Unresolved Galaxy Classifier for ESA's GAIA Mission



I. Bellas-Velidis ⁽¹⁾, M. Kontizas ⁽²⁾, P. Tsalmantza ⁽³⁾, E. Livanou ⁽²⁾, E. Kontizas ⁽¹⁾, R. Korakitis ⁽⁴⁾, A. Dapergolas ⁽¹⁾, A. Karampelas ⁽²⁾, C.Á.L. Bailer-Jones ⁽³⁾, B. Rocca-Volmerange ^(5,6), A. Vallenari ⁽⁷⁾, M. Fioc ^(5,8)

- Institute of Astronomy & Astrophysics, National Observatory Athens, PO Box 20048, 11810 Athens, Greece (2) Department of Astrophysics, Astronomy & Mechanics, University of Athens, 15783 Athens, Greece
- (3) Max-Planck-Institut fur Astronomie, Konigstuhl 17, 69117 Heidelberg, Germany
- (4) Dionysos Satellite Observatory, National Technical University of Athens, 15780 Athens, Greece
- (5) Institut d'Astrophysique de Paris, 98bis Bd Arago, 75014 Paris, France
- (6) Universite de Paris-Sud XI, I.A.S., 91405 Orsay Cedex, France
- (7) INAF, Padova Observatory, Vicolo dell'Osservatorio 5, 35122 Padova, Italy

(8) Universite Pierre et Marie Curie, 4 place Jussieu, 75005 Paris, France

Abstract: The Unresolved Galaxy Classifier (UGC), an algorithm for galaxy spectra classification, are presented. It is implemented in Java as part of the GAIA mission's ground-based pipeline software and it is developed in the frame of the Data Processing & Analysis Consortium. It is provided to analyze low-dispersion spectra of unresolved galaxies that will be observed with GAIA's BP/RP instrument. UGC is based on Support Vector Machines (SVM), a supervised learning technique. The system is trained with "labeled" galaxy spectra with a priori known the galaxy type and the astrophysical parameter values. The SVM in classification mode is used for the galaxy type, whereas the regression mode is used for the parameters. The UGC training function includes three modules: tuning of the SVM parameters cost and gamma, the SVMmodels learning and testing their performance. A separate labeled set of spectra is used for the latter. The result of the training is a set of SVM-models, each one for a single parameter regression and one for the classification. The SVM-models are then used by the UGC application module. It is applied to unlabelled galaxy spectra and estimates the galaxy spectra and the parameter values. The data necessary to train and test the UGC are provided by GAIA Object Generator library of simulated galaxy BP/RP spectra. The library is based on a set of synthetic galaxy spectra produced by us using the Pegase software. We created synthetic spectra of galaxy models over a grid of predefined galaxy types and parameters. In the first stage of development there is used a subset of the library spectra, without interstellar reddening for single pass. The tests showed very good performance of the algorithm and its applicability to the task. It is already delivered and approved by Gaia DPAC. Development of next version is in progress.

Introduction

GAIA is ESA's cornerstone mission scheduled for launch late 2011. It is a scanning satellite that will repeatedly survey in a systematic way the whole sky during its sixyear mission. It will provide astrometric, photometrical and spectrophotometrical information of all point sources up to V=20, about one billion objects. Gaia science include: stellar structure and populations, Galactic structure and evolution, extrasolar planets, solar system, galaxies and quasars, general relativity.

Its double telescope feeds three instruments: the astrometry and G-band photometry field, the multichannel BP/RP spectrophotometer, and the highresolution radial-velocity spectrograph RVS. The images from these three fields are collected in specific parts of a mosaic of 106 CCDs working in time-delay integration mode. The onboard preprocessed output will consist of one-dimensional binned images of the detected and validated point sources profiles and their spectra. It is expected a 50GB/day data flow, resulting in about 100TB uncompressed science data during the mission. An extensive, sophisticated treatment is necessary to yield meaningful information from the original unintelligent GAIA's data.

A large pan-European team of expert scientists and software developers, the Data Processing and Analysis Consortium (DPAC), submitted to ESA's Announcement of Opportunity a proposal for a comprehensive system capable of handling the full size and complexity of GAIA data. The proposal (DPAC, 2007) was approved by ESA and DPAC became officially responsible for GAIA data processing and analysis. DPAC includes six large Data Processing Centres (DPC) and is structured around nine specialists units known as Coordination Units (CU), each in charge of a specific aspect of the data processing towards the final science product.

