URSI - RCRS 2022, IIT (Indore), India, 1 - 4 December, 2022

Evaluation of different Machine Learning Models for identifications of Flares with CMEs

Hemapriya Raju*(1), Saurabh Das (1)
(1) Indian Institute of Technology Indore

Abstract

Solar eruptions such as CMEs and flares causes geomagnetic and communication disturbances on Earth. CMEs can be found in conjunction with flares, filaments, or independent. Although both flares and CMEs are understood as triggered by a common physical process magnetic reconnection, yet, the degree of association among them is unknown. We attempted to model the association of CMEs with flares through extensive Machine Learning models to study the occurrence of CMEs. Further, to improve the class separability, we have utilized the parameter change information obtained from the respective subsequent time difference. Since there is significant imbalance between the classes, we had explored approaches such as under sampling majority class, oversampling minority class and synthetically generated minority samples through SMOTE Technique. We achieved TSS around 0.81 without adding change information, and TSS around 0.92 after adding change information as additional feature on prediction of CMEs associated with flares for LDA, after addressing the class imbalance issues.

1. Introduction

Space weather is the dynamic variation in the Earth's atmosphere due to the effect of solar activity interacting with the Earth's magnetic field. These variations can be sudden, intense and eruptive due to events such as flares, Coronal Mass Ejections (CMEs) or can be steady and prolonged such as Corotating Interacting Regions (CIRs) due to open streams originating from Coronal Holes [1]. Though we knew from the past studies that occurrence of flares and CMEs are independent of each other, but the underlying physical mechanisms for both these events are thought to be from same magnetic phenomena caused by magnetic reconnection [2]. Major flares such as M and X class flares or subsequent flares are seen often associated with CMEs [3]. Flares with longer duration has greater chance of CME association varying from 26% for flares less than 2 hours to 100% for flares lasting more than 6 hours for the period 1996-1999 [3]. Though flares of any duration can be associated with CMEs. Later a statistical study has been done by Seiji et al., [4] between Flare's peak flux, its fluence and duration from the period 1996-2007 and it showed a significant correlation between flare parameters and CME's kinetic energy. The study also found that most frequent flaring cite is the center of CME span.
EUV and HMI images gave better understanding of the active region and CME characteristics. 18 features are derived from the automated AR region patches called Spaceweather HMI Active Region Patches (SHARP). The statistical studies show the contribution of individual AR features to the flares and CMEs. However, it is observed that there is no single individual parameter that shows significant variation in distinguishing these two populations. Combination of parameters helps in better classification. However, since there is no fixed combination, and performing it with large number of features manually will be tedious classification process.
Machine learning and deep learning is good in analyzing underlying patterns, and has capability to adapt and learn new features. Since understanding the underlying process of these events are highly complex, Machine learning approach can be used to study these underlying combinations. [5] explored the ML algorithms Cascade Correlation Neural Networks (CCNN) and Support Vector Machines (SVM) to study relationship between flares and the associated CMEs and achieved Area Under the Curve (AUC) up to 0.88 for two input features. Later [6] used physical features of active regions derived from photospheric vector magnetic field data obtained from Solar Dynamics Observatory (SDO) as input to the SVM and obtained True Skill Score (TSS) of 0.8±0.2.
The occurrence of flare associated with CMEs are very less in number than flares without CMEs. So, the dataset will be heavily imbalanced, which further affects the model's performance for our desired minority class. In this study we have explored several ML methods performance for prediction of CMEs. We also had studied the performance of ML models before and after addressing the class imbalance through different techniques such as under sampling, oversampling, Synthetic Minority oversampling technique (SMOTE). Further we had explored the ML model's performance with adding change information of each feature.

2. Dataset and Class imbalance

The data source is taken as physical parameters derived from Active region patches of full-disk photospheric vector magnetic field data obtained from Solar Dynamic Observatory's, Helioseismic and Magnetic Imager (SDO/HMI) images [7]. Dataset of Space weather HMI

Active Region Patches (SHARPs) are those automatically identified, tracked and derived parameters [6]. The cadence of the dataset is 12 minutes for an AR. 18 physical features are taken as input dataset for the years 2011-2018. The features are taken at 8, 12, 24, 36 and 48 hours prior to the occurrence of flare only and flare associated CME events. The respective features are shown in Fig 1. These features are derived based on the magnetic field, doppler velocity, continuum intensity and Line of sight magnetic field properties.

