
Article:
Boongoen, T., Iam-On, N. and Mullaney, J. orcid.org/0000-0002-3126-6712 (2022) Providing contexts for classification of transients in a wide-area sky survey: an application of noise-induced cluster ensemble. Journal of King Saud University - Computer and Information Sciences, 34 (8). pp. 5007-5019. ISSN 1319-1578

https://doi.org/10.1016/j.jksuci.2021.06.019

Journal of King Saud University – Computer and Information Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Providing contexts for classification of transients in a wide-area sky survey: An application of noise-induced cluster ensemble

Tossapon Boongoen [a], Natthakan Iam-On [a,*], James Mullaney [b]

[a] Center of Excellence in AI and Emerging Technologies, School of Information Technology, Mae Fah Luang University, Tasud, Muang District, Chiang Rai 57100, Thailand
[b] Department of Physics and Astronomy, University of Sheffield, UK

ARTICLE INFO

Article history:
Received 19 April 2021
Revised 16 June 2021
Accepted 23 June 2021
Available online xxxx

Keywords:
Astronomical data
Analytical method
Machine learning
Imbalance classification
Cluster ensemble

ABSTRACT

With new sensor systems that capture sky survey at high quality level, analyzing the resulting data within a limited time frame appears to be the next challenge. Specific to the GOTO project, this task proves to be crucial to discover new transients from a pool of large candidates. Initial works based on the feature-based approach design this detection as imbalance classification, where a data-level method can be used to resolve the difference in cardinality between classes. This paper presents a context generation framework to complement the previously proposed model. In particular, samples are clustered to form data contexts to which different learning strategies may be applied. To ensure the quality of data clustering, a noise-induced cluster ensemble technique that has been recently introduced in the literature is employed here. The results with simulated data and algorithms of NB, C4.5 and KNN have shown that the proposed framework can filter out some negative samples quickly, while making classification of the rest more effective. In particular, it enhances predictive performance of basic classifiers by lifting F1 scores from less than 0.1 to around 0.3–0.5. Besides, parameter analysis is also given as a guideline for its application.

1. Introduction

Sensor systems have been a major factor to various breakthroughs in science and engineering, especially in the present era of distributed system and big data. Likewise, for the field of astronomy or space science, the development of detectors as well as optical technology provides an integral contribution that leads to the discovery of major incidents. As such, acquiring data from the deep space is no longer an intricate problem. This makes the study of astronomy more interesting since the detection of transient astronomical events, simply termed "transients" by astronomers (de Buisson et al., 2015; Soraisam et al., 2018), can be managed timely. The phenomenon led to the interdisciplinary collaboration among related area of study to help astronomers investigate and evaluate the massive amount of digital information received from sky surveys, most recent, from the advanced Laser Interferometer Gravitational-wave Observatory (aLIGO, Meisner et al., 2017). The discovery of transient events is crucial, for it eventually leads to the study of rare classes of extreme events, such as, neutron stars and black holes, the tidal disruption of stars by dormant super massive black holes, or megaflares on normal main sequence stars (Wette, 2021).

The Gravitational-wave Optical Transient Observer (GOTO)[1] is among the new class of telescopes dedicated to detecting such phenomenon. Its main role in the collaborative study is to provide a visual counterpart of the detected transient events (Dyer et al., 2018). It is devoted to deliver graphic information of any event in space as noticed by gravitational observatories. GOTO is an international collaboration led by University of Warwick of UK and Monash University of Australia, with its facility housed at Roque de Los Muchachos observatory on La Palma, Canary Island. The observatory consists of an array of four state-of-the-art 0.5 m- aperture, wide-field optical telescopes which can respond to alerts coming from gravitational wave detectors, i.e., LIGO and VIRGO (Abbott et al., 2020). Basically, transient event is short-lived and may diminish in

* Corresponding author.
E-mail addresses: tossapon.boo@mfu.ac.th (T. Boongoen), natthakan@mfu.ac.th (N. Iam-On), j.mullaney@sheffield.ac.uk (J. Mullaney).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

[1] https://goto-observatory.org.

https://doi.org/10.1016/j.jksuci.2021.06.019
1319-1578/

hours or a day. This speedy transitory necessitates a rapid response in order to track down optical images. This is what GOTO has been designed for; to detect these optical signatures as quickly as possible so as to provide astronomers with as much information about these sources before they ultimately fade. Specific to GOTO observation and common to several other wide-area sky surveys, a considerable amount of images (currently around 400 of them) is collected nightly. Each of the images has approximately 20,000 astronomical sources, which results in the total number of sources just under 8 millions. These observed sources are subtracted from a so-called "reference" images, in which known sources are well studied and documented. As a result, the difference images contain a collection of roughly 40,000 to 50,000 new sources, in which a few correspond to transient events. They are then processed along the pipeline that aims to deliver readable data for further investigation. One challenge encountered by the GOTO team is to figure out a data management system efficient for the big volume of data. The other is a race with time, where an automated decision-support tool is required to identify a set of transients from those in the difference images. After that, manual observations by GOTO and collaborations can be conducted to confirm the events. This demands an interdisciplinary research where machine learning has an integral role in translating raw data into meaningful knowledge.

To initiate such a study, recent works by Tabacolde et al. (2018,) have designed the task of transient detection as a binary classification of candidate sources. In particular to the former, an oversampling technique is employed to handle the imbalance problem between the two classes (real and bogus). It is noteworthy that the proportion of minority class is less than 0.5%, much lower than those datasets investigated in both astronomy (Cabrera-Vives et al., 2017; Wright et al., 2017) and machine learning literature (Tang and He, 2017; Ofek et al., 2017). Despite the reported improvement, this data-level method has been known to promote overfitting (Lin et al., 2018). Hence, Tabacolde et al. (2018) introduce undersampling as an alternative to the previous, where the proposed clustering based model performs better than the conventional technique of RUS (Random UnderSampling, Bagui and Li, 2021; Seiffert et al., 2010). Given the findings, this paper aims to complement the previous attempts to solve the imbalance problem by producing distinct contexts for classification model development, instead of consider the whole data as one indivisible set. To achieve this goal, the recently published study of noise induced cluster ensemble (Panwong et al., 2018) is exploited to generate high-quality data clusters, which exhibit different contexts for supervised learning. In fact, the deployment of the resulting classifiers may be efficient as some clusters dedicates solely to one class, i.e., a complex classifier can be replaced by a simple rule.

Problem and Scope. The research reported within this paper aims to improve the accuracy of a classification model built to categorize object candidates as either a real source to investigate further or a bogus to simply ignore. In particular, these candidates are determined by the image differencing process in which a nightly image is subtracted from a reference, i.e., a co-added image. As such, new groups of bright pixels that have not been recorded thus far can be identified. In fact, their thumbnail images of size 21×21 pixels are extracted to form a pool of candidate images. Note that the above mentioned process is executed within the data processing pipeline of GOTO, which has been modified from that of the LSST (Large Synoptic Survey Telescope) project (Mullaney et al., 2020). Prior developing a binary classification model to differentiate between true source and bogus, an expert-driven set of features are extracted to deliver the target dataset. As an initial study before deploying within the GOTO pipeline stack, datasets are simulated to provide a realistic testbed for the proposed method against existing ones. This provides a chance to inject a rich collection of transients into actual sky images, where this is hardly obtained from any single observation. Provided this setting, it is assumed that the resulting model can be robust to various types and appearances of transient events, thus becoming sufficient to deploy in an actual pipeline in the next phase.

Contributions. The contributions made by the work presented in this paper can be summarized as follows.

- This paper presents a new framework to handle an imbalance classification problem through generation of contexts for learning model development. It makes use of noise-induced cluster ensemble to determine a clustering reference from which those contexts can be formulated. This organic combination has not been witnessed in the literature thus far, especially for astronomical data analysis.
- It reports an original set of experimental results on simulated datasets, which have been created based on the system configuration of GOTO project and generalized to other sky survey platforms. Hence, the paper provides useful findings to a wide community of astronomers and data scientists working on classification problem as a means to detect transient events. Parameter analysis is also included as guideline for the future application of this new framework.

The rest of this paper is organized as follows. Section 2 presents background and materials employed in this study including details of investigated data, feature extraction and data preparation. The proposed method is described in Section 3, in which the development of both cluster ensemble and classification contexts are emphasized. After that, Section 4 provides the results of this investigation and related discussions. The paper is concluded in Section 5 with possible future works.

2. Background and materials

This section presents background and details of the investigated data sets, which have been simulated to reflect the core characteristics of sources captured by the GOTO system. It provides a good testbed from which initial classification models can be derived and later refined with real data.