The goal of the CU8 "Astrophysical Parameters" is to provide for the observed objects classification information and estimation of specific astrophysical parameters to be included in the GAIA mission's intermediate and final database. Among these objects, it is expected to observe few millions of "unresolved" (as point sources) galaxies. The task of the Work Package CU8/WP32 "Unresolved Galaxy Classifier' (UGC) is to develop an algorithm and to implement it in Software Product for the DPAC's ground-based pipeline for classification and parameters estimation of these objects based on their BP/RP spectra.

Unresolved Galaxy Classifier Requirements

The requirements to the UGC task are defined by the factors forming the spectra emitted by the galaxies and registered by GAIA (Fig.1). An intrinsic galaxy spectrum is additionally modulated by the redshift (z), due to the galaxy radial velocity, it is influenced by the interstellar reddening (Av), its amplitude depends on the observed magnitude (Gmag) and finally it is deformed by the instrument response. It is not efficient, if possible at all, to provide an analytical solution for restoring the galaxy spectra. A promising solution is to implement a system based on supervised learning, trained with a properly selected set of template spectra.

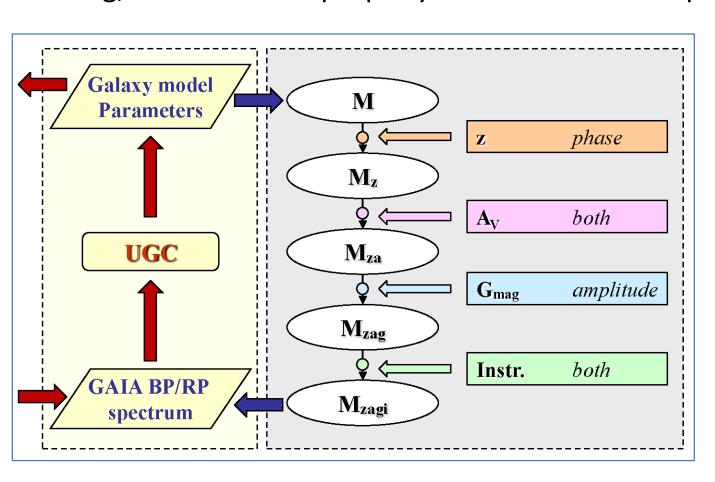


Fig. 1 The task of the UGC (left panel) based on our knowledge (right) how the galaxy spectra are formed from the intrinsic to the observed by GAIA BP/RP spectrophotometer.

Unfortunately, there is no library of observed or modeled galaxy spectra applicable to our task. We created libraries of synthetic spectra using the galaxy evolution model PEGASE.2 (Fioc M. & Rocca-Volmerange B. 1997). The model produces spectra based on a number of astrophysical parameters. Few of them are fixed by the taxonomic galaxy class (galType). From the remaining, up to four parameters (in dependence on the galType) are of primary importance (APs-gal) in modeling the spectra (Tsalmantza P., et al., 2007a). Fixing the galType to four classes and changing the APs-gal parameters in proper ranges, we created the second library of synthetic spectra that showed good coverage the two-color diagrams of observed galaxies (Tsalmantza P., et al., 2007b). These spectra were then used to produce the GAIA BP/RP simulated spectra library (for different z, Av, Gmag) UGClib2b.

The UGC system should be able to be trained with a library of template BP/RP spectra with apriori known parameters in order to provide classification and astrophysical parameters prediction for each observed galaxy spectrum during the GAIA mission. UGC should estimate the galaxy redshift and the interstellar reddening influencing the spectrum, as well.

UGC design and implementation

The Support Vector Machines (SVM), a supervised learning algorithm with nonlinear kernel function (Bennett K. & Campbell C., 2000) is used. A labeled data set containing the input (simulated galaxy spectra) and the output (apriory known specific parameter values corresponding to these spectra) is used for training. The SVM maps the input/output vectors to a higher dimensional space and creates a maximal separating hyperplane, SVM-model for this parameter. The SVM-model can then be applied to predict this parameter values when parsing unknown spectra. The "classification" mode of SVM is used to create the SVM-model for the galType parameter to classify the galaxy spectra whereas twelve models are constructed in "regression" mode for the APs-gal parameters value prediction. The SVM algorithm implemented in LIBSVM library (C-C. Chang & C-J. Lin, 2001) is used in UGC.

