steps 1, 2 and 3 are the preprocessing steps responsible for obtaining, cleaning and transforming the data for the learning task. Traditional classification methods produce models that

S. L. Díscola Junior (✉)
Department of Computer Science, Federal University of São Carlos, São Carlos, SP, Brazil

Federal Institute of São Paulo, São Carlos, SP, Brazil
e-mail: sergio.discola@ufscar.br

J. R. Cecatto
National Institute for Space Research, São José dos Campos, SP, Brazil

M. X. Ribeiro · M. M. Fernandes
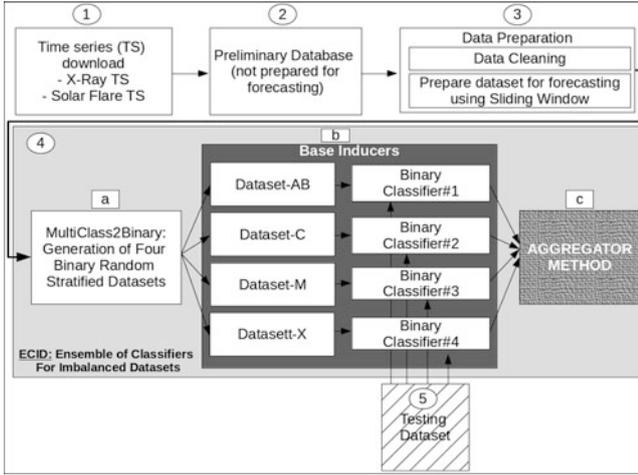Department of Computer Science, Federal University of São Carlos, São Carlos, SP, Brazil

Fig. 29.1 Overview of the proposal for solar flare forecasting: the method ECID and its pre-processing

classify instances with **"current" events**. For the forecasting purpose, it is necessary to map current values with future events to turn such classifiers in forecasting methods. Thus, our method prepares the original dataset to the forecasting by using the "Sliding Window" algorithm proposed in [8].

The prepared data are submitted to the proposed ensemble method ECID. The issue of imbalanced data is tackled by employing a stratified random sampling, which produce the datasets employed to train the base inducers. This is done by splitting the original dataset into four balanced subsamples, one for each forecasting class, using the new "Multi-Class2Binary Balancing" schema (see Fig. 29.1 – Step 4(a)).

The base inducers are weak binary classifiers (see Fig. 29.1 – Step 4(b)): the "Binary Classifier#1" generates a "Positive" result if a class "A or B" is predicted by the model and "Negative", otherwise; "Binary Classifier#2" generates a "Positive" result if a class "C" is predicted and "Negative", otherwise; and the same logic is applied for the classes "M" and "X". The output of the base inducers are submitted to the aggregator method (see Fig. 29.1 – Step 4(b)), which combines the individual votes of each inducer, producing the final multi-class forecasting. The forecasting is also multi-label when the aggregator function returns more than a class.

Step 5 provides the testing dataset to validate the model, which comprises the time series of the current day. Hence, ECID provides forecasting for the next day. A detailed description of the steps of ECID is given in following.

29.2.1 The Preprocessing Steps

As shown in Fig. 29.1, in Steps 1 and 2, the time series of X-ray intensity and the solar flare report are collected from their sources. These time series have different sample rates: the X-ray time series have a sample rate of 12 min, and the

solar flare report has a "varying" sample rate depending on the duration of the Solar Flare events. So, for one entire day, the X-ray time series produces 120 instances, and the solar flare report produces a varying amount of instances. The next task is to map the X-ray intensity and the solar flare report, as formally explained next:

- the time series of X-ray intensity is defined as $X = \{instant\,of\,X\text{-}ray\,obs,$ $X\text{-}ray\,intensity\}$, where:
  - $instant\,of\,X\text{-}ray\,obs$ is the observation instant of time;
  - $X\text{-}ray\,intensity$ is the intensity of X-ray measured given by W/m^2.
- the solar flare report contains summarized information about solar events and it is defined as $E = \{instant\,of\,Solar\,Flare,$ $Solar\,Flare\,Class\}$, where:
  - $instant\,of\,Solar\,Flare$ is the instant of time that a solar flare was observed;
  - $Solar\,Flare\,Class$ is the class of a solar flare.

