PROCEEDINGS OF SCIENCE

Search for and Classification of Short Transient Gamma-ray Events from INTEGRAL

Martin Topinka[*†]
Czech Technical University, Technická 2, Prague 6, 162 00, Czech Republic
E-mail: martin.topinka@gmail.com

A search through the INTEGRAL IBIS/ISGRI data archive revealed a large number of untriggered flares on a millisecond time scale. As the IBIS detector uses a coded mask, localisation of those flares is hindered by low count statistics in the deconvolution process. Here, selected flares are used as input for further data processing. Machine learning clustering algorithms and a search for class outliers can separate short gamma-ray bursts from soft gamma-ray repeater flares in the studied sample. Prompt discriminating between these two classes allows for the adjustment of strategies for real-time follow-up observations.

XI Multifrequency Behaviour of High Energy Cosmic Sources Workshop
25-30 May 2015
Palermo, Italy

[*]Speaker.

[†]This publication was supported by the European social fund within the framework of realising the project "Support of inter-sectoral mobility and quality enhancement of research teams at Czech Technical University in Prague", CZ.1.07/2.3.00/30.0034. Period of the project's realisation 1.12.2012–30.9.2015.

PoS(MULTIF15)064

1. Introduction

The IBIS/ISGRI γ-ray imager (20 keV $-$ 1 MeV) [1] on-board the ESA INTEGRAL satellite [2] has observed many short transients, including short gamma-ray bursts (sGRBs) [3] and soft gamma-ray repeaters (SGRs) [4] flares as they were triggered by the on-board burst alert system (IBAS) [5]. The standard data analysis pipeline `OSA` that performs deconvolution with the IBIS/ISGRI coded mask pattern has issues in reconstructing an image and subsequently a light curve for time intervals $\lesssim 1$ s because of the low count statistics per pixel. Therefore, an alternative approach, sensitive to the sudden short time excesses in the count rate above the real-time background on the detector in a Science Window[1] (ScW) has been used to detect short flare candidates [6].

2. Observational Data

The INTEGRAL observations with the total exposure time $\gtrsim 64$ Ms ($\gtrsim$ Terabytes of data), obtained in the period from Oct 2002 to Nov 2013, were processed. The search in the energy range $15-150$ keV revealed $\gtrsim 4\times10^4$ flare candidates with the count rate at least 10σ above the background on 10 ms time-scale, having passed hot pixel, cosmic ray and solar flare rejection filters. Due to the missing positional and detailed spectral information, the origin of these flares is mostly unknown. On the contrary, the nature of flares is well-assigned for the sGRBs that triggered IBAS and also for known active flaring periods of some SGRs. There are 46 SGR flares associated with SGR J1550-5418 (in October 2008 and first half of 2009) [7] and SGR 1806 [8, 9] and 4 sGRBs[2] (GRB 070707, GRB 071017, GRB 081226B, GRB 110112A) for which the data are publicly available and for which the light curve is long enough to have structure (> 30 ms) that can be used for the analysis. For each selected flare, set of representative attributes characterising the spectral, temporal and contextual information about the flares were recorded. The attributes, often called features, serve as the input data for the analysis. They are summarised in Table 1.

3. Data Analysis

3.1 Machine Learning

In the analysis described here, it is assumed that the found flares are either sGRBs or SGRs. The task is to distinguish between these two classes of objects when we observe a flare. The sGRB and SGR sources are of different origin and different physical processes may play dominant role during the flare emission. Therefore, a pattern to discriminate sGRBs from SGRs in the quick look analysis may likely exist, though it is not known. The advantage of applying machine learning is that distinguishing between the source classes in a reasonable manner is possible even without the knowledge what the true pattern is. Typically, when the class labels (sGRB or SGR in this case) of a good number of the observed flares are known, so called supervised learning is applied to train the classification algorithm on known examples. However, the small number of known

[1] The INTEGRAL observations are composed of observational blocks called Science Windows (ScWs), each of a typical duration ~ 30 minutes.

[2] http://ibas.iasf-milano.inaf.it/IBAS_Results.html

Table 1: List of features recorded for each identified flare. For certain quantities, the subscript x is an abbreviation for three energy ranges, $x \in 15-150$, $15-50$ and $50-200$ keV bands.