Parameter	Keyword	Formula
MEANGBH	Mean gradient of horizontal field	$\lvert\nabla\bar{B}_h\rvert = \frac{1}{N}\sum\sqrt{\left(\frac{\partial B_h}{\partial x}\right)^2 + \left(\frac{\partial B_h}{\partial y}\right)^2}$
MEANJZH	Mean current helicity (Bz contribution)	$\bar{H}_c \propto \frac{1}{N}\sum B_z \cdot J_z$
MEANALP	Mean characteristic twist parameter	$\alpha_{\text{total}} \propto \frac{\sum J_z \cdot B_z}{\sum B_z^2}$
MEANGBT	Mean gradient of total field	$\lvert\nabla\bar{B}_{\text{tot}}\rvert = \frac{1}{N}\sum\sqrt{\left(\frac{\partial B}{\partial x}\right)^2 + \left(\frac{\partial B}{\partial y}\right)^2}$
MEANPOT	Mean photospheric magnetic free energy	$\bar{\rho} \propto \frac{1}{N}\sum \left(B^{\text{Obs}} - B^{\text{Pot}}\right)^2$
MEANSHR	Mean shear angle	$\bar{\Gamma} = \frac{1}{N}\sum \arccos\left(\frac{B^{\text{Obs}} \cdot B^{\text{Pot}}}{\lvert B^{\text{Obs}}\rvert \lvert B^{\text{Pot}}\rvert}\right)$
SHRGT45	Fraction of Area with Shear 45°	Area with Shear 45°/Total Area
TOTPOT	Total photospheric magnetic free energy density	$\rho_{\text{tot}} \propto \sum \left(B^{\text{Obs}} - B^{\text{Pot}}\right)^2 dA$
MEANJZD	Mean vertical current density	$\bar{J}_z \propto \frac{1}{N}\sum \left(\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y}\right)$
USFLUX	Total unsigned flux	$\Phi = \sum \lvert B_z\rvert dA$
MEANGAM	Mean angle of field from radial	$\bar{\gamma} = \frac{1}{N}\sum \arctan\left(\frac{B_h}{B_z}\right)$
TOTUSJZ	Total unsigned vertical current	$J_{z_{\text{total}}} = \sum \lvert J_z\rvert dA$
ABSNJZH	Absolute value of the net current helicity	$H_{c_{\text{abs}}} \propto \lvert\sum B_z \cdot J_z\rvert$
$AREA_ACR$	Area of strong field pixels in the active region	Area = $\sum$ Pixels
R_VALUE	Sum of flux near polarity inversion line	$\Phi = \sum \lvert B_{\text{LOS}}\rvert dA$ within R mask
TOTUSJH	Total unsigned current helicity	$H_{c_{\text{total}}} \propto \sum \lvert B_z \cdot J_z\rvert$
SAVNCPP	Sum of the modulus of the net current per polarity	$J_{z_{\text{sum}}} \propto \left\lvert\sum^{B_z} J_z dA\right\rvert + \left\lvert\sum^{B_z} J_z dA\right\rvert$
MENAGBZ	Mean gradient of vertical field	$\lvert\nabla\bar{B}_z\rvert = \frac{1}{N}\sum\sqrt{\left(\frac{\partial B_z}{\partial x}\right)^2 + \left(\frac{\partial B_z}{\partial y}\right)^2}$

Fig 1. Parameters derived from SHARP Images.

We have also taken the time evolution of the feature as change information between the subsequent time lags as additional input to existing features.

Class imbalance in general is denoted when one class has more representation with respect to other. In majority of the cases, the objective is about the minority class detection. Similarly, here we are more interested in predicting whether flares will be accompanied by CMEs or not which are occasional events. Class imbalance should be addressed and performance metrics should be carefully chosen before assessing a model's performance. Machine Learning models shows elevated performance in the metrics such as accuracy due to better classification of the majority class. To overcome this bias, we had tried to address the class imbalance issues through various techniques such as under sampling majority class, oversampling minority class and Synthetic Minority oversampling Technique (SMOTE). We had compared the model's performance with respect to each sampling techniques.

Random oversampling simply duplicates randomly the datapoints of the minority class. This method will help in achieving balanced dataset when we have very small dataset with severe class imbalance. Here we cannot afford dropping the majority class. These methods are naive resampling of the existing data without assuming much about the data distribution. Repeating the datapoints may lead to reduced generalization of the model's performance in few cases where the dataset is very small.

SMOTE can be useful in such cases where it synthetically generates the minority samples based on the data points that are close in the feature space. Hence this may increase the generalization of the model's performance when new datapoints are introduced for prediction which falls under same feature space.

Rather than increasing or decreasing datapoints the model's performance for minority class can be increased by introducing class weights. This method is used to provide higher weights for the minority class which in turn signifies the importance while defining a model. This method is effective where the dataset is very small. We have utilized the above-mentioned sampling techniques to address the class imbalance issue, and compared to evaluate the effectiveness of each sampling technique in improving the model's performance.

