Astroinformatics
Proceedings IAU Symposium No. 325, 2016
M. Brescia, S.G. Djorgovski, E. Feigelson,
G. Longo & S. Cavuoti, eds.

doi:10.1017/S1743921317000266

Automatic Source Classification in Digitised First Byurakan Survey

Martin Topinka[1], Areg Mickaelian[2], Roberto Nesci[3] and Corinne Rossi[3]

[1]Dublin Institute for Advanced Studies,
31 Fitzwilliam place, Dublin 2, Ireland
email: martin.topinka@gmail.com
[2]Byurakan Astrophysical Observatory,
Byurakan, Aragatzotn, AM 0213, Armenia
[3]Universita di Roma 'La Sapienza',
Piazzale A. Moro 2, 00185 Roma, Italy

Abstract. The Digitised First Byurakan Survey (DFBS) provides low dispersion optical spectra for about 24 million sources. A two-step machine learning algorithm based on similarities to predefined templates is applied to select different classes of rare objects in the dataset automatically, for example late type stars, quasars and white dwarves. Identifying outliers from the groups of common astrophysical objects may lead to discovery of rare objects, such as gamma-ray burst afterglows.

Keywords. methods: statistical, astronomical data bases: surveys

1. Digitised First Byurakan Survey

The First Byurakan Survey (FBS) was the first systematic objective prism survey of the extragalactic sky initiated by Markarian, Lipovetski and Stepanian in between the years 1965–1980 at the Byurakan Astrophysical Observatory with the 1m Schmidt telescope and 1.5° prism (Mickaelian *et al.*, 2007). It contains 2050 Kodak photographic plates $4° \times 4°$ fields, covering 17 000 deg^2 of the Northern and part fo the Southern sky $\delta > -15°$ at high galactic latitudes $|b| > 15°$.

Each FBS plate contains about 15,000 - 20,000 low-dispersion optical spectra, yielding more than 24,000,000 objects in the whole survey. For comparison, SDSS contains about 4,000,000 spectra (SDSS R12). The spectral range is $340 - 690$ nm, with a sensitivity gap near 530 nm, dividing the spectra into red and blue parts.

The plates have been scanned and digitised which resulted in Digitised First Byurakan Survey (DFBS). The spectra have been extracted in a catalog-driven way, using object positions obtained from the USNO-A2 catalogue used as a reference down to the plate limit (17^m). The astrometric solution is within the positional error of $1''$ or less. The 2D spectral boxes have been identified and integrated to yield 1D low dispersion spectra, 141 pixels each.

The DFBS catalogue is available online in the form of a searchable SQL database (DFBS website).

2. Classification

Only a fraction of the USNO-A2 objects found in the DFBS plates has been spectrally classified so far. The machine learning (ML) object classification method proposed in this work is based on finding similarities (similarity measures) between an unknown

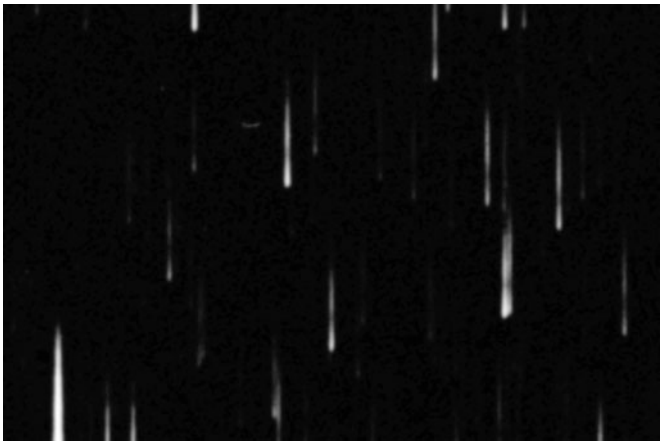

Figure 1. A typical close-up example of a photographic plate #1410 containing low-dispersion spectra from the FBS.

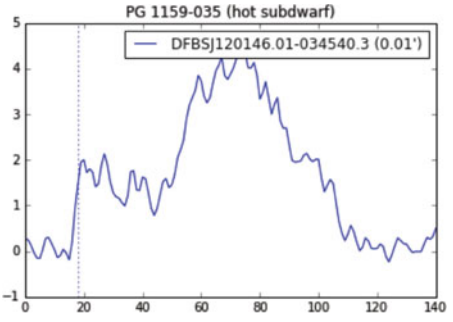

Figure 2. An example of an extracted spectrum of a hot subdwarf DFBSJ120146.01-034540.3. The red cutoff used for the wavelength calibration is marked by a dashed line. The sensitivity drop around the pixel 53 is also clearly seen.

object and predefined templates representing sources of known spectral type. The vast majority of the objects within the DFBS archive are not member of any of the classes represented by the templates. Therefore, to reduce possible mis-classification due to noise, the classification has been performed in four steps: 1) data calibration 2) specifying templates 3) filtering outliers 4) principle component decomposition and clustering.

