Gaussian Probabilistic Classification with Gaia Photometric Science Alerts data

Esteban Chalbaud

February 2022

Abstract

In this work we present the Gaussian Probabilistic Classification Method (GPCM) as a Machine Learning classifier applied to the Gaia photometric science alerts data (GPSA). We employ a k-d tree cross-matcher with the full third data release from Gaia satellite in combination with the GPSA data-base to extract magnitude and color data. Selection cuts in full Gaia data release are given by selecting sources above the galactic plane in the interval ($70° < b < 90°$) with magnitude given by $14 \leq M \leq 20$. We apply the method for classifying unknown sources inside the GPSA data by using already tagged objects inside the following classes: Supernovae type Ia (SN Ia), Supernovae type II (SN II), quasi-stellar objects (QSO) and active galactic nuclei (AGN). During the training of the classifier we find the the Tied method gives the highest accuracy with 80% of correctly classified objects. By applying the trained classifier we find that most of the unknown sources in the GPSA catalog are tagged as Quasi-stellar objects (QSO) with 52% of total the predictions falling within this category for the unknown GPSA sources.

Keywords: Machine-Learning, classifiers, Gaussian Probabilistic Classification Method

Introduction

The Gaia satellite, launched in 2013, is a photometric survey of bright object sources orbiting the Lagrange point L2. The probe explores the wavelenght in the a band-width that covers from the extended visual range between near-UV and near infrared. One of the main goals of the probe was to perform astrometry analysis of the galaxy sources while creating a 3-dimensional map of astronomical objects in the milky way [1]. Figure (1) shows the stellar density of the Milky way obtained from the third data release measurements form Gaia.

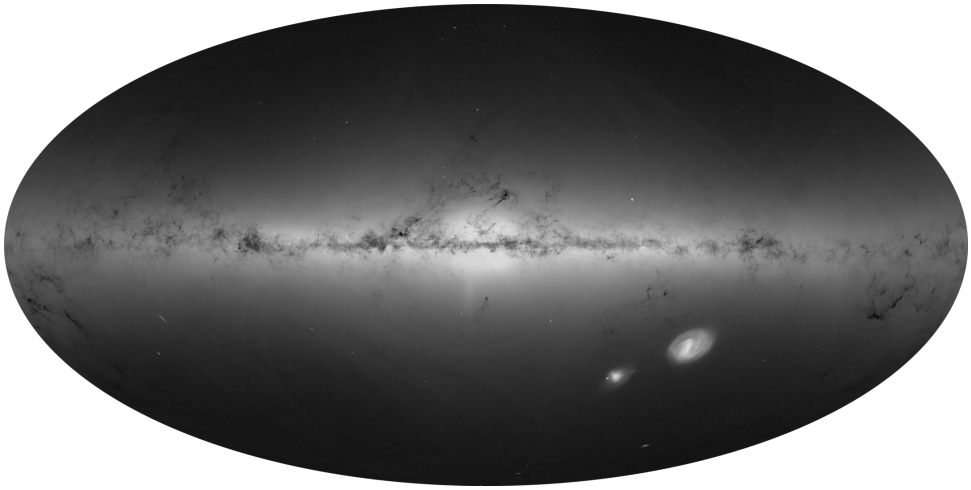

Figure 1: Stellar density map from the 3th Gaia Data Release (source: ESA)

Among the mission goals, it is expected that the Gaia mission increase the understanding on the dynamical evolution of the Milky way by measuring near 1% of the total stellar population of the galaxy. The mission is also used to understand the dynamical evolution of galaxies, the stellar evolution by using high-accurate photometry data collection to test stellar models, stellar variability and exploring parameter space for exoplanet detection in far stars [**?**].

These mission goals has been set in the context of full-sky astrometric repeated measurements of rapidly bright changing sources in the sky. These *alerts* are used as alerts of transient objects that varies significantly in velocity and observed magnitude (i.e. Supernovae, AGN, Quasars, etc). Figure (2), shows the an example of detected rapidly bright changing sources in the sky for data collected until the end of 2019.