2.1. Background and investigated data

There are software packages with a capability to synthesize astronomical images, including approximations for commonly encountered complications, e.g., background noise and the point spread functions (PSF) of sources. Specific to the current research, SkyMaker (Kauffmann et al., 2020) is employed to create the simulated images. It accepts a list of sources (i.e., stars, galaxies) containing the position (i.e., right ascension, RA, and declination, Dec) and brightness of each source. These three pieces of information is all that is required for stars (i.e, point sources). Galaxies in SkyMaker are represented by two cospatial ellipses (one for the bulge, the other for the disk), which are described by an additional seven parameters (the ratio of bulge-to-total light, bulge radius, bulge aspect ratio, bulge orientation on sky, disk radius, disk inclination and disk orientation on sky). For the simulations, a source lists is produced by querying two separate databases. For stars brighter than 17th magnitude, the USNO CCD Astrograph Catalog (UCAC) database is exploited, whereas for stars and galaxies fainter than 17th magnitude, the Sloan Digital Sky Survey (SDSS) is employed instead. This approach of combining two separate catalogues to generate our input lists was used to increase the dynamic range of our simulated images since bright stars saturate the SDSS detector and are thus under-represented in this catalogue, while UCAC does not go sufficiently deep for our purposes.

These databases are queried for all sources that would be covered by a single observation by one of the GOTO telescopes of a given patch of sky (given by its central coordinates). Each source's position (RA, Dec) was converted into a pixel coordinate (x, y) by referring to the on-sky position of the central pixel and the pixel scale (i.e., the on-sky angular size of each pixel, which is a constant 1.24 arcseconds per pixel). For the UCAC sources, the V-band magnitude was used, whereas G-band magnitudes were used for SDSS sources. At this stage, galaxies are simulated as a simple disk (i.e., not a bulge + disk combination), and thus only provide the three additional parameters that SkyMaker uses to simulate galaxy disks. With the primary goal of detecting transient sources, it is not expected that only simulating disks will have any significant impact on this study. In addition to the source list, SkyMaker also requires an input configuration file. This provides the software with information such as the type of simulation that is required (e.g., include background noise or not) and the characteristics of the telescope. For the latter, the most important of these are the saturation level of the pixels (set to 65,535), the zeropoint of the telescope (i.e., the magnitude of a star that would result in one count per second; 23.5), PSF size (see below), pixel size (1.24 arcsec per pixel), CCD size in pixels (8176×6132). To increase the realism of the simulations, the PSF size (i.e., full-width half maximum, or FWHM) is allowed to vary randomly between observations, ranging from 0.8 to 3 arcseconds.

To simulate transient sources, two observations for each patch of sky are synthesized. In the second observation, new sources are injected, randomly distributed across simulated image with brightnesses chosen randomly from a uniform distribution ranging from magnitude 14 to 19. Each simulated image is then processed using the LSST software stack (Juric, 2015; Mullaney et al., 2020), adapted to handle simulated images. Then, the output from the image differencing component of the stack is delivered as input to the data collection phase, prior data transformation and model development. Fig. 1 illustrates examples of detected sources that can be categorized as bogus (Class0) and real (Class1). These are presented as greyscale images of size 21×21 pixels.

2.2. Data preparation

Based on the common astronomical measurements made to each bright source, a data set $X = \{x_1, \ldots, x_N\}$ of N samples is characterized by 23 different attributes, $F = \{f_1, \ldots, f_{23}\}$. Note that some initial features (i.e., id, parent_id, RA, DEC, SdssCentroid_x, SdssCentroid_y) are excluded at first as they are not informative. Each instance can be defined as $x_i = \{x_{i,1}, \ldots, x_{i,23}, x_{i,c}\}$, where $x_{i,j}$ is the value of attribute $f_j \in F$ and $x_{i,c} \in \{1, 0\}$ denotes the class label. Table 1 summarizes these features in terms of their notations and descriptions. Given the data X, correlations between a feature and the two classes are investigated, together with the initial exploitation with simple classifiers. It turns out to be the case that several of these features are not informative such that the resulting classification performance tends to be inadequate. As such, the additional stage of data transformation has been designed in order to compile the existing set of features to a more discriminative one. This process is guided by domain experts that leads to the preprocessed data with the final set of 15 features, each of which is explained next. Also, see Fig. 2 for the graphical summarization. In the experiment set forth for this study, two datasets are generated, each of which goes through the sequence of data collection and preparation previously specified. See details in Section 4.

(1) PSF_Flux_Sig:

$$\frac{PSF\_flux}{PSF\_flux\_Sigma} \tag{1}$$

(2) PSF_Dipole_Flux_Pos_Diff:

$$\sqrt{D_x^2 + D_y^2}, \tag{2}$$

where D_x and D_y denote $PSF\_Dipole\_Flux\_Pos\_x - PSF\_Dipole\_Flux\_Neg\_x$ and $PSF\_Dipole\_Flux\_Pos\_y - PSF\_Dipole\_Flux\_Neg\_y$.

(3) PSF_Dipole_Flux_Pos_Sig:

$$\frac{PSF\_Dipole\_Flux\_Pos}{PSF\_Dipole\_Flux\_Pos\_Sigma} \tag{3}$$

(4) PSF_Dipole_Flux_Diff:

$$PSF\_Dipole\_Flux\_Pos - PSF\_Dipole\_Flux\_Neg \tag{4}$$

(5) PSF_Dipole_Flux_Rel:

$$\frac{PSF\_Dipole\_Flux\_Diff}{PSF\_flux} \tag{5}$$

(6) PSF_Dipole_Flux_Neg_Sig:

$$\frac{PSF\_Dipole\_Flux\_Neg}{PSF\_Dipole\_Flux\_Sigma} \tag{6}$$

(7) PSF_Dipole_Flux_x: remains *unchanged* from the original attribute set.
(8) PSF_Dipole_Flux_y: remains *unchanged* from the original attribute set.
(9) DipoleFit_Flux_Pos_Diff:

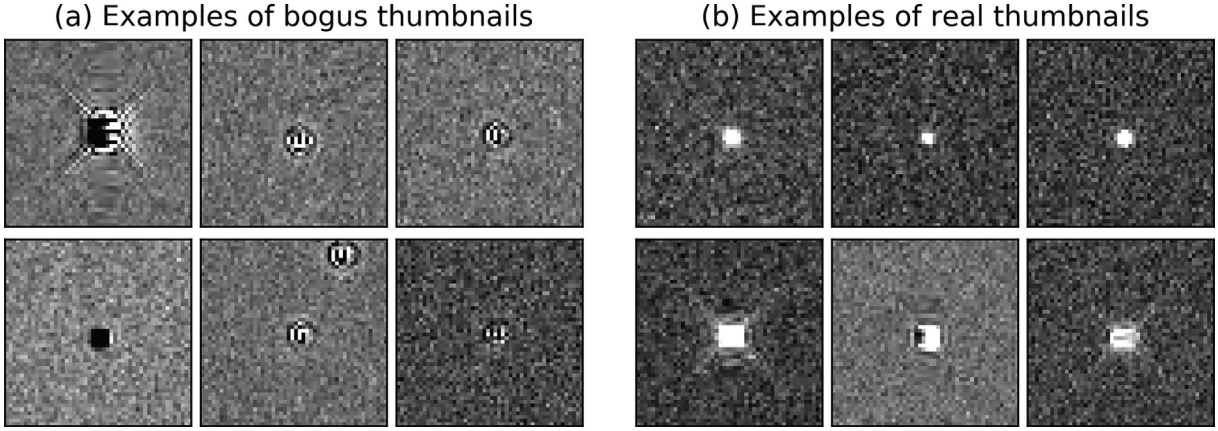

Fig. 1. Image examples of bogus and real sources, taken from Tabacolde et al. (2018).

Table 1
Details of the original set of features: notation and description.