The mapping between these time series produces the "Preliminary Database" defined as:

- a tuple $T = \{instant\,of\,X\,Ray\,observation,$ $X\text{-}ray\,intensity, Solar\,Flare\,Class\}$;
- $instant\,of\,X\,Ray\,observation$ is the instant when the X-ray was emitted by the Solar Flare;
- $X\text{-}ray\,intensity$ is the X-ray emitted by the Sun in $instant\,of\,X\,Ray\,observation$ and,
- $Solar\,Flare\,Class$ is the class of the Solar Flare occurred in that instant.
- the "Preliminary Database" is $M^{XE} = \cup T$.

In Step-3, the data is cleaned by discarding tuples with troubled values and, then, the "Sliding Window" approach [8] is applied to map current instances to future events (classes of solar flares), enabling the learning model to forecast. The Sliding Window approach builds a set formally defined as:

- $|M^{XE}|$ is the number of instances of the "Preliminary Dataset", M^{XE};
- $S^{SFD} = \{x_i(t), x_i(t+1), \ldots,$ $x_i(t + currentWindowSize),$ $maximumClassOfFutureWindow,$ $1 \le i \le |M^{XE}|\}$, where S^{SFD} is the "Slided Solar Flare Dataset", $x_i(t)$ is the $X - ray\,intensity$ of the current Window, $currenWindowSize$ is the size of the current window, and $maximumClassOfFutureWindow$ is the maximum solar flare class occurred in the future window (for a detailed explanation of current and future window, see [8]);

29.2.2 The Proposed Method ECID (Ensemble of Classifiers for Imbalanced Datasets)

The preprocessed data is submitted to ECID that employs a modified bootstrap strategy: ECID builds multiple learning models of the same type from different subsamples of the training dataset. Specifically, it splits the training dataset in subsamples using a strategy that we named *MultiClass2Binary*.

"MultiClass2Binary" strategy is the key to provide a strong multi-class forecasting for 4 classes of solar flares (AB, C, M and X). This strategy builds 4 distinct balanced datasets using stratified random undersampling. The sampling strata are the solar flares classes, and the sampling schema is detailed next. The first dataset is composed of 50% of the tuples classified as classes *A* or *B* and 50% of remaining; the second one is composed of 50% of tuples class C and 50%, the remaining; the third is balanced in the same way for class M, and the last one for class X.

In Step-4(b) Fig. 29.1, each dataset is applied to a binary inducer from the same type (eg. a decision tree algorithm). Accordingly, the **Binary Classifier#1** provides the forecasting for class *A* or *B*, **Binary Classifier#2**, for class C, **Binary Classifier#3**, for class M, and, **Binary Classifier#4**, for class X. We propose to employ all inducers of the same type because preliminary experiments did not demonstrated improvements by combining different types of inducer. However, the combination of different inducers types could also be performed.

In Step-4(c), a testing dataset is applied to each model to produce specific results for each solar flare class. For solar flare prediction, the sample rate of the input is 12 min. Different from traditional approaches, which the input of testing is a single tuple, the testing input is a set of 120 tuples (collected in a day to forecast the most high class for the next day). The Aggregator method combines the 120 results obtained in the last task, providing unified multi-class forecasting for a given day.

The Aggregator method is based on a voting count schema. Table 29.1 shows a simplified example of the output produced by the base inducers, which are binary classifiers. The "instant" column corresponds to the daily 12 min sample vote by each inducer. Columns z_Y (where $z \in \{AB, C, M, X\}$) is set to 1 if the correspondent inducer produced positive forecasting for the given class. Consider the tuple with "Day_1_00:00", for example, if **Binary Classifier#1** produced a "Positive" forecasting, the AB_Y column is set to 1; if **Binary Classifier#2** produced a "Positive" forecasting, the C_Y column is set to 1; and so on. Note that, the forecasting is originally given in a 12 min rate, but the goal of this work is to produce a daily forecasting. So, the Aggregator combines the votes of the base inducers producing a daily result.