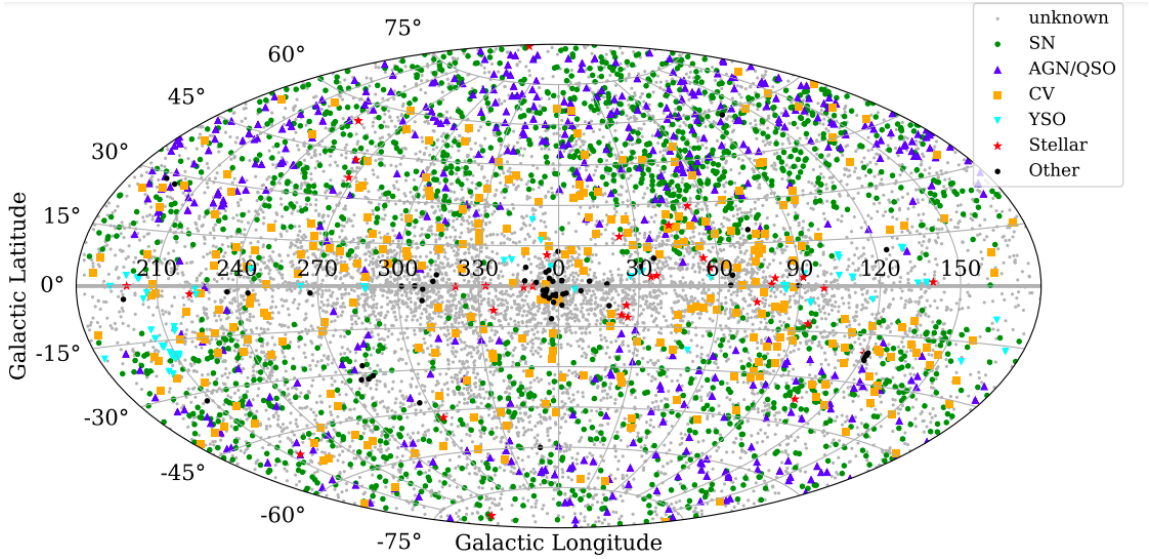

Figure 2: Gaia alerts data from measurements recorded until the end of 2019 for several known sources in the sky SN, AGN, etc. [2]

The Gaia alerts photometric data, is intended to be a catalog of object with rapid variation in brightness in the sky classified as transient phenomena. The collection of the data allows to identify objects in the sky that behaves in a broad category of object described by current stellar evolution models (if they are in the galaxy) or extra-galactic objects (i.e. SN, AGNs, etc.). A complete description of the transient data analysis can be found here [2].

Despite most of the detection of the sources in the catalog can be clearly associated with known sources of objects generating a transient in the data, it can be possible that classification of the sources cannot be performed specially if they are given in the region of the satellite spectrum where identification cannot be easy. The identification and posterior classification of stellar or extra-galactic objects might be interesting to test stellar evolution models in the milky way.

Considering this, in this work we show the so called Gaussian Probabilistic Classification Method as an statistical technique for classification data within four classes: Supernovae type Ia (SN Ia), Supernovae type II (SN II), quasi-stellar objects (QSO) and active galactic nuclei (AGN), from the measured transients events. We show in this work that we can classify unknown sources inside this catalog by using already classified objects by only using their color properties and measured magnitude for sources located in a region of the sky that is centered in the North galactic pole with 20° radii.

This work is organized as follow:

We firstly present the data set used in the analysis showing the GPSA recorded until February 18th of 2022, we show the cross-matching technique used to recover the color properties by using the full data release catalog with selection cuts given by a magnitude interval of $14 \leq M \leq 20$ inside the galactic North pole with colatitude $70^\circ < b < 90^\circ$.

After cross-matching the data we present the Gaussian Mixture Model as a technique used for classifying the data, from where we concentrate in four different constructions of the classifier based in the covariance matrix shape.

Finally we train the classifier by using the already tagged data inside the GPSA catalog from where we employ the trained classifier to show that unknown sources can be classified by using the four dominant classes inside the GPSA catalog: Supernovae type Ia (SN Ia), Supernovae type II (SN II), quasi-stellar objects (QSO) and active galactic nuclei (AGN). Lastly, in the conclusions we summarize by showing that most of the unknown objects are classified as QSOs for the most accurate classifier.

Data analysis

In this analysis we employ two data sets from the Gaia satellite: The full Gaia third data release and the Photometric Alert Science data (GPSA). In this section we describe each data set used in the analysis and show the cross-matching technique employed for identifying the features inside the GPSA catalog inside the full data release.