Notation	Description
PSF_flux	Measure of brightness of source within Point Spread Function (PSF)
PSF_flux_Sigma	Measure of uncertainty associated with PSF_flux
PSF_Dipole_Flux_Pos_x	Position in X-dimension of positive part of dipole
PSF_Dipole_Flux_Pos_y	Position in Y-dimension of positive part of dipole
PSF_Dipole_Flux_Pos	Brightness of positive part of dipole
PSF_Dipole_Flux_Pos_Sigma	Uncertainty associated with measurement of PSF_Dipole_Flux_Pos
PSF_Dipole_Flux_Neg_x	Position in X-dimension of negative part of dipole
PSF_Dipole_Flux_Neg_y	Position in Y-dimension of negative part of dipole
PSF_Dipole_Flux_Neg	Brightness of negative part of dipole
PSF_Dipole_Flux_Neg_Sigma	Uncertainty associated with measurement of PSF_Dipole_Flux_Neg
PSF_Dipole_Flux_x	Average X-dimension position of dipole (i.e., average of negative and positive positions)
PSF_Dipole_Flux_y	Average Y-dimension position of dipole (i.e., average of negative and positive positions)
DipoleFit_Flux_Pos_x	Position in X-dimension of positive part of dipole fit
DipoleFit_Flux_Pos_y	Position in Y-dimension of positive part of dipole fit
DipoleFit_Flux_Pos	Brightness of positive part of dipole fit
DipoleFit_Flux_Pos_Sigma	Uncertainty associated with measurement of DipoleFit_Flux_Pos
DipoleFit_Flux_Neg_x	Position in X-dimension of negative part of dipole fit
DipoleFit_Flux_Neg_y	Position in Y-dimension of negative part of dipole fit
DipoleFit_Flux_Neg	Brightness of negative part of dipole fit
DipoleFit_Flux_Neg_Sigma	Uncertainty associated with measurement of DipoleFit_Flux_Neg
DipoleFit_Flux_x	Average X-dimension position of dipole fit (i.e., average of negative and positive positions)
DipoleFit_Flux_y	Average Y-dimension position of dipole fit (i.e., average of negative and positive positions)
DipoleFit_Flux	Measure of overall brightness in dipole fit - magnitude (i.e., ignoring whether it is positive or negative) of entire dipole

Fig. 2. Details of data transformation, taken from Tabacolde et al. (2018).

$$\sqrt{C_x^2 + C_y^2}, \tag{7}$$

provided that C_x and C_y are *DipoleFit_Flux_Pos_x* – *DipoleFit_Flux_Neg_x* and *DipoleFit_Flux_Pos_y* – *DipoleFit_Flux_Neg_y*, respectively.

(10) DipoleFit_Flux_Pos_Sig:

$$\frac{DipoleFit\_Flux\_Pos}{DipoleFit\_Flux\_Pos\_Sigma} \tag{8}$$

(11) DipoleFit_Flux_Diff:

$$DipoleFit\_Flux\_Pos - DipoleFit\_Flux\_Neg \tag{9}$$

(12) DipoleFit_Flux_Rel:

$$\frac{DipoleFit\_Flux\_Diff}{DipoleFit\_Flux} \tag{10}$$

(13) DipoleFit_Flux_Neg_Sig:

$$\frac{DipoleFit\_Flux\_Neg}{DipoleFit\_Flux\_Neg\_Sigma} \quad (11)$$

(14) DipoleFit_Flux_x: remains *unchanged* from the original attribute set.
(15) DipoleFit_Flux_y: remains *unchanged* from the original attribute set.

3. Proposed method

Based on the recent study of Tabacolde et al. (2018) with GOTO data classification, the problem of class imbalance can be handled more effectively using the cluster-based undersampling as compared to the RUS technique and oversampling counterpart. However, this methodology is still applied to the dataset as a whole without realizing that there may be different learning contexts within a given data. In general, a single classification model is rarely accurate across data subsets, thus requiring a unique classifier built for each of the possible contexts. To allow context-based learning, the proposed framework first makes use of a trustworthy clustering algorithm to generate clusters (i.e., data contexts) from the data $X \in R^{N\times D}$, where N and D denote the size of samples and features, respectively. For this research, the concept of noise-induced cluster ensemble (i.e., ensemble clustering) is exploited to deliver these cluster based contexts. Then, context-specific classification schemes can be formulated to categorize new instances. These stages are elaborated in the following sections.

3.1. Context generation using cluster ensemble

The noise-induced cluster ensemble (Panwong et al., 2020; Iam-On, 2020; Panwong et al., 2018) that has proven more accurate than other ensemble models is specifically exploited for this purpose. Given a dataset X and a desired ratio of noise $\alpha\% \in \{1,2,\ldots,100\}$, the clustering process can be described below.

Step1. To start with, generate a set of A variations of the original dataset $X \in R^{N\times D}$, i.e., $(X'_1, X'_2, \ldots, X'_A)$, in such a way that each of these $X'_j, j = 1\ldots A$ contains randomly selected α^* positions of noise. Note that the localization of noise is determined using the salt-and-pepper method. In addition, the number of these locations $\alpha^* \in \{1,\ldots,N\}$ is subjected to numbers of samples, features and α. Formally, α^* can be estimated as follows.

$$\alpha^* = \left\lceil \frac{ND\alpha}{100} \right\rceil \quad (12)$$

where D is the number of features, i.e., $D = 15$ for the current research.

Step2. For each of the variation $X'_j \in \{X'_1,\ldots,X'_A\}$, the identified positions are filled with noise that is a random value within the feature domain. Before moving to the actual step of noise injection, domains of all D features are standardized. For a feature $f_p, p = 1\ldots D$, the normalized value $x_{i,p} \in [0,1], i = 1\ldots N$ is estimated from the initial value $x^*_{i,p}$ by the following equation.

$$x_{i,p} = \frac{x^*_{i,p} - min_p}{max_p - min_p} \quad (13)$$

provided that min_p and max_p correspond to the minimum and maximum values occurring in the dataset X for the feature f_p. For a data variations $X'_j, j = 1\ldots A$, each selected position specific to attribute f_p is filled in with a noise value. In particular, it is randomly selected as a continuous value within the normalized interval of $[0,1]$. This can be regarded as a special case of normal distribution, which provides better performance than conventional ensemble methods (Panwong et al., 2018).

Step3. After filling in noise values, those perturbed data variations or matrices will be exploited to produce base clusterings using the classical k-means technique and the Random-k strategy (Boongoen and Iam-On, 2018). To be more precise, the number of clusters (k) is randomly selected from the range $\{2,3,\ldots,\sqrt{N}\}$. This is constrained to $\{2,3,\ldots,50\}$ if $\sqrt{N} > 50$. With the data matrix $X'_j, j = 1\ldots A$, k-means is applied for Y trials to create a set of solutions $\{\pi_1(X'_j), \pi_2(X'_j), \ldots, \pi_Y(X'_j)\}$.

Step4. Having completed the previous step for all the perturbed data matrices, the resulting partitions have to be aggregated and represented in a meaningful format. To this end, the pairwise-similarity matrix of Fred and Jain (2005) is used to combine those base clusterings. Each entry $\theta_{uv} \in [0,1], u, v \in \{1,\ldots,N\}$ in the similarity matrix Θ denotes the similarity between instances $x_u, x_v \in X$. Based on a base clustering $\pi_e(X'_j)$ where $j = 1\ldots A, e = 1\ldots Y$, the similarity $\theta_{uv}(\pi_e(X'_j))$ between x_u and x_v is 1 if they are assigned to the same cluster, 0 otherwise. Given all the Y base clusterings generated from the perturbed data matrix $X'_j, j = 1\ldots A$, the similarity can be concluded as follows.

$$\theta_{uv}(X'_j) = \frac{\sum_{e=1\ldots Y}\theta_{uv}(\pi_e(X'_j))}{Y} \quad (14)$$

Provided that, the similarity θ_{uv} is calculated by the following.

$$\theta_{uv} = \frac{\sum_{j=1\ldots A}\theta_{uv}(X'_j)}{A} \quad (15)$$

Step5. After $\Theta \in [0,1]^{N\times N}$ being formulated, a consensus function can be applied to create the final clustering π^*. For this research, k-means is exploited as to set the benchmark for more complex alternatives, in addition to its simplicity and efficiency. Note that the method of Mehar et al. (2013) is employed to automatically find the optimal number of cluster (K) for the dataset represented by Θ. To make this procedure more concisely defined and reproducible, the following algorithm named **Noise-Induced-Ensemble** summarizes all the five processing steps explained above.

3.2. Context specific classification models

Given the desired clustering result $\pi^* = \{C^*_1,\ldots,C^*_K\}$ and the corresponding set of centroids $z^* = \{z^*_1,\ldots,z^*_K\}$, these K clusters is considered for the formation of classification contexts as follows.

- If a cluster $C_e \in \pi^*$ is pure with samples belonging to one class only, a specific data context $CTX_e \subset X, CTX_e = C_e$ and the relation $CTX(z_e) = CTX_e$ are formed. Supposed that $\pi^* = \{C^*_1, C^*_2, C^*_3\}$ and the cluster C^*_1 is pure, the resulting context CTX_1 and relation $CTX(z_1) = CTX_1$ are initiated. And specific to CTX_1, a simple classification rule $CL_1 = 0$ can be created if all samples in the context CTX_1 belong to class 0, or $CL_1 = 1$ if all samples in the context CTX_1 belong to class 1, otherwise.
- On the other hand, other clusters that are not pure are combined to the same context $CTX_\delta \subset X$. Based on the same example in which two clusters C^*_2 and C^*_3 are not pure, they are aggregated to form $CTX_\delta = C^*_2 \cup C^*_3$, with the two corresponding relations $CTX(z_2) = CTX_\delta$ and $CTX(z_3) = CTX_\delta$ being specified. After establishing this, a specific classifier CL_δ is generated for the context CTX_δ using the classification algorithm t.