Four functions are provided by UGC (Fig.2). Specific SVM parameters are necessary to be tuned to get their optimal values. Then, the SVM-model can be created in the training process. Both, the tuning and the training, use one and the same labeled data set. The trained model is then tested with another labeled data set, not used in the training. These three functions should be applied for each galaxy parameter. The obtained SVM-models are finally used in the GAIA application function to classify unknown galaxy spectra and to estimate the corresponding galaxy parameters.

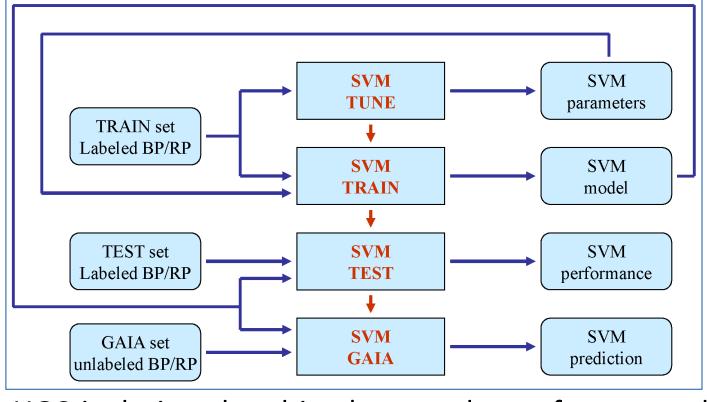


Fig. 2 Functional diagram of the UGCv3. The first three functions should be used in these sequence for each galaxy parameter, the fourth can be used after completing them.

UGC is designed and implemented as software product in Java following the DPAC requirements (Levoir T. et al., 2007) and guidelines (O'Mullane et al., 2007). The UGC functions are separate tasks (modules), that are accessed through a checking module (Fig.3). The tuning, training and testing tasks works with auxiliary labeled galaxy spectra simulated by synthetic and/or semi-empirical libraries. The GAIA application task works with the unlabeled spectra from the GAIA pipeline that are already classified as galaxy spectra by another CU8 package.

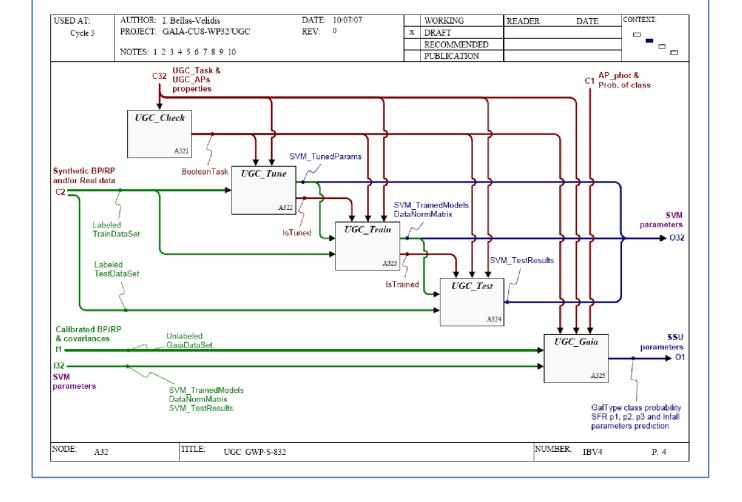


Fig. 3 SADT diagram illustrating the modular structure of the UGCv3. The modules dependence and the I/O are shown as well.

In the UGC development we are following the DPAC approach of 6-month cycles, starting from a very simple implementation towards the final complete software product. Here we describe the Cycle3 UGCv3 software. It is mainly intended to test the applicability of the SVM algorithm to galType classification and APs-gal parameterization of unresolved galaxy spectra.