As it is mentioned in [2], the photometric alert science data contains the sources with significant change from a constant magnitude, while the scanning of the entire sky is performed. The catalog of objects contains the Name of the source (unique name for internal identification), time of observation, time of publication, angular coordinates (i.e. Right ascension and declination) in the IRCS frame, magnitude, mean historic magnitude computed from previous measurements, historic scattering in variation of magnitude of the source, and the class of the source previously classified by the collaboration. Figure (3) shows distributions for the typical variables inside the data for the Supernovae Ia class of objects inside the data.

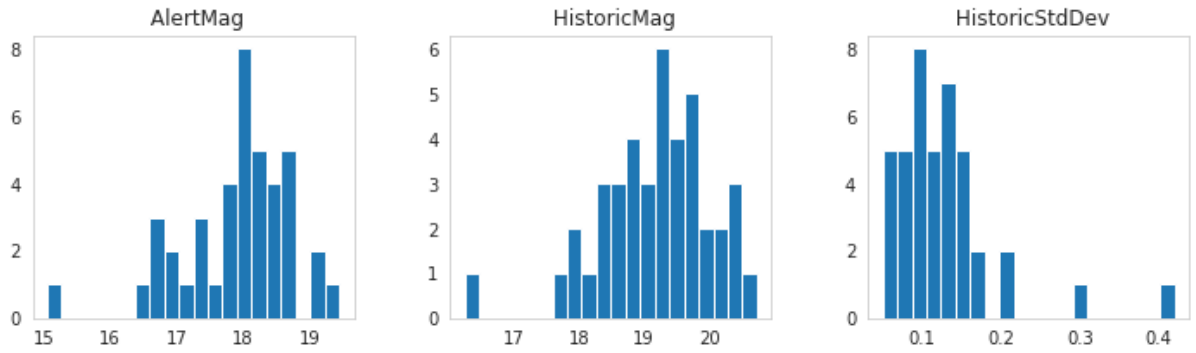

Figure 3: Distributions of Supernovae Ia inside the GPSA data for absolute magnitude, historic magnitude observed by satellite scans, and the historic dispersion for the magnitude changes in the sources.

Additionally we use the full Gaia data release available in: (https://gaia.aip.de/query/). The description of the data release can be found in [3]. From a collection of 1811709771 sources, it is possible to use cuts in the data to obtain distributions in color plane space for the sources in Figure (4) we show an example of a subset of full Gaia data when there is no constrain in magnitude but randomly selected 10^6 objects.