Feature name	Description
Spectral properties	
max_sig	maximal significance in σ above background in the flare
fluence_x	total fluence (in counts) in the flare in $15-150$ keV
h2s	hard to soft ratio between 15-50 keV and 50-200 keV
spectral lag	shift of the maxima between $50-200$ keV and $15-50$ keV
Light curve	
duration	total duration of the flare in $15-150$ keV
variability_x	normalised variance per pixel to average variance per pixel in a given ScW
bg_jump	the count difference between the first and last point of the flare
$p_{\sigma,x}$	the relative fraction of the points above 10σ detection threshold
Pulse shape	
asymmetry	rise time to decay time ratio
half-symmetry	mirror difference folded around midpoint
mad_x	median absolute deviation
std_x, skew_x, kurtosis_x	higher moments
Contextual	
Time	Date and time of the pulse start
RA	Right Ascension of the current pointing
decl	Declination of the current pointing
# neighbours	# of pulses with the signif. $\geq 10\sigma$ above the background within the same ScW
bg_ mean	average background level in the ScW
bg_ var	background variations in the ScW
Galactic	localised within $\pm 15^\circ$ from the Galactic plane
GTI flag	Flare occurred within GTI

sGRBs in the entire studied sample (4) would make the supervised learning strongly biased and inefficient. The highly unbalanced class abundance would result in high 42/46 = 91% overall classification accuracy if the majority class of SGRs were predicted primitively in all cases. This approach would fail to find any sGRB. Therefore, unsupervised learning, for which labels are not known or ignored, is used to detect clusters in the N-dimensional vector space in which each flare is represented by a vector built from the flare features. Thus, each dimension corresponds to one of the features (e.g. duration or variability etc.) The Euclidean distances between the points in this space are calculated and the techniques of K-mean and spectral clustering algorithms are used to detect clusters of points, assuming that nearby points (representing flares) would likely be of similar origin. Besides it, an additional test in which sGRBs are considered to be outliers from the SGR class, is performed [10, 11].

3.2 Dimensionality Reduction

Besides the high CPU demands, the initial high dimensionality of the vector space of the flare

features can confuse the algorithm by noisy dimensions – the dimensions that represent features with no or very low relevancy for the classification itself. Moreover, the aggregating clustering methods can suffer from the "curse of dimensionality" degeneracy in d-dimensional space where $d \gtrsim 10$.

Much of the contextual information is not relevant to the classification purpose. Therefore, only the logarithm of the number of flares in the neighbourhood and boolean flag whether a flare is close to the Galactic plane are kept. To smear off the importance of these two contextual features further down, the distance in these two dimensions is suppress by factor of 10 when calculating the metrics. To remove unwanted redundancy, only one feature in any highly correlated pair of features (by means of the Pearson correlation coefficient $|r| > 0.3$) is kept.

The principle component analysis (PCA) projection is often efficiently used to lower the dimensionality of the problem. However, PCA detects a linear combination of features with the largest variance, rather than emphasises the feature of the highest importance for the classification task. Therefore, the supervised machine learning algorithm of Random Forests is used here to score the features importance based of their relevance to the classification. While the simple decision tree algorithm recursively divides the vector space to achieve the best possible class splitting, Random forest is a Monte Carlo ensemble algorithm that tackles the over-fitting problem that is often encounter in the classical decision tree classifier. In the Random forest algorithm, a large number of randomly perturbed decision trees are generated with one of the features missing. The result is then averaged over many tosses. Higher classification accuracy with the feature missing statistically implies lower significance of the omitted feature. The results of the feature importance are shown in Fig. 1. The first 8 most significant features are used for the clustering analysis. This covers $\sim 70\%$ of the overall information in the input dataset.

3.3 K-Mean Clustering

K-mean clustering is a simple but powerful algorithm. The iterative algorithm is described in the caption of Fig. 2. To unify the scales in each dimension, the input data are normalised to have zero mean and unit standard deviation.

The number of clusters must be fixed a priori. Two clusters are expected by the nature of the problem. The best number of underlying clusters can be seen through the compactness of the clusters, defined by the Silhouette score metrics

$$s \equiv \frac{b-a}{max(a,b)}, \tag{3.1}$$

where a are refers to the mean distances between a sample data point and all other points in the same class, while b marks the same with respect to the next nearest cluster. The Silhouette score as a function of number of clusters is shown in Fig. 3. To mitigate the algorithm sensitivity to the choice of the initial cluster centres, the mean result from a large number of random Monte Carlo initial setups is used.