Algorithm: Noise-Induced-Ensemble (X, α, η, A, Y, K)

$X \in [0,1]^{N\times D}$, a normalized dataset of Nsamples and D features;
α, a desired ratio of noise between 1 to 100%;
η, a choice of algorithm to create ensemble members & final clustering, e.g., k-means;
A, a number of perturbed data matrices generated by injecting noise into X;
Y, a number of ensemble members generated from each perturbed data matrix
K, a number of clusters preferred in the final clustering
(1) **For** each data perturbation j = 1 ...A
(2) $X'_j \leftarrow X$
(3) Randomly select α^* positions in X'_j (see Eq. 12)
(4) **For** each selected entry $x_t \in X'_j, t = 1 \ldots \alpha^*$
(5) $x_t \leftarrow$a random value in [0, 1]
(6) **For** each ensemble member $\pi_e(X'_j), e = 1 \ldots Y$
(7) $\pi_e(X'_j) \leftarrow \eta(X'_j, k), k$ is randomly selected from $\{2, \ldots, \sqrt{N}\}$
(8) Generate the pairwise matrix Θusing Eqs. 14 and 15
(9) Create final clustering $\pi^* \leftarrow \eta(\Theta, K)$
(10) **Return** $\pi^* = (C^*_1, C^*_2, \ldots, C^*_K)$

As for the prediction made to a test or unseen sample $y \in [0,1]^{1\times D}$, the selection of context-based classification model can be summarized by the following steps and associated algorithm.

Step1. Firstly, the sample under question y will be mapped to the cluster ensemble that is used to create the pairwise similarity matrix Θ, such that the resulting representation of this sample is transformed to $y' \in [0,1]^{1\times N}$. See details of this process in the **Mapping-New-Sample** algorithm.

Step2. Having obtained the new representation y' of a test sample, find distances $d(y', z^*_q)$ between y' to all centroids $z^*_q \in z^*$, using the Euclidean metric of:

$$d(y', z^*_q) = \sqrt{\sum_{o=1\ldots N}(y'_o - z^*_{q,o})^2} \tag{16}$$

Step3. Then, select the centroid $z' \in z^*$, where the distance between y' and z' is the minimum from those estimated in Step 2.

$$z' = \arg\min_{\forall z^*_q \in z^*} d(y', z^*_q) \tag{17}$$

Step4. And finally, find out the data context from the relation $CTX(z')$, then select the appropriate classifier or rule to generate the predicted class of y.

Algorithm: Mapping-New-Sample (y, Π)

y, a new sample where $y \in [0,1]^{1\times D}$;
y', the tansformed representation of y where $y' \in [0,1]^{1\times N}$;
Π, a cluster ensemble with A × Y members, with Y clusterings are; generated from each $X'_j, j = 1 \ldots A$;
sim(a,b,π), a function that returns 1 if samples a and b they are assigned; to the same cluster in clustering $\pi$and 0, otherwise;
(1) **For** each sample $x_i, i = 1 \ldots N$
(2) $y'_i \leftarrow 0$
(3) **For** each clustering $\pi_g \in \Pi$
(4) $y'_i \leftarrow y'_i + \sim(y, x_i, \pi_g)$
(5) $y'_i \leftarrow \frac{y'_i}{A\times Y}$
(6) **Return** $y' = (y'_1, \ldots, y'_N)$

4. Performance evaluation

This section presents the design of empirical study, which aims to assess and compare accuracies between the proposed method and other relevant techniques. It is followed by a report of results with discussion that provides other useful theoretical and practical issues.

4.1. Experimental design

Table 2 provides details of the two datasets exploited in this study, each of which is described in terms of numbers of samples belonging to the two classes (i.e., Class1 and Class0 that correspond to real transients and bogus samples, respectively) and corresponding percentages. Based on both percentages of Class1 samples that are around 0.3%, these datasets provide a great challenge to the research community of imbalance classification. At the same time, this illustrates an actual scenario of discovering transient events, which rarely happen and appear in a survey. Other experimental settings are summarized as follows.

- For the application of noise-induced cluster ensemble, the noise ratio of $\alpha = 8\%$ is investigated as suggested by the original work (Panwong et al., 2018). For each dataset, the localization and noise injection trials (A) and the clusterings created from each perturbed matrix (Y) are all set to 20. In addition, the k-means clustering technique is employed to create both ensemble members and the final clustering.
- Having obtained the target clustering result, the contexts and associated relations are formed in accordance with the steps identified previously. Specific to the context CTX_δ, three classical classifiers are exploited as the preferred algorithm t: NB (Naive Bayes with the Gaussian kernel function), C4.5 (Decision Tree with the maximum depth of 10) and KNN (k-Nearest Neighbors, where k = 1), respectively. These settings form a basis to compare the proposed framework with its baseline, where the whole dataset is considered as one context of CTX_δ. Note that KNN is included here to represent the result obtained by a lazy learning model where a distance metric is simply used to determine the predicted class from a nearest neighbor. Similar to KNN where all features contribute to the estimation of a prediction output, NB approaches this using a different concept of conditional probability, which is later simplified by the assumption of independency among features. In contrary, C4.5 differentiate the significance among features, i.e., which one should be used to assess a sample under examination first, and which are later. A decision tree is built to form branches of such an order, which allows a classification to be made based on a subset of original features. This collection of classification algorithms also present two different approaches to analyzing a numerical dataset, which are usually included in many comparative studies of classification problem (Alghobiri, 2018). On one hand, refined domains of numeric features are exploited as they are for the estimation of distance metric used by KNN. On the other, they are reduced to intervals by C4.5 and NB to simplify sample-class relations. More complex alternatives like classifier ensemble (Dong et al., 2020) and a deep learning model (Dong et al., 2021) may be explored in the future work.

Table 2
Description of examined dataset: numbers of class-specific samples and percentages.

Dataset	No. of all samples	Class 0 samples	Class 0 percentage	Class 1 samples	Class 1 percentage
Data1	5,989	5,973	99.733	16	0.267
Data2	6,771	6,753	99.734	18	0.266

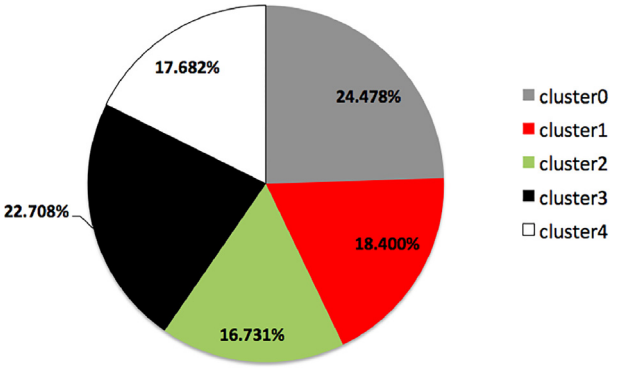

Fig. 3. Percentages of data distribution among different clusters in Data1.

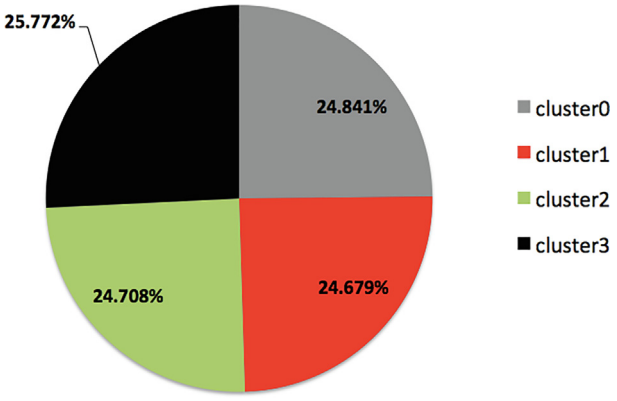

Fig. 4. Percentages of data distribution among different clusters in Data2.

- To allow a robust comparison, 20 trials of 10-fold cross validation are employed to determine F1 measure $\in [0, 1]$, where the rate of 1 indicates the most effective classifier with no false positive nor false negative. It can be defined by the following equations, where TP = true positive, FP = false positive, TN = true negative and FN = false negative.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{18}$$

where $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$.