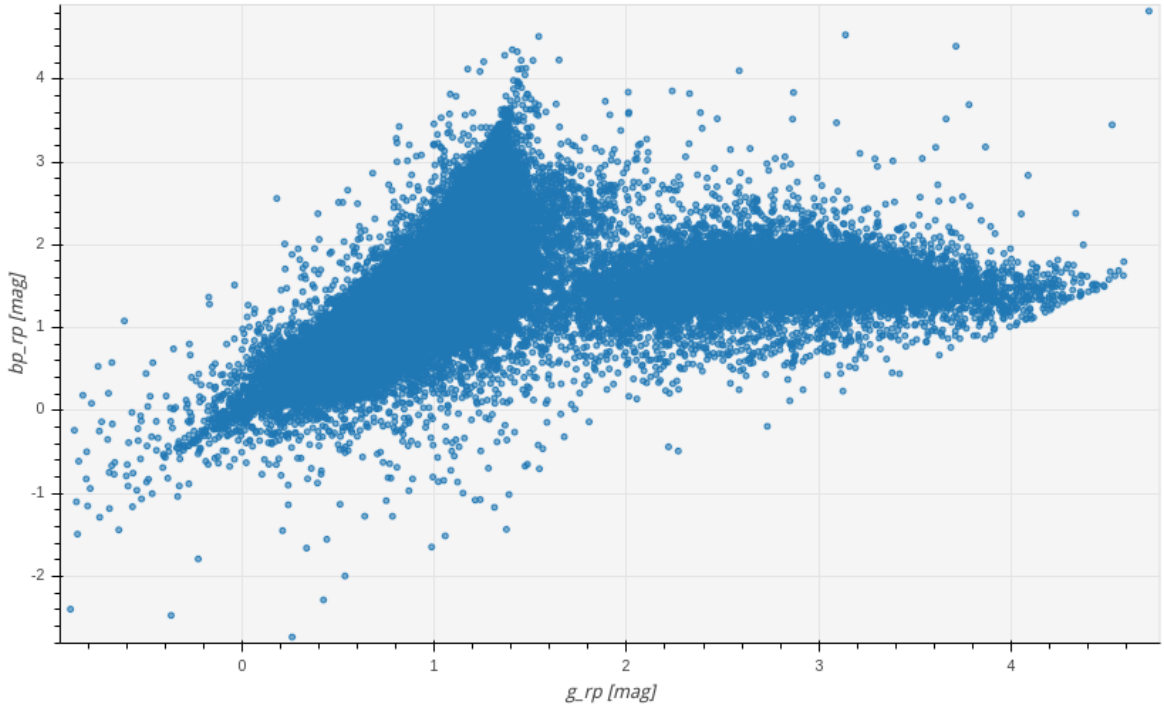

Figure 4: Example of the data contained in the data release for the color plane space.

It is important to point out that, since we are interest in classifications of the sources without using the already features inside the GPSA data another data needs to used for a given source. In order to solve this problem, we use a cross-matching between catalogs to use features that are different from those used inside the GPSA catalog for building the class of the source (the classification for this case uses the historic magnitude and its standard deviation and it is actually employed by the collaboration for alert classification [2]).

Figure (5) shows how the cross-matching algorithm is employed to cross-match the sources inside the GPSA catalog into the full Gaia data release. The source in the catalog is compared against a grid that has been obtained by recursively dividing the grid until every object is contained in a single division. The functioning of the method for matching two sources can be described by the following steps:

- Look for an object with the median right ascension, split the catalogue into objects left and right.
- Find the objects with the median declination in each partition, split the partitions into smaller partitions but in up and down directions.
- Find the objects with median right ascension in each of the partitions, split the partitions again into smaller cells of objects left and right.
- Repeat the previous two steps until each partition until each cell contains only one object in it.
- Calculate the angular distance.

For this analysis we have used the library `match-to-catalog-sky` inside the `Astropy` library.

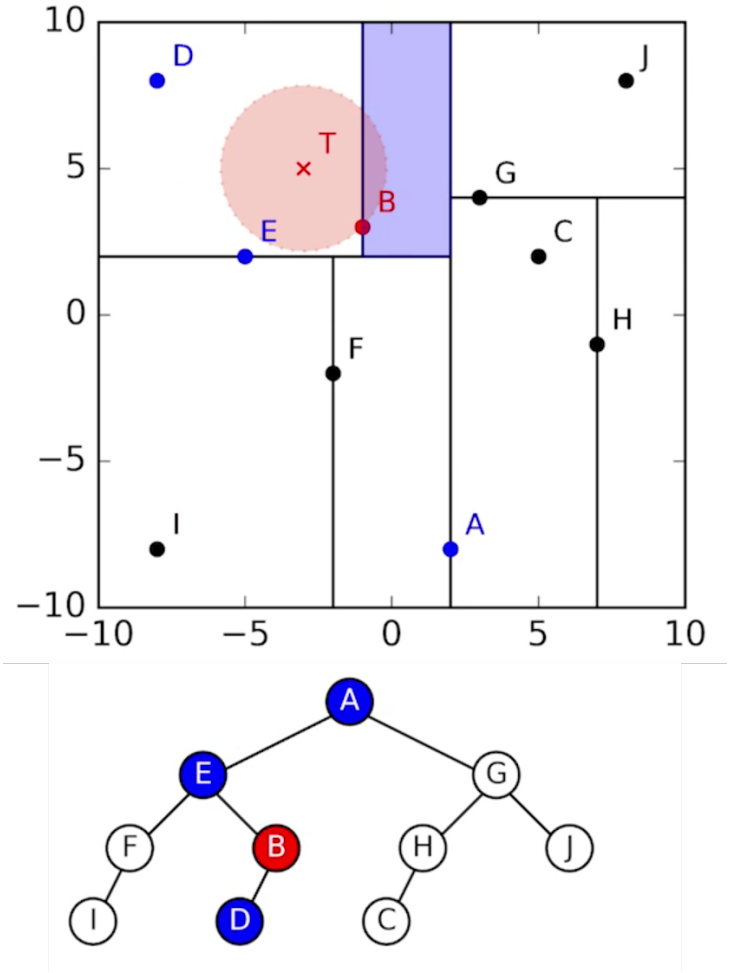

Figure 5: Diagram explaining the cross-matching technique used in this analysis for a given source inside the catalog A compared with the catalog B and its corresponding k-d tree classification for matching the source. Notice that cells start to appear as long as more division are performed in the data. (credits: Coursera)