3.4 Spectral Clustering

Unlike the K-mean clustering, the spectral clustering joins data points based on the graph connectivity rather than a distance to a cluster centre, overriding the K-means clustering limitation

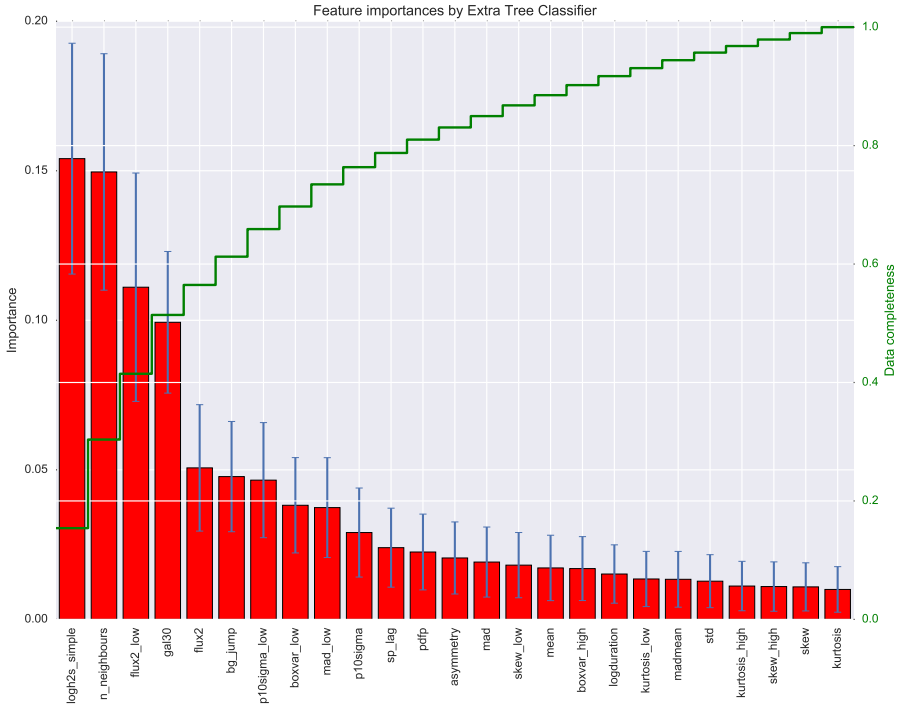

Figure 1: Relative feature importance given by the simulation of 1000 realisations of random forests. Error-bars denote the fluctuation given by the Monte Carlo process. The green line plots the completeness of the dataset reconstructed only with the most significant features (the fraction of information kept compared to the initial dataset).

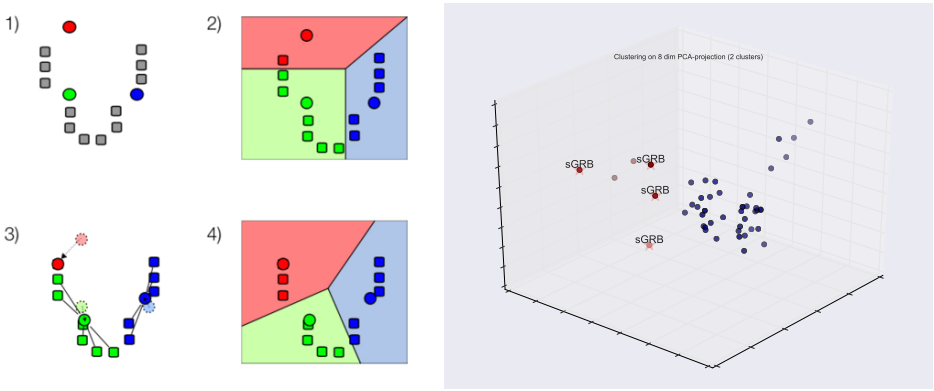

Figure 2: *Left:* The scheme of the K-mean clustering algorithm. 1) The initial cluster centres are set randomly. 2) The points are assigned to belong to the nearest cluster centre. 3) The cluster centres are then recalculated as centre of masses of the assigned data points. 4) Steps 2 and 3 are repeated until a convergence has been achieved. *Right:* The result from the K-mean clustering for 2 clusters. The true GRBs are annotated. The PCA projection into 3 dimensions is used for visualisation purposes only, not for the data pre-processing, the clustering happens in the 8-dimensional vector space.