2.1 Training methodology

Training of the ML model is carried out as follows. We analyze 8 ML model's performance that are Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Random Forest (RF), Decision Tree (DT), Gradient Boost (GB), Logistic Regression (LR), Adaboost, XGBoost. The dataset is split into kfolds using stratified kfold to ensure proper proportion of classes in training and test dataset. The test set has been carefully separated from training set before sampling process and it is not involved in any of the training. Fig 2 , represents the structure of the training we had carried out in this article.

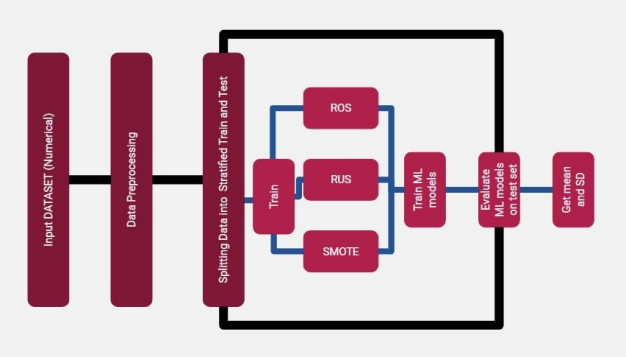

Fig 2. Overall training workflow of each model.

The sampling is done only to the training dataset. This sampled dataset is fed to the Machine learning models for training. From the dataset of kfold, the model is trained on k-1 folds which is sampled, and tested on the remaining one-fold which has been kept apart without sampling. The corresponding metrics are calculated for the respective fold. This is repeated iteratively, till the model has finished testing for all kfolds. At the end of kfolds, mean and standard deviation is calculated to set the error bar for the model. In this way, model's performance evaluation has been carried out.

3. Results and Discussions

3.1 Metrics

In this work, we have a binary classification problem, where the correct and misclassified binary predictions are recorded using confusion matrix. We consider True positive (TP) class as when flares are accompanied by CMEs in both prediction and actual. False Positive (FP) cases are when the actual class is flares without CMEs and predicted class is flares with CMEs. False Negative (FN) when the actual class is flares with CMEs and the model predicts wrongly as flares without CMEs. True Negative (TN) refers when both actual and prediction class are flares without CMEs. Based on the confusion matrix, the model's performance is evaluated based on the TSS metric which is not affected by dataset imbalance. We had tried in a similar way to [6] article, to check the performance of all models on different time lags 8h, 12h, 24h, 36h, 48h before the flare appears.

We have evaluated the performance of the model in two ways.

1) By utilizing only, the 18 SHARP features.

2) By adding the change information for each feature from the previous time lag.

In the first method, where we have utilized 18 SHARP features, we find that the performance of the SVM is good compared to other models. The corresponding figure is shown in Fig 3. We find that with SMOTE sampling, TSS performance metric is 0.84±0.12 with probability of detection rate of 86%, and less false alarm and false detection rate. LDA also performs good next to SVM with ROS and SMOTE sampling methods. LDA achieves TSS of 0.81 with PoD rate of 88%, False Alarm Rate (FAR) of 7% and FDR of 27%. Random forest and Decision trees perform worst with TSS of 0.35. The best performance of the models is obtained from 36h time lag.

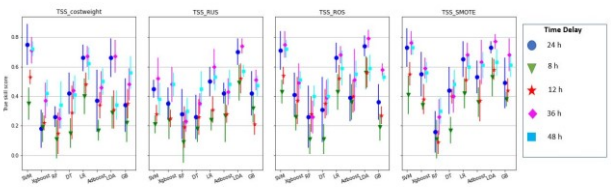

Fig 3. TSS score comparison with different ML models with 18 features for different sampling techniques.

The Probability of Detection (PoD) is more in case of LDA than SVM, which explains that the LDA is able to detect the occurrence of CME 88% of the time correctly whenever flare occurs. But comes with a cost of higher FAR. We have also trained the models without any sampling methods, and tried to get the model's performance. Without sampling, SVM's TSS performance decreases to 0.51 which shows that the results are biased towards the majority class.