Methodology

In the present work we use the an Gaussian Probabilistic Classification Method based in a Gaussian Mixture Model [4] as a classifier of the data. The method uses Gaussian distributions to model the posterior distribution for a given data, by using its features, for a given class assigning a probability of the data for belonging to a given class. The method is a machine learning clustering algorithm since the data is classified once the probabilities overpasses a given threshold for a given class.This threshold is calculated by computing the Bayesian Information Criterion to assess the number of clusters in the data. By using a given number of Gaussian variables, the probability density for the data x belonging into a class C_k is given by:

$$P(x|C_k) = \sum_{q=1}^{Q} a_{k,q}\phi(x|m_{k,q}V_{k,q}) \tag{1}$$

where ϕ is a multivariate normalized multivariate Gaussian distribution with mean value m and covariance matrix V. The coefficients $a_{k,q}$ corresponds to the mixing coefficients that should be positive and normalize to 1. Notice that since the method depends in the covariance matrix used to compute probability distributions the classification will depend on how the covariance is built. In this analysis we focus in four combinations for covariance matrix construction based in how the classes objects (i.e. SN Ia, SN II, AGN, QSO) in the analysis we have used four ways of building the covariance matrix:

- Full: In this analysis each component has its own covariance matrix computed from the data. In this case it corresponds to case where covariances for color space (magnitude, g-rp and bp-g) are calculated independently.
- Tied: In this case the covariance matrix contains all possible correlations in the data by using a general covariance matrix in the analysis. For this case all the components share the same covariance matrix.
- Diagonal: Each component has a diagonal matrix for each component in the analysis, neglecting internal cross-correlations in the data for each component.
- Spherical: The covariance matrix is diagonal for each component (magnitude, g-rp and bp-g) and also a single valued, i.e. each gaussian has only one variance.

The Machine Learning methodology consists in calculating the weights and the mean values in the distribution in equation (1) by training the classifier while computing the Bayesian Information Criterion coefficient to obtain the thresholds and classify the data. The calculation is performed by using Expectation-maximization technique [5], from where the parameters are obtained.

Results

From the performed analysis we obtain that most of the sources inside the GPSA catalog are cross-matched inside the full Gaia release for the given cuts in the data, when the angular distance inside the cross-matcher is fixed within the angular resolution of the satellite. i.e. 1". Figure (6) shows the matching of the GPSA inside the full GDR for the magnitude region between $14 \leq M \leq 20$.