Figure 3: The silhouette score measures the compactness of the identified clusters. The score is plot as a function of the number of clusters. The highest values for $k = 2$ and $k = 3$ suggests that the optimal number of clusters is 2 or 3.

to the d-dimensional spheroid shape of the clusters. The connections between the data points x_i are defined by the affinity matrix $A_{ij} \simeq \exp\left(-\alpha \parallel x_i - x_j \parallel^2\right)$ where α is a constant. Alternatively, with a threshold applied $A_{ij} = 0$, if $\parallel x_i - x_j \parallel^2 \geq R$, where R is the hard cut off. Then, the graph Laplacian is constructed $L = D - A$, where D is the diagonal degree matrix measuring the degree of each node. In this dual graph representation each cluster will occupy a diagonal block. Eigenvalues of the system $Lv = \lambda v$ are calculated. The k eigenvectors corresponding to the k lowest eigenvalues define k-dim subspace in which clusters can be found by using the standard K-means clustering technique.

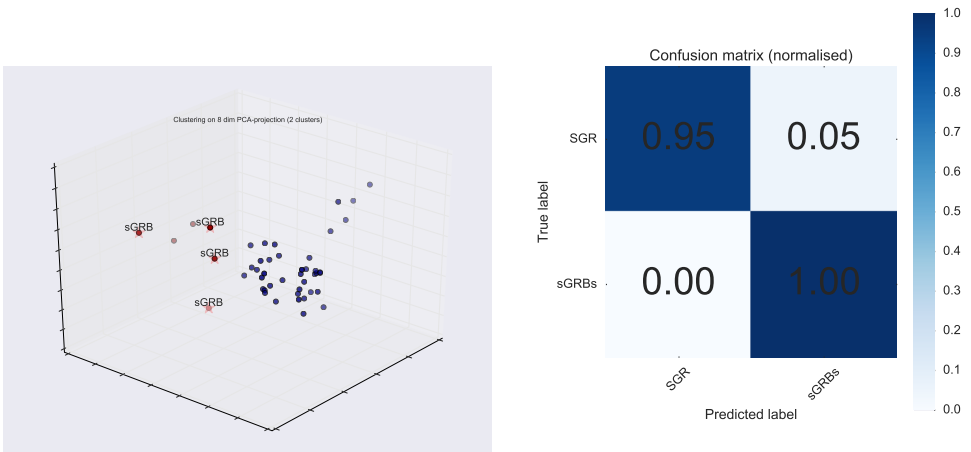

Figure 4: *Left:* The clusters identified by the spectral clustering algorithm. The true GRBs are annotated. The PCA projection is again used for visualisation purposes only, not for the data preprocessing. *Right:* The normalised confusion matrix of the spectral clustering algorithm. The numbers depict the actual fraction of true/false positive/negative associations in each class when compared to the true labels.

3.5 Outliers

In the classification task, rare events such as sGRBs can be viewed as outliers in the SGRs class. The outlier score S_{out}, defined as the relative distance to a data point from the centre identified by K-mean clustering, is calculated and normalised to the mean distance in the sample

$$S_{out} \equiv \frac{distance(x, centroid_C)}{median\left(\forall_{x \in C} distance(x, centroid_C)\right)} \tag{3.2}$$

Data points further than 3σ threshold from the SGR cluster centre are marked as outliers.

Alternatively, the traditional support vector machine (SVM) algorithm can be modified to detect outliers. The SVM algorithm tries to set the separation hyperplane between two classes of data points in the vector space of the features maximising the margin from the separation surface. In the extreme case, one class can consist a single test point. The relative size of the margin around this point determines whether or not the data point is considered as an outlier.