4.2. Results and discussion

After the initial stage of applying the noise-induced cluster ensemble to the two datasets under examination, optimal numbers of clusters for Data1 and Data2 are 5 and 4, with percentages of data distribution among different clusters being illustrated in Figs. 3 and 4, respectively. The Cluster0 in both cases are similarly pure with samples of Class0 only. Henceforth, simple rules can be formed to classified a new instance, whose distance to z_0^* is the shortest among available centroids, as a member of Class0. Specific to Data1, samples belonging to Cluster1, Cluster2, Cluster3 and Cluster4 are combined to form the context CTX_δ. Likewise, for Data2, samples in CTX_δ are from Cluster1, Cluster2 and Cluster3. With respect to the F1 metric, Fig. 5 presents the comparison of those scores obtained by the baseline and context-based counterpart, across three classification algorithms identified earlier. It is clearly shown that the proposed framework usually delivers a more effective classifier than the baseline model, i.e., average F1 values from the 10-fold cross validation are improved from 0.0091 to 0.0162 with NB, from 0.0000 to 0.0325 with C4.5, and from 0.1667 to 0.3636, respectively. Similarly, the average F1 scores achieved with Data2 is given in Fig. 6, which confirms the effectiveness of the context-based strategy. In particular, the scores of the baseline are lifted from zeros to 0.3478 and 0.5000 by the coupling of data contexts with KNN and C4.5.

From previous illustrations, NB appears to be the least accurate among three classification techniques with the best F1 measures of 0.0162 for Data1 and 0.0178 for the other dataset. This observation is caused by the sparseness of data represented as zero conditional probabilities between features and the minority class that has been exploited within this model. As such, through the smoothing mechanism that replaces zeros with small numbers, the resulting probability of Class1 can be much lower than that of Class0, hence a lack of ability to recognize real transients. Nonetheless, the proposed approach is able to reduce the number of samples belonging to Class0 (around 24–25% for the entire datasets, see Figs. 3 and 4), thus partly decrease the difference between class-specific proba-

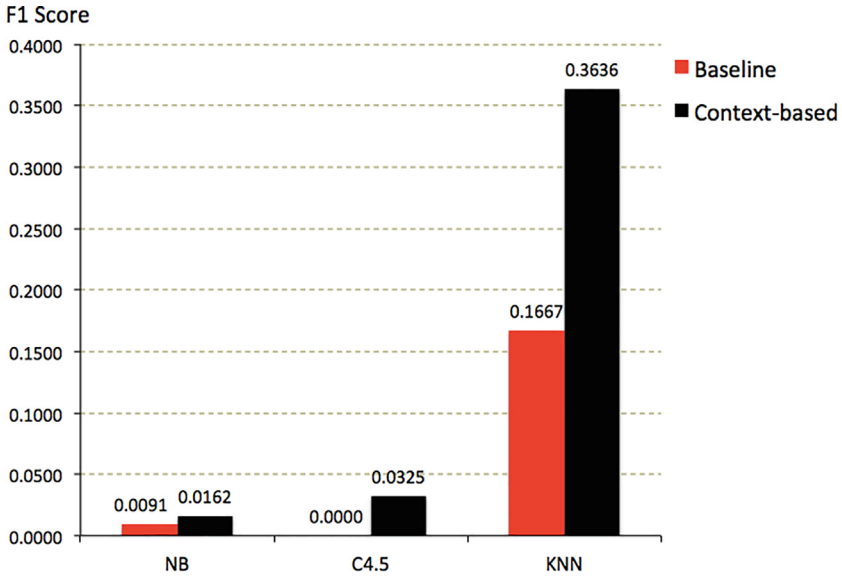

Fig. 5. F1 scores obtained by different classification models with Data1. These are averages summarized from 20 trials of 10-fold cross validation.

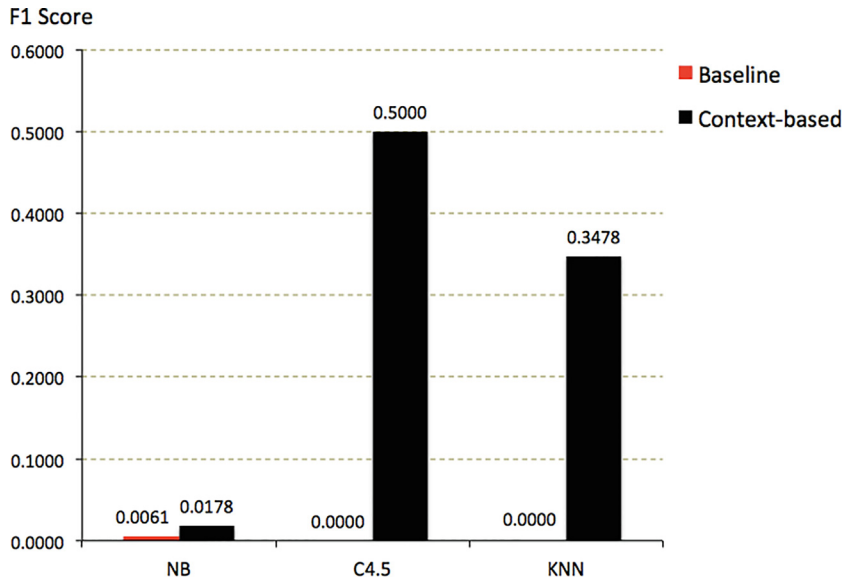

Fig. 6. F1 scores obtained by different classification models with Data2. These are averages summarized from 20 trials of 10-fold cross validation.

Table 3
Precision (PR) and Recall (RC) obtained with all investigated classification models for Data1. These are averages summarized from 20 trials of 10-fold cross validation with corresponding values of standard deviations being given in (brackets).

Classifier	PR(Baseline)	PR(Context-based)	RC(Baseline)	RC(Context-based)
NB	0.0046	0.0082	0.5625	0.5625
	(0.0036)	(0.0032)	(0.1201)	(0.1062)
C4.5	0.0000	0.0187	0.0000	0.1250
	(0.0000)	(0.0079)	(0.0000)	(0.0842)
KNN	0.2500	0.6667	0.1250	0.2500
	(0.0974)	(0.1013)	(0.0883)	(0.1006)

Table 4
Precision (PR) and Recall (RC) obtained with all investigated classification models for Data2. These are averages summarized from 20 trials of 10-fold cross validation with corresponding values of standard deviations being given in (brackets).

Classifier	PR(Baseline)	PR(Context-based)	RC(Baseline)	RC(Context-based)
NB	0.0031	0.0092	0.1667	0.2778
	(0.0030)	(0.0022)	(0.1001)	(0.0924)
C4.5	0.0000	1.0000	0.0000	0.3333
	(0.0000)	(0.0000)	(0.0000)	(0.0721)
KNN	0.0000	0.4444	0.0000	0.2857
	(0.0000)	(0.0871)	(0.0000)	(0.0722)

bilities with every feature. This leads to cases that Class1 probability becomes more comparable to that of the other, and as a result, F1 measures are marginally improved from the baseline alternative. Between C4.5 and KNN, the former seeks to find highly discriminative features to present a root node and others in the upper layer of a decision tree. However, in Data1, correlation measures between features and target classes are rather low, thus demoting the effectiveness of this method, which may make use of some but not all available features. This is in line with the result of KNN that achieves a better result by simply including all features to find a nearest neighbor. With Data2 where correlations between features and classes get higher than before, C4.5 has become more effective as some individual features alone are informative to classify new samples, compared to the aggregation of all features pursued by the KNN counterpart.

Besides the overview given above, Tables 3 and 4 provides more details about precision and recall measures (i.e., both averages and corresponding standard deviations from 10-fold cross validation) obtained by different classification models for Data1 and Data2. Based on the former, the context-based framework is able to improve both precisions and recalls using C4.5 and KNN. As for the other dataset, the proposed mechanism is exceptional such that both measures have been significantly improved over those of the baseline. For instance, the recall scores of both C4.5 and KNN that are simply nothing before have become more desirable for the discovery in astronomy (i.e., it is preferred that a model is capable of recalling all the real sources). Another interesting finding from these results is that reducing the size of the original data to the context CTX_{δ} not only makes the prediction more efficient, but also allows a classifier like C4.5 that determines significance of features more effective. With the arguments made thus far, the clustering-led context generation proves to be useful and gives a good foundation for further development. For the context CTX_{δ}, the undersampling method introduced by Tabacolde et al. (2018) can be exploited prior the creation of a classifier.