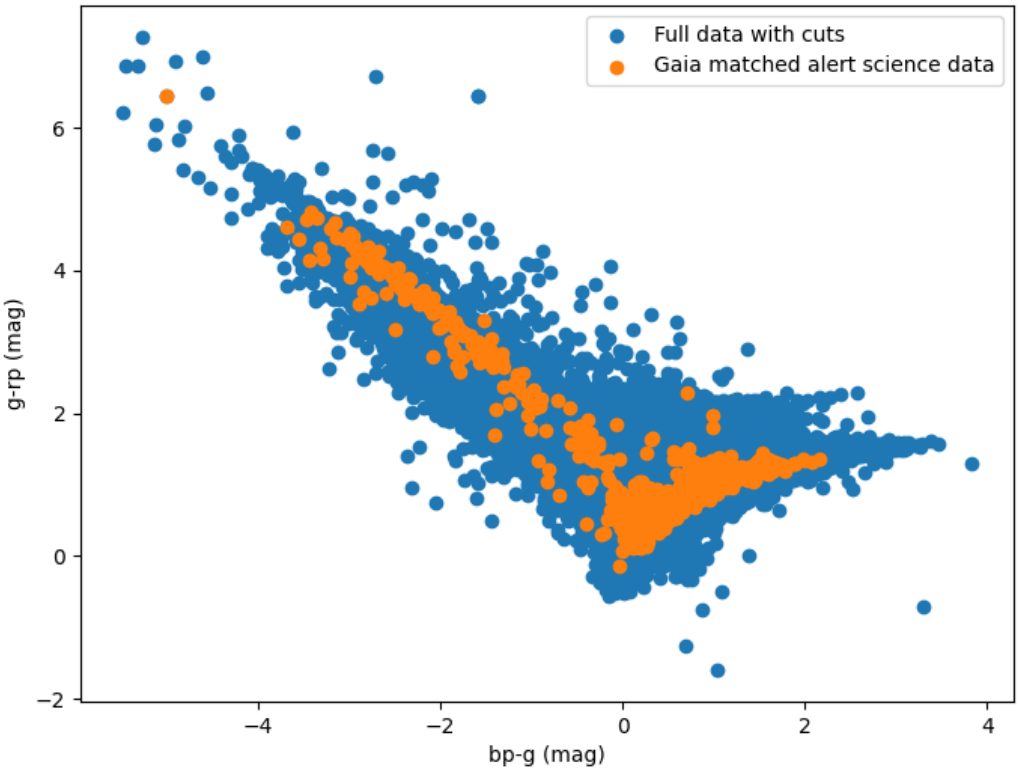

Figure 6: Sources measured by the Gaia satellite for the data cut selection described in the previous section (blue) against the source from the GPSA (orange) after the cross-match with the catalogs.

From the figure we notice that by using the selection cuts in the full data release we mostly reproduce the observed color limit boundary pointed out in [6]; however, the limit region is not exact due to the lack of color corrections in the analysis having few sources outside this limit. Additionally, the color limit boundary appears to be an important source systematic in classifying the object than belongs to this region in the color plane since this region correspond to an area from where there is strong stellar contaminants in data. Notice that most of the sources inside the GPSA catalog are indeed contained nearby this region.

As it is pointed out in the previous section, we perform a naive classification of the unknown sources inside the GPSA data by using the already classified alert science data from the catalog from the collaboration [3, 2] using a different technique. In this analysis we focus in the classes for the sources that are associated with magnitude and color (i.e. bp-g and g-rp plane) From the already cross-matched sources. By training our GPCM these features are used inside the classifier to obtain reliable classifications from the already known source inside the cross-matched data. Figure (7), shows the results of classifying the known region inside the GPSA catalog and the accuracy of the classifier when several covariance methods are used in the analysis.