Table 29.1 Example of the base inducers individual forecasting

Instant	AB_Y	C_Y	M_Y	X_Y
Day_1_Instant_00:00	1	1	0	0
Day_1_Instant_00:12	…	…	…	…
Day_1_Instant_00:24	…	…	…	…

After calculating the daily sum of the z_Y columns of Table 29.1, the Aggregator decides among the classes, considering the three most voted ones as formally defined: Let *firstYes, secondYes, thirdYes* be the number of votes obtained, respectively, by the first three classes most daily voted in the aggregated table.

The parameters p_1, p_2 and p_3 are the least vote frequency acceptable to the class to be considered for *firstYes, secondYes*, and *thirdYes*, respectively, i.e. parameters p_1, p_2 and p_3 denote the daily minimum percentage of votes required for the class to be chosen. We found empirically that the best results are obtained using $p1 = 20\%$, $p2 = 5\%$ and $p3 = 2\%$. These parameters values show that as the class voting becomes smaller, the class becomes rarer, so it should be kept in the results. In fact, as the class frequency becomes smaller, it will probably be less foreseen, then the importance of the minority class should be increased.

29.3 Experiments

Experiments 1–6 were performed to validate the proposed method. All the experiments:

- used the same solar dataset composed of the time series of X-ray intensity emitted by Sun and the solar flare report from the period 2010 to 2017;
- employed the data pre-processing described in Sect. 29.2.1;
- used 70% of the dataset for training and 30% for testing to validate the method. More specifically, the training dataset have 67678 instances (50% of class AB, 35% of class C, 13% of class M, and 2% of class X). The dataset is extremely imbalanced, and according to astrophysicist, the most relevant class are M and X, which are the most infrequent ones.

For experiments 1 to 3, we employed a traditional classification method alone to produce the forecasting model: Experiment 1 used IBK; Experiment 2 employed SVM; and Experiment 3 employed J48. The implementation of these methods was obtained from Weka [16].

For experiments 4 to 6, we employed the proposed ensemble method ECID (see Sect. 29.2.2). The stratified undersampling performed in the Step 4(a) of ECID (see Fig. 29.1) is shown in Fig. 29.2. The training dataset contains 33839
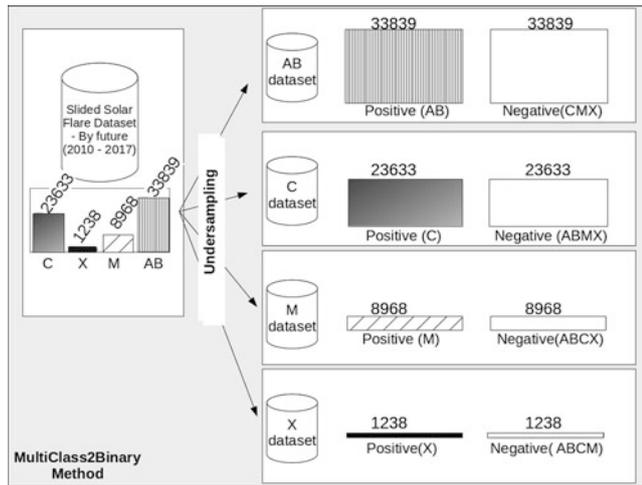
Fig. 29.2 ECID sampling (see Fig. 29.1)

Fig. 29.3 Experiments – arithmetic mean

tuples labeled as class "A or B" (called "AB"), 23633 labeled as "C", 8969 as "M" and 1238 as class "X". The method randomly undersamples the original dataset according to each class. For example, the balanced dataset-AB is composed of 33839 tuples labeled "A" or "B" of the original dataset, these tuples will be re-labeled as "Positive(AB)", the other 33839 tuples are composed by a subsample of the original tuples labeled as classes "C", "M" and "X", which will also be re-labeled as "Negative(CMX)". The experiments 4–6 used this strategy for undersampling, but they vary the base inducers in each experiment: in Experiment 4 the base inducers were IBK; in Experiment 5 were SVM; and in Experiment 6 the base inducers were J48.