For the interpretation of experimental results thus far, averages across multiple trials are exploited for simplicity. This initial assessment approach follows the central limit theorem (CLT) sug-

gesting that the observed statistics in a controlled experiment can be justified to the normal distribution. However, to obtain a more robust comparison between context-based classifiers and their baseline models, the number of times that one is 'significantly better' and 'significantly worse' (of 95% confidence level) than the others are investigated next. Let $\mu(i,j,t)$ be the average of F1 scores, across the t-th run of n-fold cross validation (n is 10 for the current research) for a technique $i \in TC$ (TC contains a context-based classifier and its baseline model), on a specific dataset $j \in DAT$ (DAT consists of Data1 and Data 2). Formally, $\mu(i,j,t)$ can be defined as follows:

$$\mu(i,j,t) = \frac{1}{n}\sum_{\eta=1}^{n} F1_{\eta}(i,j,t), \tag{19}$$

where $F1_{\eta}(i,j,t)$ denotes the $F1$ score obtained from the η-th fold within the t-th run of method i, on the dataset j. The comparison of means obtained from a single trial of cross validation may be misleading, as the difference between means may not be statistically significant at times. As such, it is more reliable to make a decision based on the 95% confidence interval for the mean $\mu(i,j,t)$. Such an interval is defined by the following.

$$\left[\mu(i,j,t) - 1.96\frac{Std(i,j,t)}{\sqrt{n}},\quad \mu(i,j,t) + 1.96\frac{Std(i,j,t)}{\sqrt{n}}\right], \tag{20}$$

where $Std(i,j,t)$ denotes the standard deviation of $F1$ measures across n-folds cross validation of the t-th trial, for a technique i over a dataset j. The statistical significance of the difference between any two methods $i, i\prime \in TC$ over any dataset $j \in DAT$ is found if there is no intersection between their confidence intervals of $\mu(i,j,t)$ and $\mu(i\prime,j,t)$. For any dataset j, a classifier i is significantly *better* than the other model $i\prime$ when

$$\left(\mu(i,j,t) - 1.96\frac{Std(i,j,t)}{\sqrt{n}}\right) > \left(\mu(i',j,t) + 1.96\frac{Std(i\prime,j,t)}{\sqrt{n}}\right) \tag{21}$$

Following that, the frequency that one technique $i \in TC$ is significantly *better* than the other across all experimented trials and datasets, i.e., $B(i)$, is calculated by the next equation.

$$B(i) = \sum_{\forall j \in DAT}\ \sum_{\forall t=1\ldots20}\ \sum_{\forall i' \in TC, i\prime \neq i} better_j(i,i\prime,t), \tag{22}$$

where

$$better_j(i,i',t) = \begin{cases} 1 \ if \left(\mu(i,j,t) - 1.96\frac{Std(i,j,t)}{\sqrt{n}}\right) > \left(\mu(i\prime,j,t) + 1.96\frac{Std(i\prime,j,t)}{\sqrt{n}}\right) \\ 0 \ otherwise \end{cases} \tag{23}$$

Likewise, the frequency that one technique $i \in TC$ is significantly *worse* than the other, i.e., $W(i)$, is estimated as follows.

$$W(i) = \sum_{\forall j \in DAT}\ \sum_{\forall t=1\ldots20}\ \sum_{\forall i\prime \in TC, i\prime \neq i} worse_j(i,i\prime,t), \tag{24}$$

where

$$worse_j(i,i\prime,t) = \begin{cases} 1 \ if \left(\mu(i,j,t) + 1.96\frac{Std(i,j,t)}{\sqrt{n}}\right) < \left(\mu(i\prime,j,t) - 1.96\frac{Std(i\prime,j,t)}{\sqrt{n}}\right) \\ 0 \ otherwise \end{cases} \tag{25}$$

Based on this statistical evaluation approach, Fig. 7 presents (B-W) statistics, i.e., the difference between frequencies of better and worse, which compare each of four context-based classifiers to their baseline models. Given the number of trials as 20, the range of this (B-W) is between −20 and 20, where the minimum occurs as a context-based classifier is significantly worse than its baseline for all 20 trials of 10-fold cross validation, and the maximum happens as it is constantly better than the baseline model. More details are provided in Table 5, in which frequencies of both better and worse are presented for different classification models. Based on this assessment, the proposed framework is usually more effective than a baseline for both datasets examined herein. In addition, the improvement made to NB is less significant than other two cases, with (B-W) values being around 10. Similar statistics for C4.5 and KNN are 14 and 18 with Data1, 20 and 20 for Data2, respectively. This observation supports the discussion made earlier that NB is still constrained by a problem of data spareness, despite of the help of context-based implementation to decrease the difference of feature-specific probabilities between classes.

Continue from the statistical evaluation emphasized previously, it is interesting to further explore both KNN and C4.5 with respect to their parameter settings, i.e., the number of nearest neighbors (K) and the maximum depth (Depth), respectively. For this purpose, additional assessments are conducted by repeating the aforementioned experiment for different parameter values. Specific to KNN, Fig. 8 shows F1 measures which have been obtained with different values of K $\in \{1,2,3,4,5,6\}$, and categorized by datasets. Note that, just like before, these scores are averages from 20 trials

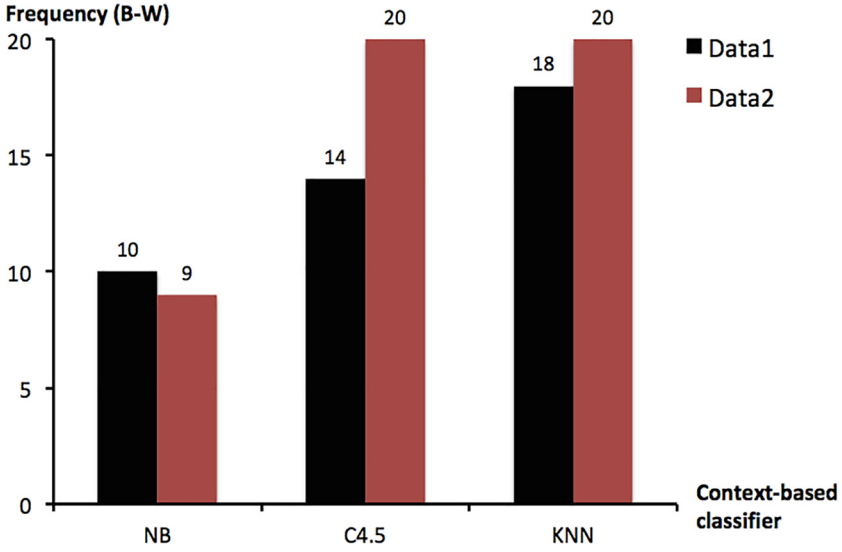

Fig. 7. Frequency of (B-W) obtained by four context-based classifiers, each of which is compared to its baseline model.

Table 5
Better and worse frequencies obtained by different classification models, categorized by two datasets investigated in this study. Note that these are acquired from a comparison between each context-based classifier and its baseline counterpart.

Dataset	Classifier	Modeling approach	Better (frequency)	Worse (frequency)
Data1	NB	Baseline	2	12
		Context-Based	12	2
	C4.5	Baseline	0	14
		Context-Based	14	0
	KNN	Baseline	0	18
		Context-Based	18	0
Data1	NB	Baseline	1	10
		Context-Based	10	1
	C4.5	Baseline	0	20
		Context-Based	20	0
	KNN	Baseline	0	20
		Context-Based	20	0

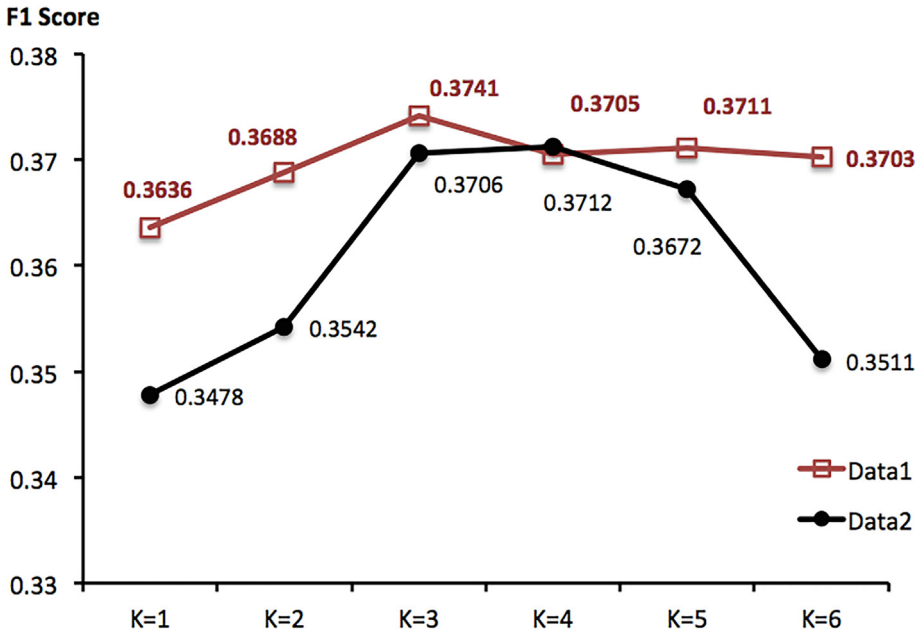

Fig. 8. F1 scores obtained by KNN with different values of $K \in \{1, 2, 3, 4, 5, 6\}$. Note that these are averages from 20 trials of 10-fold cross validation.