UGCv3 Results

In this cycle we limited UGC to GAIA BP/RP simulated noiseless spectra of galaxy models of fixed magnitude, without redshift and without reddening, registered in a single pass. There are 28885 such spectra in the UGClib2b library representing the four galType models. About 2800 represents "early-type" galE models obtained with random variation (in a specific range) of only two APs-gal parameters, Ep1 and Ep2. There are about 10500 "spiral-type" gals spectra with random values of three parameters (Sp1, Sp2, Sp3).and 1500 spectra are for "irregular-type" gall models with another set of parameters (Ip1, Ip2, Ip3). Finally, 14000 spectra represent "starburst-type" models with variation of four parameters, Bp1, Bp2, Bp3 and Bp4.

We tuned, trained and tested one SVM-model in classification mode for the galType and twelve in regression mode for the corresponding APs-gal parameters. In each case, the total set of applicable library spectra was randomly split in two equal subsets a training and a testing one. The performance of each SVM-model is analyzed using the results of an "internal" test (with the training set) and of the real test with the testing set. These results are presented in the following two tables.

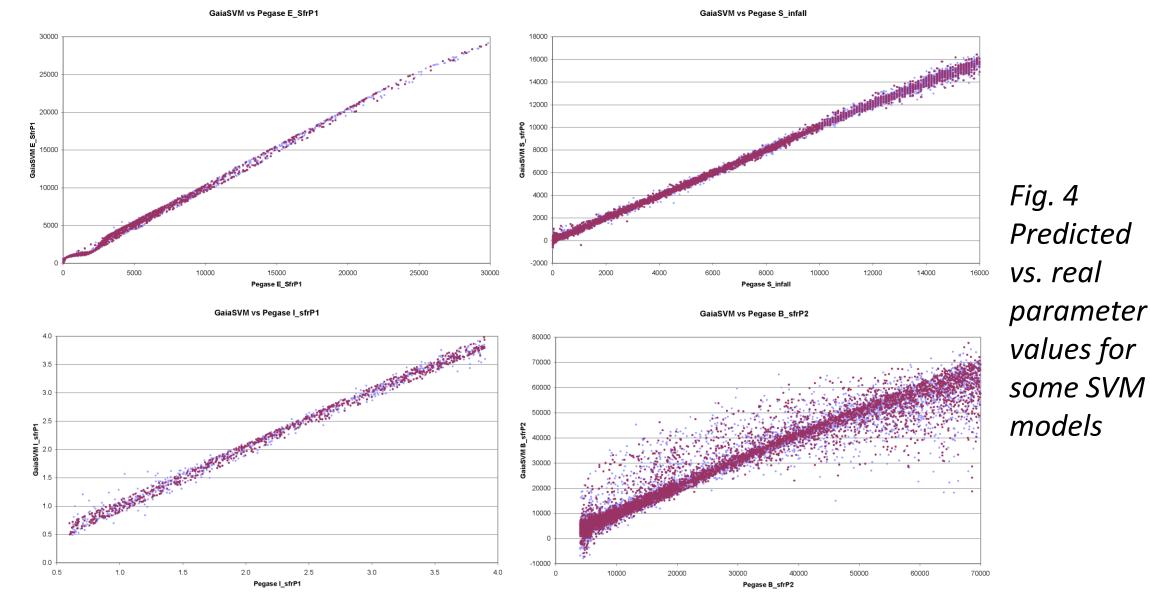
Galaxy	galT	ype classi	fier SVM-	-model	galType classifier SVM-model				Train & Test	
spectra	performance in % of the train-data				performance in % of the test-data				sets number	
class	galE	galS	galI	galB	galE	galS	galI	galB	of spectra	
galE	99.81	0.19	0.00	0.00	99.67	0.33	0.00	0.00	1408 1408	
galS	0.00	100.00	0.00	0.00	0.00	100.00	0.00	0.00	5284 5285	
galI	0.00	0.00	100.00	0.00	0.00	0.15	99.85	0.00	750 750	
galB	0.00	0.00	0.00	100.00	0.00	0.00	0.00	100.00	7000 7000	

Internal test accuracy = 99.98% External test accuracy = 99.97%

The table above shows the confidence matrices in both the internal and the external tests of the galType SVM-model. It is evident the very good performance in both cases to predict the galaxy type.