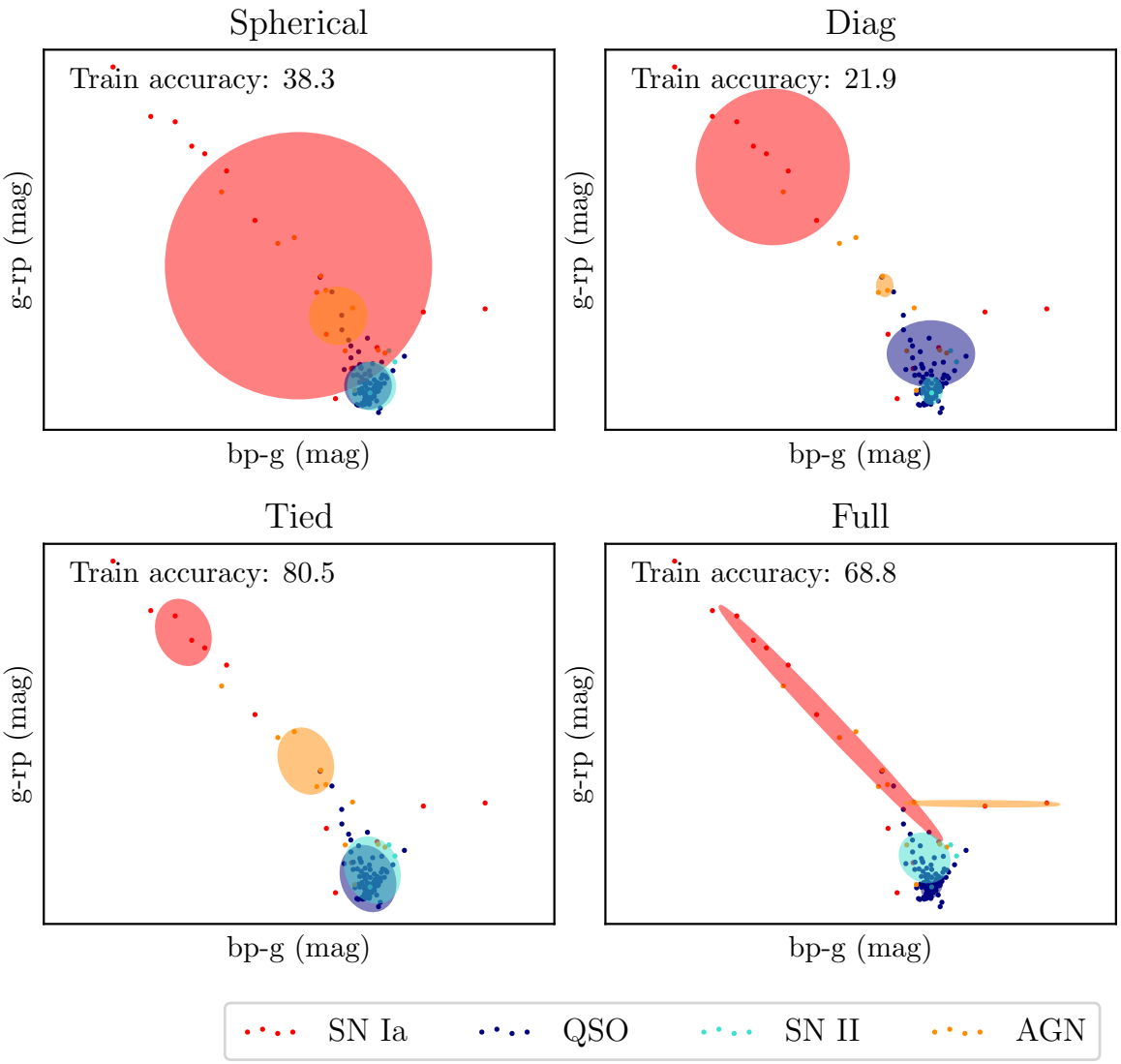

Figure 7: Training classification of the known part of sources inside the GPSA data for several covariance matrices used inside the Gaussian Probabilistic Classification method (i.e. Spherical, Diag, Tied and Full). Ellipses contains the 68% confidence level for the posterior distribution of the classified sources within the classes considered in the analysis SN Ia, QSO, SN II and AGN.

The results shows a broad change in the accuracy in the predictions from the GAPS known region depending on the covariance matrix is built. As it is described in the previous section, depending on the covariance used in the analysis, correlations between different classes of objects (i.e. SN Ia, SN II, QSO and AGN) can be captured or not.

From figure (7), we observe that the tied covariance matrix case contains the most accuracy in the data when the predictions from the method are compared with the tagged classification already contained inside the GPSA catalog. We can interpret this result by the usual Bayesian inference from where matrices that combines correlations and cross-correlations in data (as the tied covariance indeed is able to perform) are expected to ensure better constraints or in this case better accuracy in classification of the data.

