doi:10.20396/revpibic.v0i0.id

XXVI Congresso de Iniciação Científica Unicamp
17 a 19 de outubro Campinas | Brasil

A Comparative Study on Hierarchical Clustering Techniques Applied to Solar Flare Forecasting.

Mirelle C. Bueno, Guilherme P. Coelho, Ana E. A. da Silva.

Abstract
Due to the harmful effects that high intensity solar flares may cause, several research groups are dedicated to the task of predicting this phenomenon. Given this scenario, the present project applied and compared hierarchical clustering techniques as a preprocessing step to solar flare forecasting, in order to verify whether this approach leads to improvements.

Key words:
Hierarchical Clustering; Solar Flare Forecasting; Space Weather.

Introduction

Solar activity plays a key role in the dynamics of the planet. Therefore, it is necessary to study and constantly monitor such phenomena. Among the various solar events, the so-called solar flares are particularly relevant, as they are responsible for several effects in the terrestrial Ionosphere. Solar flares occur when the magnetic energy accumulated in the solar atmosphere is abruptly released[1]. This phenomenon can cause a series of harmful effects, ranging from short circuits in power distribution systems to interruptions in satellites and telecommunications systems[2]. In this scenario, the present work developed a prediction tool for the occurrence of solar flares one day ahead, combining a hierarchical agglomerative clustering approach, known as AGNES[3], with the following predictors: k-NN, Naive Bayes and J48. Two variants of AGNES were evaluated, namely *Single* and *Complete-Linkage*.
This tool uses the groups identified by AGNES in the following way:

- For a new data sample to be classified, it is verified to which group the new sample of data belongs.
- The classifiers, specifically trained with data from that cluster, is used to classify this new sample.

Therefore, the groups generated by the clustering algorithm will be responsible for training the predictors. The main objective of this work was to evaluate if the pre-clustering improves the accuracy of the prediction of solar explosions. In addition, it was analyzed which of the methods (*Complete-linkage* and *Single-Linage*) bring more benefits to the forecast.

Results and Discussion

The obtained results indicated a great difference between *Complete* and *Single-Linkage* methods: the first approach led to clusters with a better equilibrium between the number of samples of each class (but with a dominance of samples from the negative class), while groups generated by *Single-Linkage* did not present a good distribution of the number of samples (and a prevalence of the positive class was observed in the groups).
Table 1 presents the results of the data classification step, obtained through the classifiers in three different situations: without pre-clustering of the training data, with pre-clustering done by *Single-Linkage* and with pre-clustering done by *Complete-Linkage*. The experiments were made according to a block cross-validation approach, with 5 iterations.
It can be seen in Table 1 that the highest values of F-measure were obtained by the k-NN classifier without the pre-grouping step. Only the Naïve Bayes classifier presented gains with pre-clustering, more specifically with the *Complete-Linkage* approach.

Table 1. Results of the data classification experiments, in bold the highest values of F-measure and accuracy.

Pre-clustering Method	***Accuracy***	***F-measure***	**Classifier**
None	**0.932**	**0.533**	k-NN
Complete-Linkage	0.884	0,229	
Single-Linkage	0.869	0.289	
None	0.817	0.279	Naïve Bayes
Complete-Linkage	**0.838**	**0.382**	
Single-Linkage	0.695	0.305	
None	0.346	**0.346**	J48
Complete-Linkage	**0.361**	0.303	
Single-Linkage	0.352	0.309	

Conclusions

The obtained results allowed us to conclude that pre-clustering may lead to gains only with specific classifiers. However, the best classification results were obtained by the k-NN classifier, without the pre-clustering step. Therefore, it is up to the user to evaluate whether pre-clustering should be used.

[1] HOLMAN, G. D. *Solar Flare Theory*.Hesperia. 2007. Disponível em: http://hesperia.gsfc.nasa.gov/sftheory/index.htm Acesso em: 9 Dez. 2016
[2] LIMA, S. D. S. *Geomagnetic Storms Origin and consequence*. Dissertation (Graduation in Physics) - Science and Technology Center. Ceará: State University of Ceará, 2012.
[3] KAUFMAN, L.; ROUSSEEUW, J. P. *Finding Groups in Data* **:** *An Introduction To Cluster Analysis*. 1. ed. Canada: Wiley-Interscience,1990.p.199-275.