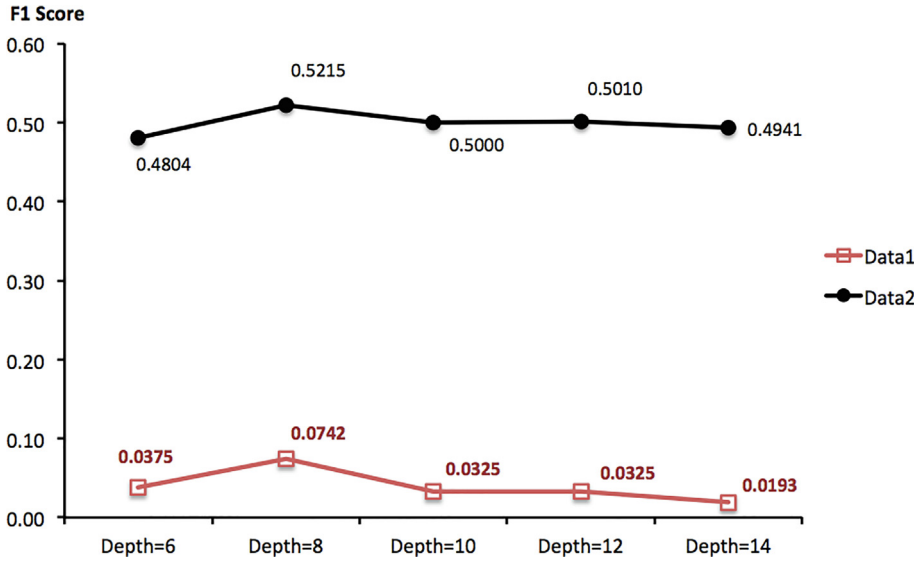

Fig. 9. F1 scores obtained by C4.5 with different values of Depth $\in \{6, 8, 10, 12, 14\}$. Note that these are averages from 20 trials of 10-fold cross validation.

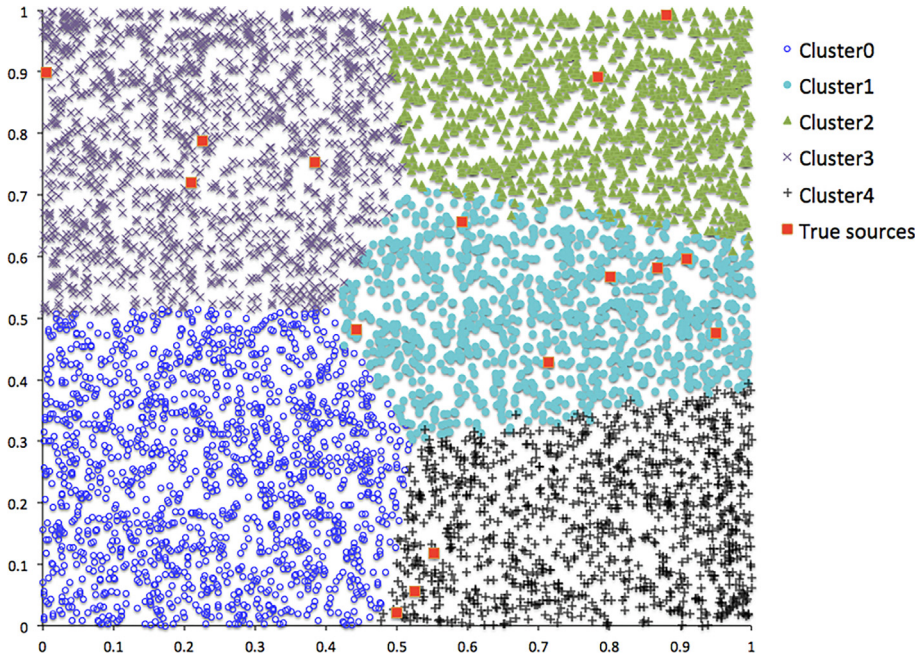

Fig. 10. A scatter plot of samples in Data1, categorized into five clusters where x-axis and y-axis correspond to the two attributes of DipoleFit_Flux_x $\in [0, 1]$ and DipoleFit_Flux_y $\in [0, 1]$. Note that the true sources are highlighted by red blocks.

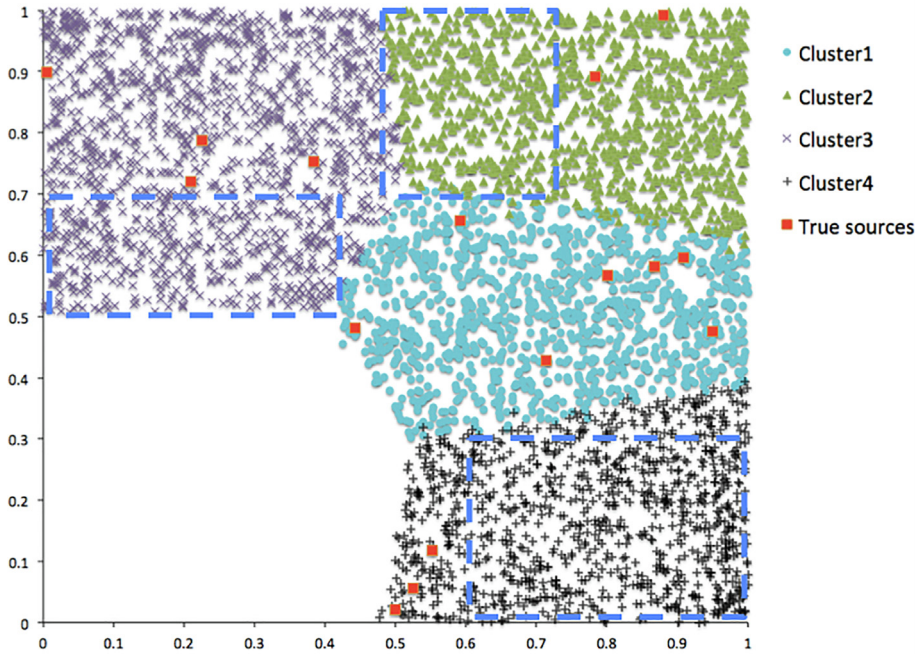

Fig. 11. A scatter plot of samples in Data1, categorized into five clusters where x-axis and y-axis correspond to the two attributes of DipoleFit_Flux_x $\in [0, 1]$ and DipoleFit_Flux_y $\in [0, 1]$. Note that the blue dashed blocks with samples of class0 only are those areas that can be further filtered out.

of 10-fold cross validation. With this figure, the optimal K of 3 should be employed instead of the current choice of K = 1, such that F1 is improved from 0.3636 to 0.3741 in Data1, and from 0.3478 to 0.3706 in Data2. Likewise, Fig. 9 presents F1 values obtained with various depths of decision tree, i.e., Depth $\in \{6, 8, 10, 12, 14\}$. It is noteworthy that the greatest value of Depth is 15 that is the number of available features, is not included here as it does not lead to an improvement to Depth of 14. In accordance with this illustration, the optimal setting of Depth = 8 can slightly lift the model accuracy from the current choice of Depth = 10, i.e., from 0.0325 to 0.0742 in Data1, and from 0.5000 to 0.5215 in Data2. Given this insight, classification performance can be maximized, either though the use of recommended values or the empirical framework exploited to generate these two figures.

It is also important to point out that the current work provides a new benchmark for researches in both areas of general data mining as well as astronomical data analysis. As such, most of the studies focus on assessing a new method or a collection of them over datasets in which the minority class occupies 1% to 30% of the whole

training set. Very few such as Nanni et al. (2015) has a try with the case of minority class being between 0.5% to 1%. Given this, the investigation with the two datasets in this paper (minority class is less than 0.5%) provides a rare opportunity to assess how well the proposed and existing techniques can cope with the extreme scenario. Fig. 10 presents the scatter plot of samples belonging to five clusters in Data1, with the true sources being highlighted in red marks. Note that x-axis and y-axis of this plot correspond to the two attributes of DipoleFit_Flux_x $\in [0, 1]$ and DipoleFit_Flux_y $\in [0, 1]$, which exhibit the best correlations to the known classes. It is obvious from this figure that cluster0 can be effectively identified with minimal overlapping with the rest. As a result, a classification rule built upon the context CTX_0 delivers good performance. However, the context CTX_δ covering samples from the other four clusters obtains a small number of true sources sparsely distributed over this space. With this illustration, one may realized that a single model might not be equally effective for the sub-problems found in this context. Intuitively, it may be possible to break each of the four clusters further down such that the areas specified with dashed-blue blocks in Fig. 11 are filtered out. Of course, the level of imbalance will be even better with this iterative-like clustering, hence the accuracy of classification model. Nonetheless, a stopping criterion of such a process seems to be critical as to prevent the event of overfitting.