	SVM			Performance					
	model			TRAII	N task	TEST task			
Id	cost	γ	Name	Units	Range	StDev	Error	StDev	Error
Ep1	27	2^{-4}	E_SfrP1	My	10-30000	0.067	371	0.067	375
Ep2	2^{15}	2^{-7}	E_SfrP2	M_{\odot}	0.2-1.5	0.060	0.021	0.059	0.021
Sp0	2^{17}	2^{-10}	S_infall	My	5-16000	0.043	203	0.043	203
Sp1	2^{10}	2^{-3}	S_SfrP1	-	0.3-2.4	0.119	0.069	0.148	0.086
Sp2	2^{19}	2^{-7}	S_SfrP2	My/M⊙	5-30000	0.168	1465	0.181	1574
Ip0	2^{17}	2^{-5}	<i>I_infall</i>	My	5000-30000	0.129	843	0.192	1261
Ip1	2^{13}	2^{-4}	I_SfrP1	-	0.6-3.9	0.065	0.064	0.092	0.091
Ip2	2^{15}	2^{-5}	I_SfrP2	My/M⊙	4000-70000	0.141	2894	0.146	3002
B p0	2^{17}	2^{-5}	B_infall	My	5000-30000	0.246	1672	0.279	1897
Bp1	2^{17}	2^{-6}	B_SfrP1		0.6-3.9	0.106	0.100	0.120	0.113
Bp2	2^{14}	2^{-7}	B_SfrP2	My/M⊙	4000-70000	0.236	4848	0.252	5180
<i>Bp3</i>	2^{14}	2^{-7}	B_SfrP3	My	1-250	0.049	3.46	0.051	3.55

The overall results for the performance of the twelve regression SVM-models are shown above via the standard deviation of the differences between the real and the predicted values of the parameter (normalized) and the corresponding error in parameter's units. The values indicate a very good performance. In the case of Sp2, Bp0 and Bp2 it is likely to improve it with better tuning. This is also illustrated on figures below (Fig. 4), showing the predicted vs. real values of selected parameters.



Discussion

The performance of the SVM-models is tested and shows that the SVM algorithm is applicable for the requirements of the UGC software package. Improvements in the SVM-parameters tuning and models training are necessary in few cases. Having in mind the UGCv3 limitations, tests should be carried out with noisy spectra in a range of Gmag. The next steps will include development of algorithms considering the redshift and the reddening. We also plan to test the implementation of Artificial Neural Networks algorithm in UGC.

References

K. P. Bennett, C. Campbell, 2000, Support Vector Machines: Hype or Hallelujah?, SIGKDD Explorations, Vol. 2, 12

Chich-Chung Chang & Chich-Jen Lin, 2001, LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

DPAC, 2007, Proposal for the Gaia Data Processing, GAIA-CD-SP-DPAC-FM-030-2 Fioc M. & Rocca-Volmerange B., 1997, PEGASE: a UV to NIR spectral evolution model

of galaxies, Astron. Astrophys., v. 326, 950 T. Levoir, J-C. Damery, J. Hoar, F. Jocteur-Monrozier, T. Lock, A-T. Nguyen, W. O'Mullane, P. Parsons, 2007, Gaia DPAC. Product Assurance Plan, GAIA-C1-SP-CNES-

W. O'Mullane, J. Hoar, T. Levoir, F. De Angeli, A-T. Nguyen, 2007, Software

Engineering Guidelines for DPAC, GAIA-C1-UG-ESAC-WOM-011-1

P. Tsalmantza, Kontizas M., C. A. L. Bailer-Jones, B. Rocca-Volmerange, R. Korakitis, E. Kontizas, E. Livanou, A. Dapergolas, I. Bellas-Velidis, A. Vallenari, M. Fioc, 2007a, Towards a library of synthetic galaxy spectra and preliminary results of classification and parametrization of unresolved galaxies for GAIA, Astron. Astrophys., v. 470, 761 P. Tsalmantza, B. Rocca-Volmerange B., M. Kontizas, I. Bellas-Velidis, E. Kontizas, C. A. L. Bailer-Jones, R. Korakitis, A. Dapergolas, E. Livanou, M. Fioc, A. Vallenari, 2007b, The second library of synthetic galaxy spectra with Pegase2 code for GAIA purposes, GAIA-C8-TN-UOA-PAT-002