The metrics used to analyze the results must be carefully interpreted. Analyzing only accuracy (ACC) may incur in error due to the majority classes have higher probability to be predicted then the minority ones. In that case, the overall ACC may perform well, but the most relevant classes (the minority) should be poorly predicted. Then, for a better analysis, the True Positive Rate (TPR) and the Precision for each class may be balanced. In this case, it means that either the model truly predicted most classes. Some metrics aim to give numeric values for this "balance", such as **TSS** which is the difference between recall and False Positive Rate (FPR) [4], and **F-Measure**, that relates Precision and Recall. Our method was validated using the joint analysis of accuracy (ACC), True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), TSS and F-Measure. As the results obtained are multi-class, and sometimes multi-label, we took some cautions:

(1) First, it was calculated individual metrics for each class (AB, C, M and X) by considering "Positive" for a specific class, and "Negative" as all the remaining classes;
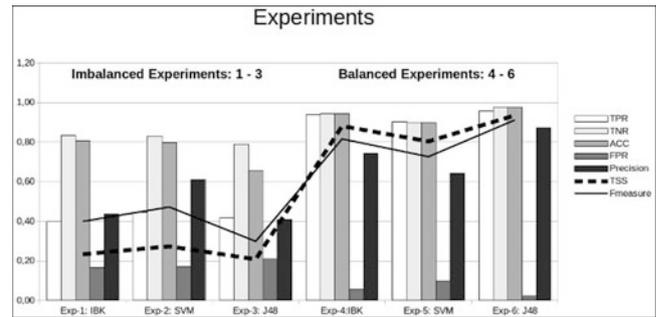
(2) Second, the arithmetic mean of each metric for all classes was obtained to compare the experiments performed (shown in Fig. 29.3).

As observed in Table 29.2, the classifiers alone produced relatively good results for predictions to classes AB and C (which are the majority classes), and very poor results for classes M and X (the minority and most important classes). The highest results for classes AB and C were 72.8% of TPR and 64.1% of Precision in Experiment-2 (SVM), and 76% of TPR and 70.8% of Precision for class C in Experiment-1 (IBK). In the other hand, it reached 34.2% of TPR and 12.4% of Precision for class M, and 73.3% of TPR and 10.8% of Precision for class X. Although a good TPR was obtained for class X, a very poor Precision drastically decreased the level of reliability of the forecasting model. TSS and F-measure metrics corroborate these results, which were reasonable for classes AB and C, and very poor for classes M and X.

Table 29.3 shows the results of Experiments 1–6, which used our proposed method ECID. The results obtained for the minority, but most important, classes were higher than the previous experiment using the classifiers alone. The **best results for class X** were achieved in Experiment 6 using J48 as the base inducers of ECID: 86.6% of TPR and 76.4% of Precision, TSS of 0.85 and an F-Measure of 0.812. Also, the predictions of class M resulted in 100% of TPR and 80% of Precision. Predictions of classes AB and C achieved results of more than 95%.

Figure 29.3 shows the increase of the individual metrics from the first three experiments (not using ECID) and the last three ones (using ECID). TPR had an increase of more than 50%, TNR, an increase of 12.5%, FPR, a decrease of 33%, and the Precision an increase of 56%. As shown in Fig. 29.3, TSS and Precision increased with the proposed method.