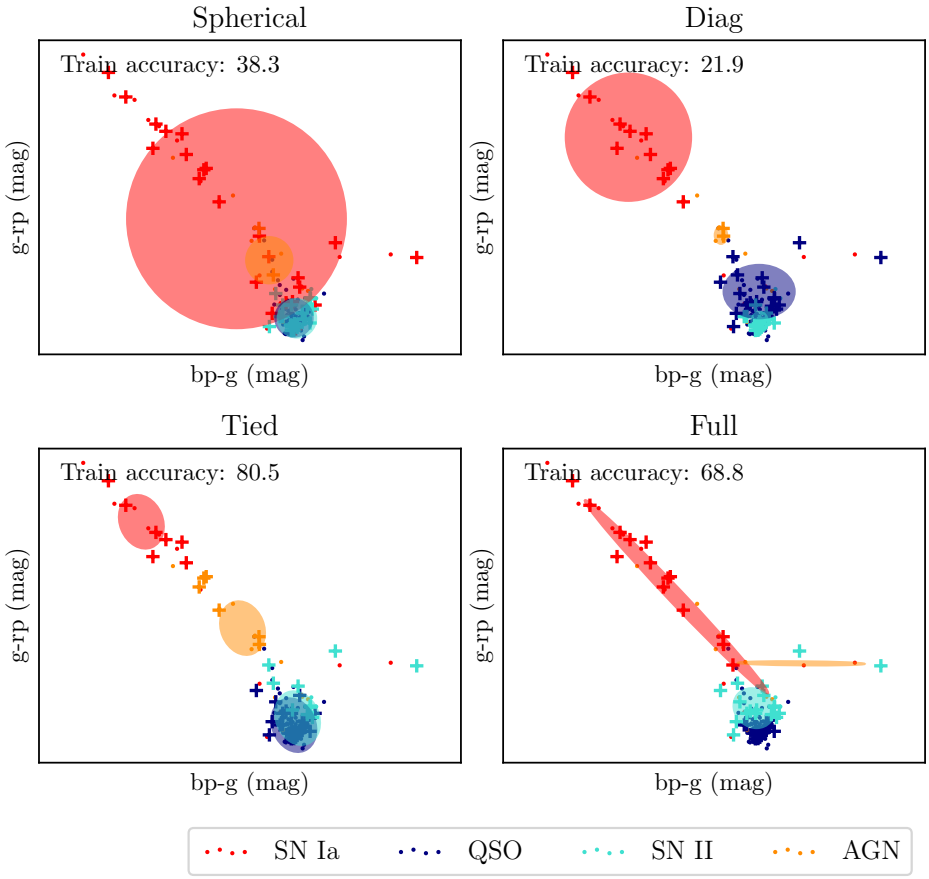

Figure 8: Classification of the unknown sources (cross-shaped marker) inside the GPSA data for several covariance matrices used inside the Gaussian Probabilistic Classification method (i.e. Spherical, Diag, Tied and Full). Classified sources are less than $2-\sigma$ from the confidence ellipses.

On the other hand, if we apply the trained classificator with the cross-matched part of the GPSA catalog that is tagged as uknomn, we obtain that for most of the covariance

cases studied here agrees in the classification of the sources in the region way before the color edge limit region. Figure (8) shows this apparent behavior for all of the covariances used in the analysis except for the one with the higher accuracy when the unknown sources are added (cross-shaped marker). Notice that classified unknown sources are less than $2-\sigma$ away from the centers of the confidence ellipses.

Covariance type	SN Ia	QSO	SN II	AGN	Total
Spherical	23	1	22	2	48
Diagonal	11	18	17	2	48
Tied	7	25	10	6	48
Full	14	16	18	0	48

Table 1: Classifications of data from the different covariance matrices used inside method after training. The outputs are given when the classifier is applied into the unknown sources inside the GPSA catalog after full cross-matching with the 3th Gaia data release.

Summary and Conclusions

We have shown a Machine Learning classifier technique following [6], by applying it to GPSA data by using several covariance constructions inside the algorithm. We have shown that using three features in the data associated with color $g-rp$ and $bp-g$ in addition with the source magnitude measured by the satellite. It is possible to train a classifier with an accuracy above 70% and using it to classify sources inside the GPSA data that currently are not classified as one of the classes of objects shown here (i.e Supernovae type Ia (SN Ia), Supernovae type II (SN II), quasi-stellar objects (QSO) and active galactic nuclei (AGN)).

We obtain that for the explored classifications in this work the one with the highest accuracy corresponds to the Tied covariance inside the GPCM method, that uses a global covariance matrix along with all the features in the data. From this first trial in classifying theses sources we obtain that 15% os the sources corresponds to Supernovae type Ia, 52% to quasi-stellar objects, 21% to Supernovae type II, and 12% to active galactic nuclei for a cross-matched population of 48 objects that were correctly identified inside the full Gaia data release from the cross-matcher with a angular distance difference of $1"$.

Finally, we point out that it is possible to use or combine different features in data and including different maximal angular distance in the cross-matching sources to improve the accuracy of the estimators by having more sources in the analysis or distinct features in the explored classes that allow to improve the classification.

References

[1] T. Prusti, J. de Bruijne, A. Brown, A. Vallenari, C. Babusiaux, C. Bailer-Jones, U. Bastian, M. Biermann, D. Evans, L. Eyer, F. Jansen, C. Jordi, S. Klioner, U. Lammers, L. Lindegren, X. Luri, F. Mignard, D. Milligan, C. Panem, and S. Zschocke, "The gaia mission," *Astronomy and Astrophysics*, vol. 595, 09 2016.

[2] S. Hodgkin, D. Harrison, E. Breedt, T. Wevers, G. Rixon, A. Delgado, and al, "Gaia photometric science alerts," *Astronomy and Astrophysics*, vol. 652, 06 2021.

[3] C. Ordenovic, A. Brown, A. Vallenari, T. Prusti, J. de Bruijne, C. Babusiaux, O. Creevey, L. Eyer, S. Klioner, and G. Collaboration, "Gaia early data release 3. summary of the contents and survey properties," *Astronomy and Astrophysics*, 11 2020.

[4] C. Fraley and A. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, pp. 611–631, 06 2002.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, pp. 1–38, 1977.

[6] C. A. L. Bailer-Jones, M. Fouesneau, and R. Andrae, "Quasar and galaxy classification in gaia data release 2," *Monthly Notices of the Royal Astronomical Society*, vol. 490, p. 5615–5633, Oct 2019.