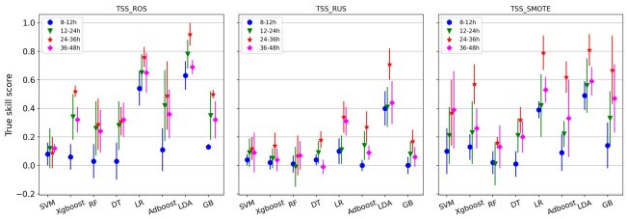

Fig 4. TSS score comparison with different ML models with change information as additional features for different sampling techniques

Hence either having higher cost weight for the minority class [6] or balancing it with oversampling minority class increases the TSS to 0.84. However, we can observe that the LDA's performance doesn't changes drastically with and without sampling. LDA performance is observed good only at time lag of 36h, and reduces its TSS performance to 0.5 for other successive time lags. SVM doesn't show this variation at different successive time lags. SVM's TSS performance consistently increases from 8h to 36h, and then decreases at time lag of 48h gradually. In both models, 36h performance shows that it can be optimum time lag to monitor the flare-CME associations.

For the next method, we tried to include the change information between the successive lags as additional features to the existing dataset and the respective results are shown in Fig 4. We trained the model in the similar way with different sampling methods and noted down the performance of each model. For SVM we had tuned hyperparameters gamma to 0.0088 from 0.075. LDA shows increased performance of TSS score upto 0.92±0.07 under ROS and SMOTE oversampling methods. LDA identifies the occurrence of CME perfectly on average of 96% of the time under ROS and SMOTE sampling method with very less FAR of 4%. SVM with SMOTE sampling performs maximum TSS score of 0.68±0.1.SVM's probability of detection is 74% for SMOTE with FAR of 6%. The above-mentioned performance metrics are when difference between 24h and 36h time lag are added as additional features.

Overall, we can see that LDA performs better under all sampling methods in par with svm, with lesser error bar. SVM performs better than all models with TSS score of 0.83 under Random oversampling method without including change information. Next to these two models, simple LR method performs better with TSS score around 0.80 in ROS and SMOTE method.

4. Conclusion

In this paper we tried to analyze the performance of different ML models on prediction of whether a flare will be associated with CMEs or not. We had tried to evaluate the model's performance by addressing the class imbalance issue. Also, we tried to check whether the model's performance increase when we include the change

information of the respective features. We achieved a better model performance by LDA than SVM using sampling techniques and change information. We find that SVM performs better with class weight assigned if the class imbalance issue is not addressed. However, we saw that performance of majority of the models were increased after the classes were balanced. This was achieved through oversampling minority classes randomly and through synthetically generating minority samples using datapoints that fall under close to same feature. We find that the LDA and SVM shows overall increased performance when the class imbalance is addressed. Since class imbalance is significant issue to be addressed while applying ML to real world scenarios, In this article we had tried to study the model's performance while trying to address the class imbalance issue using various sampling methods. We further tried to evaluate the ML methods performance, when additional features of change information are added to the existing dataset. We observe that the change information increases the model's performance from 0.81 to 0.92. Further we notice that the model's performance is better at time lag of 36 hours which indicates that this may be the optimum time lag to observe whether a CME will be associated with Flare or not. In future, we would like to generate the features using GAN, as the newly generated features will have more characteristics than SMOTE or ROS, and that would lead us in better generalization of the model.

5. Acknowledgements

The authors thankfully acknowledge the use of data courtesy of the SDO/AIA science teams for providing SHARP input data.

6. References

[1] Carolus J. Schrijver, Driving major solar flares and eruptions: A review, Advances in Space Research, Volume 43, Issue 5, 2009, Pages 739-755, ISSN 0273-1177, doi:10.1016/j.asr.2008.11.004.

[2] X.-L. Yan, Z.-Q. Qu, D.-F. Kong, Relationship between eruptions of active-region filaments and associated flares and coronal mass ejections, MNRAS Society, Volume 414, Issue 4, July 2011, Pages 2803–2811, doi:10.1111/j.1365-2966.2011.18336.x

[3] Andrews, M. A Search for CMEs Associated with Big Flares. Solar Physics 218, 261–279 (2003). doi:10.1023/B:SOLA.0000013039.69550.bf

[4] Qahwaji, R., Colak, T., Al-Omari, M. et al. Automated Prediction of CMEs Using Machine Learning of CME–Flare Associations. Sol Phys 248, 471–483 (2008). doi:10.1007/s11207-007-9108-1

[5] Yashiro, S., & Gopalswamy, N. (2008). Statistical relationship between solar flares and coronal mass ejections. Proceedings of the International Astronomical Union, 4(S257),233-243. doi:10.1017/S1743921309029342

[6] Bobra, M. G ; Ilonidis, S. Predicting Coronal Mass Ejections Using Machine Learning Methods. ApJ 821. doi:10.3847/0004-637x/821/2/127

[7] Pesnell, W.D., Thompson, B.J. & Chamberlin, P.C. The Solar Dynamics Observatory (SDO). Sol Phys **275**, 3–15 (2012). https://doi.org/10.1007/s11207-011-9841-3