5. Conclusion

This paper has presented a new framework to provide data contexts from which different classification strategies can be established. The proposed idea is unique and different from a conventional approach to develop a single model to solve all possible sub-problems within the data under examination. In particular, these contexts are derived using the recently published method of noise-induced cluster ensemble that shows exceptional performance across several benchmark data collections. Based on the empirical study with two simulated datasets generated within the GOTO project and three well-known classification algorithms, the context-based framework usually leads to better predictive quality than the baseline counterpart. It is also noteworthy that the new method is generalized such that the resulting contexts can be coupled with existing techniques to solve the imbalance problem.

Despite the reported improvement, one possible way to take this framework forward is to repeatedly apply the clustering to filter out those samples of the majority class. This is rather similar to the bi-level learning mechanism employed for the task of face detection (Boongoen et al., 2016). As mentioned earlier, figuring out effective stopping criteria would be a challenge. Another significant work that may reveal an important factor to enhance the proposed framework further is applications of different ensemble matrices, consensus functions and aggregation operators (Boongoen, 2017). A great deal of alternatives can be found in the literature, which has been developed over the past two decades (Boongoen and Iam-On, 2018; Pattanodom et al., 2016; Iam-On and Boongoen, 2015). Besides, possible applications of fuzzy reasoning (Fu et al., 2010) and clustering-based data discretization (Sriwanna et al., 2017) can also be further studied to add the explanation aspect to the desired classification process. Moreover, it will be interesting to combine this context generation with convolutional neural networks (CNN). This follows the recent trend of exploiting deep learning technology in the astronomy domain (Wright et al., 2017). For instance, Deep-HiTS (Cabrera-Vives et al., 2017) that is a rotation-invariant convolutional neural network model has been introduced to classify images of transient candidates for the High cadence Transient Survey (HiTS).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is partly supported by MFU, STFC GCRF 2018 project: From stars to Baht Phase II (the collaboration between MFU and University of Sheffield), and Newton Institutional Link 2020-21 project (the collaboration between MFU and Aberystwyth University).

References

Abbott, B.P., Abbott, R., T.D.A., et al., 2020. Prospects for observing and localizing gravitational-wave transients with advanced ligo, advanced virgo and kagra. Living Reviews in Relativity 23, 3.

Alghobiri, M., 2018. A comparative analysis of classification algorithms on diverse datasets. Eng. Technol. Appl. Sci. Res. 8, 2790–2795.

Bagui, S., Li, K., 2021. Resampling imbalanced data for network intrusion detection datasets. J. Big Data 8, 6.

Boongoen, T., 2017. A brief review of fuzzy aggregation. NKRAFA J. Sci. Eng. Technol. Innov. 13, 59–66.

Boongoen, T., Iam-On, N., 2018. Cluster ensembles: a survey of approaches with recent extensions and applications. Comput. Sci. Rev. 28, 1–25.

Boongoen, T., Iam-On, N., Undara, B., 2016. Improving face detection with bi-level classification model. NKRAFA J. Sci. Eng. Technol. Innov. 12, 52–63.

de Buisson, L., Sivanandam, N., Bassett, B., Smith, M., 2015. Machine learning classification of SDSS transient survey images. Mon. Not. R. Astron. Soc. 454, 2026–2038.

Cabrera-Vives, G., Reyes, I., Forster, F., Estevez, P., Maureira, J., 2017. Deep-HiTS: rotation invariant convolutional neural network for transient detection. Astrophys. J. 836, 97.

Dong, S., Wang, P., Abbas, K., 2021. A survey on deep learning and its applications. Comput. Sci. Rev. 40, 100379.

Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q., 2020. A survey on ensemble learning. Front. Comput. Sci. 14, 241–258.

Dyer, M.J., Dhillon, V.S., Littlefair, S., Steeghs, D., Ulaczyk, K., Chote, P., Galloway, D., Rol, E., 2018. A telescope control and scheduling system for the Gravitational-wave Optical Transient Observer (GOTO). In: Proceedings of International Conference on Observatory Operations: Strategies, Processes, and Systems, pp. 124–137.

Fred, A.L.N., Jain, A.K., 2005. Combining multiple clusterings using evidence accumulation. IEEE Trans. Pattern Anal. Mach. Intell. 27, 835–850.

Fu, X., Boongoen, T., Shen, Q., 2010. Evidence directed generation of plausible crime scenarios with identity resolution. Appl. Artif. Intell. 24, 253–276.

Iam-On, N., 2020. Clustering data with the presence of attribute noise: a study of noise completely at random and ensemble of multiple k-means clusterings. Int. J. Mach. Learn. Cybern. 11, 491–509.

Iam-On, N., Boongoen, T., 2015. Diversity-driven generation of link-based cluster ensemble and application to data classification. Expert Syst. Appl. 42, 8259–8273.

Juric, M., 2015. The LSST data management system. ArXiv e-prints, arXiv:1512.07914.

Kauffmann, O.B., Le-Fevre, O., et al., 2020. Simulating JWST deep extragalactic imaging surveys and physical parameter recovery. Astron. Astrophys. 640, A67.

Lin, C., Hsieh, T., Liu, Y., Lin, Y., Fang, C., Wang, Y., 2018. Minority oversampling in kernel adaptive subspaces for class imbalanced datasets. IEEE Trans. Knowl. Data Eng. 30, 950–962.

Mehar, A.M., Matawie, K., Maeder, A., 2013. Determining an optimal value of k in k-means clustering. In: Proceedings of IEEE International Conference on Bioinformatics and Biomedicine, pp. 51–55.

Meisner, A., Bromley, B., Nugent, P., Schlegel, D., Kenyon, S., Schlafly, E., Dawson, K., 2017. Searching for Planet Nine with coadded wise and neowise-reactivation images. Astron. J. 153, 65.

Mullaney, J.R., Makrygianni, L., et al., 2020. Processing GOTO data with the Rubin Observatory LSST Science Pipelines I: Production of coadded frames. In arXiv (Instrumentation and Methods for Astrophysics) (pp. 1–17)..

Nanni, L., Fantozzi, C., Lazzarini, N., 2015. Coupling different methods for overcoming the class imbalance problem. Neurocomputing 158, 48–61.

Ofek, N., Rokach, L., Stern, R., Shabtai, A., 2017. Fast-cbus: a fast clustering-based undersampling method for addressing the class imbalance problem. Neurocomputing 243, 88–102.

Panwong, P., Boongoen, T., Iam-On, N., 2018. Improving consensus clustering with noise-induced ensemble generation: A study of uniform random noise. In: Proceedings of International Conference on Machine Learning and Computing, pp. 390–395.

Panwong, P., Boongoen, T., Iam-On, N., 2020. Improving consensus clustering with noise-induced ensemble generation. Expert Syst. Appl. 146, 113–138.

Pattanodom, M., Iam-On, N., Boongoen, T., 2016. Hybrid imputation framework for data clustering using ensemble method. In: Proceedings of Asian Conference on Information Systems, pp. 86–91.

Seiffert, C., Khoshgoftaar, T., Hulse, J.V., Napolitano, A., 2010. Rusboost: a hybrid approach to alleviating class imbalance. IEEE Trans. Syst. Man Cybern. Part A 40, 185–197.

Soraisam, M.D., Gilfanov, M., Kupfer, T., Prince, T.A., Masci, F., Laher, R.R., Kong, A., 2018. Multiwavelength approach to classifying transient events in the direction of M31. Astron. Astropyhs. 615, 1–9.

Sriwanna, K., Boongoen, T., Iam-On, N., 2017. Graph clustering-based discretization of splitting and merging methods (graphs and graphm). Human-centric Comput. Inf. Sci. 7, 1–39.

Tabacolde, A.B., Boongoen, T., Iam-On, N., Mullaney, J., Sawangwit, U., Ulaczyk, K., 2018. Transient detection modeling as imbalance data classification. In: Proceedings of IEEE International Conference on Knowledge Innovation and Invention, pp. 180–183.

Tabacolde, A.B., Boongoen, T., Iam-On, N., Mullaney, J., Sawangwit, U., Ulaczyk, K., 2018. Transient detection modelling for gravitational-wave optical transient observer (goto) sky survey. In: Proceedings of International Conference on Machine Learning and Computing, pp. 384–389.

Tang, B., He, H., 2017. Gir-based ensemble sampling approaches for imbalanced learning. Pattern Recogn. 71, 306–319.

Wette, K., 2021. Geometric approach to analytic marginalisation of the likelihood ratio for continuous gravitational wave searches. Universe 7, 174.

Wright, D., Lintott, C., Smartt, S., Smith, K., Fortson, L., Trouille, L., 2017. A transient search using combined human and machine classifications. Mon. Not. R. Astron. Soc. 472, 1315–1323.