29.4 Related Works

In literature, Solar flare forecasting methods usually aggregate more than one class as the "Positive" and "Negative",

Table 29.2 Experiments 1–3 – Metric Details

	Experiment-1: IBK				Experiment-2: SVM				Experiment-3: J48			
	AB	C	M	X	AB	C	M	X	AB	C	M	X
TPR	0,635	0,760	0,197	0	0,728	0,649	0,131	0,266	0,309	0,286	0,342	0,733
TNR	0,833	0,542	0,962	1	0,741	0,606	0,971	0,998	0,978	0,757	0,597	0,825
ACC	0,757	0,638	0,853	0,971	0,736	0,625	0,851	0,977	0,693	0,549	0,561	0,822
TSS	0,469	0,302	0,160	0	0,469	0,255	0,103	0,264	0,288	0,043	−0,06	0,558
FPR	0,166	0,457	0,037	0	0,258	0,393	0,028	0,001	0,021	0,242	0,402	0,174
Precision	0,708	0,566	0,468	0	0,641	0,565	0,434	0,8	0,914	0,482	0,124	0,108
F-Measure	0,670	0,649	0,277	0	0,681	0,604	0,202	0,4	0,462	0,359	0,182	0,189

Table 29.3 Experiments 4–6 – Metric Details

	Experiment-4: IBK				Experiment-5: SVM				Experiment-6: J48			
	AB	C	M	X	AB	C	M	X	AB	C	M	X
TPR	0,900	1	0,916	0,933	0,892	0,910	0,875	0,933	0,966	1	1	0,866
TNR	0,978	0,906	0,945	0,943	0,907	0,893	0,902	0,898	0,985	0,962	0,974	0,983
ACC	0,942	0,942	0,942	0,942	0,900	0,900	0,900	0,900	0,977	0,977	0,977	0,977
TSS	0,879	0,906	0,861	0,876	0,799	0,804	0,777	0,831	0,952	0,962	0,974	0,850
FPR	0,021	0,093	0,054	0,056	0,092	0,106	0,097	0,101	0,014	0,037	0,025	0,016
Precision	0,973	0,870	0,628	0,5	0,892	0,844	0,477	0,358	0,983	0,943	0,8	0,764
F-Measure	0,935	0,930	0,745	0,651	0,892	0,876	0,617	0,518	0,975	0,971	0,888	0,812

Table 29.4 Comparison among literature works and ECID

	TPR	TNR	ACC	FPR	Precision	TSS	F-Measure
ECID	0,96	0,98	0,98	0,02	0,87	0,94	0,91
Nishikawa (=X)	0,90	0,99	0,99	0,0003	0,89	0,91	0,89
Nishikawa (>=M)	0,91	0,99	0,99	0,002	0,92	0,91	0,92
Bobra (>=M)	0,71	0,98	0,97	–	0,80	0,70	0,75
Li (>= M)	0,73	0,78	0,77	–	–	–	–

turning it a binary classification model. Instead of this, our method provides the contribution to give multi-class (and multi-label) forecasting, because, in a single day, it is important to identify the correct class of solar flare that occurs in a day. Additionally, as the astrophysicists do not fully understand these phenomena, each work considers a different set of solar features as input. Taking those facts into account, we can cite some relevant works that we named:

- *Nishizuka*: In [3], it was developed a solar flare forecasting method that labels: (1) $classes = X$ as "Positive", in a first experiment, and (2) $classes \geq M$ as "Positive", in the last set of experiment;
- *Bobra*: In [4], it was developed a solar flare forecasting method that labels $classes \geq M$ as "Positive";
- *Li*: In [5], it was developed a solar flare forecasting method that labels $classes \geq M$ as "Positive" according to an equation of flare importance;

Table 29.4 presents the metrics obtained by the *Nishizuka, Bobra and Li* forecasting methods and our proposed method ECID:

As shown in Table 29.4, ECID achieved TSS and Precision higher than *Nishikawa* and the other studies. If we consider just the forecasting of class "X", our method obtained a TSS of 0,85 and an F-measure of 0,81. These results are slightly lower than *Nishikawa* ones, but our method also predicted flares of classes "AB", "C" and "M" consistently, instead of *Nishikawa*, which performed a binary classification. Thus, we believe that our proposed approach is an important contribution for the task of solar flare forecasting.