2.1. *Data Calibration*

Extracted spectra may be misplaced. For any subsequent machine learning classification it is important to calibrate wavelengths for each digitised spectrum to have consistent input feature set. Therefore, the position of a steep drop in sensitivity in the red part of a spectrum was used for wavelength calibration. Hartmann-like formula $\lambda = 190 + 21930/(k - irb + 43.86)$, where k is the pixel position along the spectrum and irb the position of the red cutoff, is used to obtain the wavelength λ in nm. Notice that the pixel to wavelength transformation does not preserve equidistant binning. The accuracy cannot be better than 1 pixel. The dispersion is about 1.5 nm/pix at 370 nm, 2.6 nm/pixel at $H\gamma$ and 10 nm/pix at the red edge. The spectral resolution is at least 2 times worse, as the photographic grains occupy $1.5 - 2$ pix. An illustrative spectrum of a hot subdwarf DFBSJ120146.01-034540.3 (aka PG 1159-035) is shown in Fig. 2

https://doi.org/10.1017/S1743921317000266 Published online by Cambridge University Press

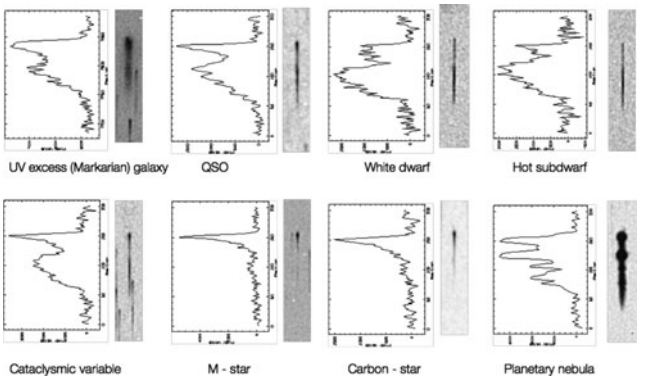

Figure 3. Eight spectra of sources representing template members considered in the classification analysis with their 1D and 2D spectra.

The actual apparent magnitude of a source does not contribute to the spectral classification, the flux at the red peak was used to normalise all spectra.

The classification method was tested on a sub-sample containing 404 270 sources in 24 plates using objects of magnitude between $15 - 16.5$ mag in the R band with $S/N > 8$. Only well extracted, non-overlapping, not corrupted spectra were used in the analysis.

2.2. *Templates*

The primary goal of the classification is to find rare or unusual objects. Eight different object classes have been predefined in the way that the key representative members in each class were identified as they occurred in the DFBS: UV excess (Markarian) galaxies: Mkn 266, Ka 163; quasars (QSOs): CSO 409, PG 1634+706, HS 1626+6433; white dwarves: PG 0109+111, PG 1449+168, PHL 962; hot sub-dwarves: PG 1159-035, PG 1638+147; cataclysmic variables: DQ Her, UX CVn, V1193 Ori; M-stars: BIS 029, BIS 196; Carbon-stars: BIS 001, BIS 036, BIS 184; planetary nebulae: NGC 2242, NGC 6210, NGC 7009, see Fig. 2

For the purpose of template definition, the spectra of the class representatives were averaged. For simplicity, these class groups were merged in three groups only: 'QSO' (containing QSO and Markarian galaxies), 'white dwarves' (containing white dwarves and hot sub-dwarves) and late type stars 'M-stars' (containing M-stars and carbon stars). Planetary nebulae were not considered in the classification.

The advantage of generating templates from the survey itself to obtaining templates from simulations is that the same data acquiring process has been used to obtain both the templates and unknown objects. This lower the importance of exact calibration and determining classification biases coming from instrumental effects. Such templates have the same noise attributes as the unknown spectra.

2.3. *Pre-filtering*

In the preselection phase three distance measures were considered: the cross-correlation distance, the Euclidian distance and the Euclidian distance in the space of integrated spectrum over red and blue parts of the spectrum.

The cross-correlation yields the correlation coefficient r_{ik} defined as

$$r_{ik} = \frac{\Sigma_j (D_{ij} - \hat{D}_i)(T_{kj} - \hat{T}_i)}{\sqrt{\Sigma_j (D_{ij} - \hat{D}_i)^2}\sqrt{\Sigma_j (T_{kj} - \hat{T}_i)^2}} \tag{2.1}$$

https://doi.org/10.1017/S1743921317000266 Published online by Cambridge University Press

where $\hat{D}_j$ and $\hat{T}_j$ are mean values over the j-th variable, D and T stand for the unknown spectrum and the spectrum of a given template.

A point-wise Euclidian distance between two spectra is calculated using

$$d_{ij}^2 = \Sigma_j (D_{ij} - T_{kj})^2 \tag{2.2}$$

where both spectra were first normalised to have zero mean and unit variance.

To mitigate the wavelength calibration displacement and a possible unwanted background fluctuation, the 0, +1 and -1 pixel displaced spectra are used to adopt for a possible inaccuracy during the digitalisation similarly to Bratsolis *et al.* 2000. Only the minimal distance between the template and the shifted spectra was considered. Blue and red parts of the spectrum are analysed separately yielding two new sets of features.