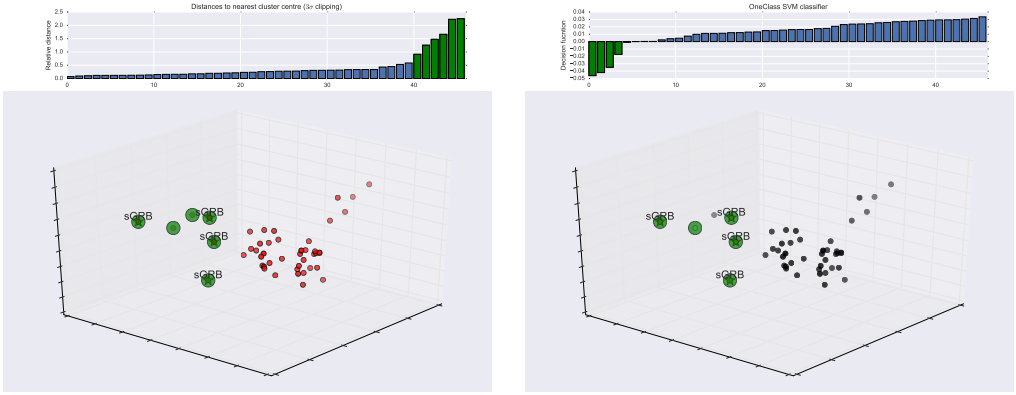

Figure 5: *Top left:* Outlier score measured as distance from the SGR flare cluster centre. The data points above 3σ threshold are marked green. Bottom left: The PCA projection of the data points with the outliers with the respect to the distance from the cluster centre are marked in green. Short GRBs are annotated by a star. *Top right:* Outlier score as a logarithm of the inverse of the relative size of the SVM margin. Negative values mean further outliers. *Bottom right:* The PCA projection of the data points with the outliers identified by the single class SVM algorithm are marked green. Short GRBs are annotated by a star.

4. Conclusions

The most relevant features to perform the classification are hard-to-soft ratio, Galactic position flag, intrinsic variations, loneliness (number of flares in the vicinity) and the flux in counts in the low energy band. However, none of the features exceeds 15% of importance individually. Based on the Silhouette score, the 8-dim feature space representation suggests 3 clusters (Fig. 3). Cross-check with the true labels reveals the group of sGRBs, SGRs and the branch of some SGR giant flares in the vector space, clearly distinguishable from the sGRB group (Fig. 2 and Fig. 4). Assigning the cluster with majority of true SGRs in it to the SGRs cluster and the same for sGRBs, the overall classification accuracy reaches 96%. A similar result is achieved from the outlier analysis. This gives a power to discriminate in first order the sGRB from the SGRs without the need for detailed spectral information. None of the 4 sGRBs was misclassified as an SGR flare. There are 1-2

isolated peculiar SGR flares that occupy similar region in the parametric space as sGRBs. The very marginal case of the sGRB subclass, the least distant outlier from the SGR class, is GRB 071017 which has been an unusually soft sGRB [12, 13].

Although, the machine learning concept is general, the clustering may be unique for the given instrument and the process used to collect the data. The presented method can be further tested on new sGRBs and SGR flares detected by INTEGRAL or be modified to be applied on a dataset from a different mission.

References

[1] F. Lebrun, J.P. Leray, P. Lavocat et al., *ISGRI: The INTEGRAL Soft Gamma-Ray Imager*, *A&A*, **411**, L141 (2003)

[2] C. Winkler et al., *The INTEGRAL mission*, *A&A*, **411**, L1–L6 (2003)

[3] E. Nakar, *Short-hard gamma-ray bursts*, *Physics Reports*, **442**, Issue 1-6:166-236 (2007)

[4] S. Mereghetti, *The strongest cosmic magnets: soft gamma-ray repeaters and anomalous X-ray pulsars*, *A&A Reviews*, **15**, 255-287 (2008)

[5] S. Mereghetti, D. Gotz, J. Borkowski et al., *The INTEGRAL Burst Alert System*, *A&A*, **411**, L291 (2003)

[6] M. Topinka, *Studies of Cosmic Gamma-Ray Bursts and a Search for Other Transient Events Detected by the INTEGRAL*, PhD thesis (2011)

[7] Mereghetti, S., Gotz, D., Weidenspointer, et al., *Strong Bursts from the Anomalous X-Ray Pulsar 1E 1547.0-5408 Observed with the INTEGRAL/SPI Anti-Coincidence Shield*, *ApJ*, **696**, L74 (2009)

[8] Hurley E.P, et al., *An exceptionally bright flare from SGR 1806-20 and the origins of short-duration γ-ray bursts*, *Nature*, **434**, 1098–2005 (2005)

[9] K. Hurley, *Soft gamma repeaters*, *Advances in Space Research*, **47**, 1337 (2011)

[10] F. Pedregosa, G. Varoquaux, A. Gramfort et al., *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research, **12**, 2825 (2011)

[11] Ž. Ivezic, A. J. Connolly, J. T. VanderPlas and A. Gray, *Statistics, Data Mining and Machine Learning in Astronomy*, Princeton University Press (2014)

[12] M. Topinka, A. Martin-Carrillo, S. Meehan, L. Hanlon and B. McBreen, *New outbursts from GRB 071017 and 1E1547.0-5408 discovered in an automated search for SGR-like events in the INTEGRAL archive*, 8th INTEGRAL Workshop. The Restless Gamma-ray Universe, *104* (2010)

[13] S. Mereghetti, P. Esposito, A. Tiengo et al., *The magnetar candidate AX J1818.8-1559*, *A&A*, **546**, A30 (2012)