29.5 Conclusions

Most works for solar flare forecasting performs a binary prediction. Therefore, in that scenario the problems of imbalanced data and high similarity between adjacent classes are minimized. Our contribution was to deal with these problems by proposing the new ensemble method ECID, producing a multi-class forecasting. For each solar flare class, ECID employs a stratified random sampling for the training of base one-class inducers, strengthen their sensitivity to one-class. Using a modified bootstrap approach, the aggregator

method combines the inducers results enabling a strong multi-class forecasting, which can also be multi-label in case of indiscernible classes. The results obtained showed that our proposal is well-suited to solar flare forecasting, achieving 86.6% of TPR and 76.4% of Precision for class X, 100% of TPR and 80% of Precision for class M, 100% of TPR and 94.3% of Precision for class C, and 96.6% of TPR and 98.3% of Precision for class AB. For future work, we intend to add more solar features in the process.

References

1. Holman, G.D.: The mysterious origins of solar flares. Sci. Am. **294**(4), 38–45 (2006)
2. Basu, S., Basu, S., MacKenzie, E., Bridgwood, C., Valladares, C.E., Groves, K.M., Carrano, C.: Specification of the occurrence of equatorial ionospheric scintillations during the main phase of large magnetic storms within solar cycle 23. Radio Sci. **45**(5), 1–15 (2010)
3. Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., Watari, S., Ishii, M.: Solar flare prediction model with three machine-learning algorithms using ultraviolet brightening and vector magnetograms. Astrophys. J. **835**(2), 156 (2017)
4. Bobra, M.G., Couvidat, S.: Solar flare prediction using SDO/HMI vector magnetic field data with a ML algorithm. Astrophys. J. **798**(2), 135 (2015)
5. Yu, D., Huang, X., Hu, Q., Zhou, R., Wang, H., Cui, Y.: Short-term solar flare prediction using multiresolution predictors. Astrophys. J. **709**(1), 321–326 (2010)
6. Gallagher, P.T., Moon, Y.-J., Wang, H.: Active-region monitoring and flare forecasting I. Data processing and first results. Sol. Phys. **209**(1), 171–183 (2002)
7. Li, R., Zhu, J.: Solar flare forecasting based on sequential sunspot data. Res. Astron. Astrophys. **13**(9), 1118–1126 (2013)
8. Discola Jr, S., Cecatto, J., Fernandes, M., Ribeiro, M.: SeMiner: a flexible sequence miner method to forecast solar time series. Information **9**(1), 8 (2018) [Online]. Available: http://www.mdpi.com/2078-2489/9/1/8
9. García, S., Herrera, F.: Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy. Evol. Comput. **17**(3), 275–306 (2009)
10. Wallace, B.C., Small, K., Brodley, C.E., Trikalinos, T.A.: Class imbalance, redux. In: 2011 IEEE 11th International Conference on Data Mining, pp. 754–763. IEEE (2011)
11. Han, H., Wang, W.-Y., Mao, B.H.: Borderline-smote: a new oversampling method in imbalanced data sets learning. In: Advances in Intelligent Computing, pp. 878–887. Springer, Berlin/Heidelberg (2005)
12. Sagi, O., Rokach, L.: Ensemble learning: a survey. Wiley Interdisciplinary Rev. Data Min. Knowl. Discov. **8**(4), e1249 (2018)
13. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **42**(4), 463–484 (2012)
14. Rätsch, G., Onoda, T., Müller, K.-R.: Soft margins for adaboost. Mach. Learn. **42**(3), 287–320 (2001)
15. Ahmed, W., Qahwaji, R., Colak, T., Higgins, P.A., Gallagher, P.T., Bloomfield, D.S.: Solar flare prediction using advanced feature extraction, machine learning, and feature selection. Sol. Phys. **283**(1), 157–175 (2013)
16. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: an update. SIGKDD Explor. **11**(1), 10–18 (2009)