Smoothing the spectra by 3 and 5 bin width window moving average and repeating the similarity measure yields another features.

The rough integrated red and blue flux defined as the sums over the red and blue regions of the spectra are used as two more new features.

Having several values of the similarity measures to each template object, a new vector space is constructed and the objects are sorted by the distance to the nearest template. Since the majority of the objects are regular stars different from all the templates a threshold $D_{th} = 99.99\%$ is set. Only the objects nearer to any of the templates than D_{th} pass to the next step of the analysis.

2.4. *Principal Component Analysis*

Once the rare spectral class member candidates have been preselected in the previous step, the Principal Component Analysis (PCA) is applied to their original spectra to down-size the dimensionality of the problem from the 141 dimensions and to achieve more sensitive classification, transforming the space of the spectra into 10 dimensions, avoiding the curse of dimensionality issue (Ivizic *et al.* 2014). Due to uneven level of noise at each pixel the combination of the signal and the noise is expected and therefore the weighted PCA is applied. The PCA solution solution is found iteratively based upon the Expectation-Maximisation algorithm.

2.5. *Clustering*

Unsupervised clustering based on minimal spanning tree (Ivizic *et al.* 2014) applied on the PCA transformed spectra was then used to find neighbour-hooding objects. The resulted tree is trimmed with the threshold set to 0.05 of the average distance in the dataset. The 2D projection of the tree and tree islands containing the original template members are identified and marked in Fig. 4. The first 5 identified candidates in each group (QSO, white dwarves and late type stars) were inspected visually to approve wether they belong to the class of objects suggested by the algorithm.

3. Conclusion

Preliminary results of the ML classification based on clustering around known cluster centres defined by the set of templates showed promising result on the tested sub-sample of 404,270 spectra ($\sim 17\%$ of the total volume of DFBS). The match within the chosen trimming threshold is 100%. Confirmation of the identified candidates as a function of the tree trimming threshold is going to be studied in the up-coming work.

Due to the large size of the full dataset splitting the dataset to chunks is required. Parallel versions of the algorithms (parallel PCA, minimum tree spanning and merging

https://doi.org/10.1017/S1743921317000266 Published online by Cambridge University Press

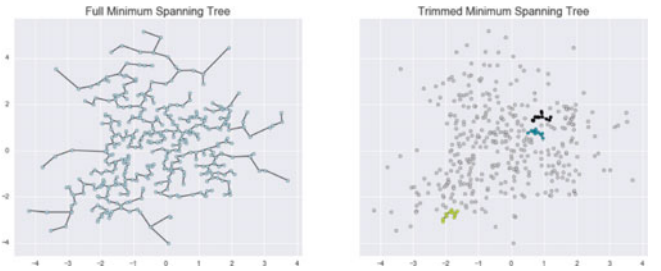

Figure 4. Full (left) and trimmed (right) minimum spanning tree shows clustering around templates as cluster centres for QSOs (blue colour), carbon and M stars (black) and white dwarves (yellow). Two dimensional PCA projection of the feature space is used for visualisation purposes.

and thresholding during the preselection phase) have been tested successfully. Therefore, extending the analysis to the entire DFBS is feasible and will be a subject of an on-going work. The most computationally expensive "big data" part of the algorithm lies in the preselection phase which is run in parallel for small chunks of data.

There is a risk that the leading PCA component in the PCA decomposition does not reflect the class division and would not be sensitive to the classification task. In classification cases for which a good size training sample for a sub-class exists (e.g. M and Carbon stars) the bagging algorithm with positive cases only (PU learning) will be applied. The PCA performs poorly if data are contaminated with high level of noise. To extend the analysis for fainter or noisy sources, more robust feature extraction method, such as Bernoulli Restricted Boltzmann machine model, will be used.

Identifying outliers (small isolated clusters) from the groups of common astrophysical objects may lead to discovery of rare objects, e.g. unusually blue objects such as gamma-ray burst afterglows.

The machine learning techniques learnt in this work can be applied on other datasets of similar characteristics, e.g. on the GAIA quick look low-resolution spectra and transient alerts, such as gamma-ray burst follow-ups.

References

Mickaelian, A. M., Nesci, R., Rossi, C., *et al.*, *A&A*, 464, 1177, 2007.

SDSS R12 webiste `http://sdss.org`

DFBS website `http://ia2.oats.inaf.it`

Bratsolis E., Bellas-Velidis I., Dapergolas, A., Kontizas E. and Kontizas M., Automatic extraction and classification of low-dispersion objective prism stellar spectra, *A&A*, Suppl. Ser., 2000

Ivezic, Z., Connolly, A. J., VanderPlas J. T., & Gray A., Statistics, Data Mining and Machine-Learning in Astronomy, Princeton University Press, 2014

https://doi.org/10.1017/S1743921317000266 Published online by Cambridge University Press