Unresolved Galaxy Classifier for ESA/Gaia mission: Support Vector Machines approach

Ioannis Bellas-Velidis[1], Mary Kontizas[2], Anastasios Dapergolas[1], Evdokia Livanou[2], Evangelos Kontizas[1], Antonios Karampelas[2]
[1] Institute for Astronomy and Astrophysics, National Observatory of Athens, PO Box 20048, 11810, Athens, Greece
[2] Department of Astrophysics, Astronomy & Mechanics, Faculty of Physics, University of Athens, 15783, Athens, Greece
ibellas@astro.noa.gr
(Submitted on 26.03.2012; Accepted on 29.04.2012)

Abstract. A software package Unresolved Galaxy Classifier (UGC) is being developed for the ground-based pipeline of ESA's Gaia mission. It aims to provide an automated taxonomic classification and specific parameters estimation analyzing Gaia BP/RP instrument low-dispersion spectra of unresolved galaxies. The UGC algorithm is based on a supervised learning technique, the Support Vector Machines (SVM). The software is implemented in Java as two separate modules. An offline learning module provides functions for SVM-models training. Once trained, the set of models can be repeatedly applied to unknown galaxy spectra by the pipeline's application module. A library of galaxy models synthetic spectra, simulated for the BP/RP instrument, is used to train and test the modules. Science tests show a very good classification performance of UGC and relatively good regression performance, except for some of the parameters. Possible approaches to improve the performance are discussed.
Key words: Methods: data analysis; Techniques: miscellaneous; Galaxies: general

Introduction

Gaia is a cornerstone mission of ESA, scheduled for launch in 2013. It is a scanning satellite that will repeatedly survey in a systematic way the whole sky during its six-year mission. The scientific payload will provide astrometric, photometrical and spectroscopic 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 AF, the BP/RP spectrophotometer, and the high-resolution radial-velocity spectrograph RVS. The images from these three fields are collected in particular parts of a mosaic of 106 CCDs that works in a time-delay integration mode. The on-board preprocessed output will consist of one-dimensional binned images of the detected and validated point-like sources profiles and their spectra. It is expected a 50 GB/day data flow, resulting in about 100 TB uncompressed science data during the mission. An extensive, sophisticated treatment is necessary to yield meaningful information from the original unintelligent data.

A large pan-European team of expert scientists and software developers, the Data Processing and Analysis Consortium (DPAC), submitted a proposal for a comprehensive ground-based system, capable of handling the full size and complexity of Gaia data. The proposal (DPAC, 2007) was approved by ESA and the DPAC became officially responsible for Gaia-mission data processing and analysis. DPAC includes six large Data Processing Centers (DPC) and is structured around nine 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 an estimation of specific astrophysical parameters to be included in the missions intermediate and final database. Among these objects, it is expected few millions of unresolved (point-like) galaxies to be observed. The main task of the top-level work-package CU8/WP832 "Unresolved Galaxy Classifier" (UGC) is to develop an algorithm for classification and parametrisation of the unresolved galaxies and to implement it in a software system, part of the DPAC's ground-based pipeline.

The requirements for the UGC functionality are described in the first section. The data library used for the development is presented in the second section. The currently developed algorithm and its implementation in software modules are shortly presented in the next section. In the fourth section, the classification and regression performance of the application module are discussed. Conclusions derived from the tests and key-points for further development of UGC are given in the last section.

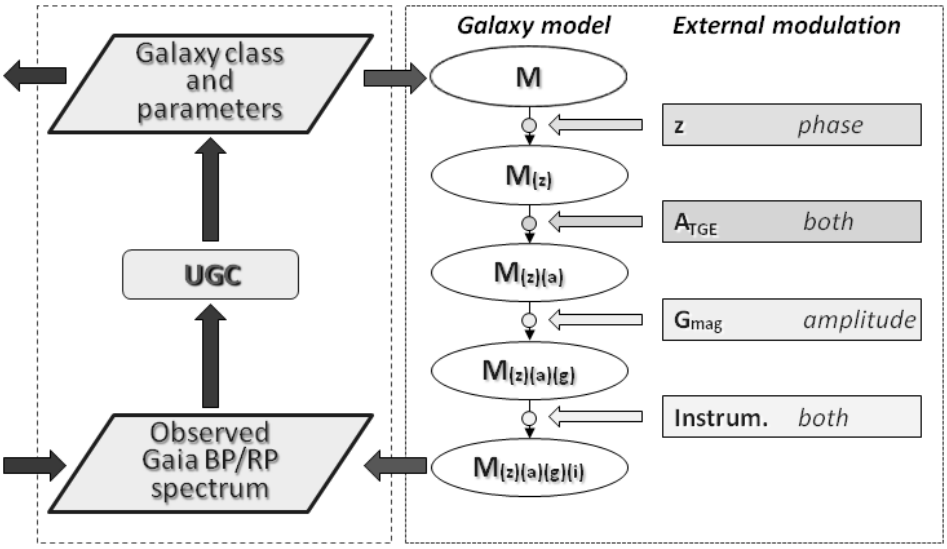

Fig. 1. The main task of the Unresolved Galaxy Classifier (left panel) and a model of the external factors influencing the observed galaxy spectra (right panel).

1. Galaxies spectra classification and regression requirements

The UGC system shall be able to provide for each galaxy, observed during the Gaia mission, a taxonomy classification and a particular astrophysical parameters estimation. This should be based on the galaxies spectra observed with the low-resolution Gaia BP/RP instrument. The requirements to the UGC task are set by the factors forming the spectra emitted by the galaxies

and registered by Gaia (Fig.1). An intrinsic galaxy spectrum is modulated by the source's redshift and by the total extinction in our Galaxy (TGE) towards the source. The spectrum amplitude depends on the objects distance, the observed magnitude, and is deformed by the instrumental response and noise.

The requirements that UGC must meet are additionally set by the algorithm selected to be implemented. It is not possible to provide an analytical solution to directly restore and parametrise the registered spectra. A promising solution is to implement an Artificial Intelligence system based on supervised learning. Such a system is firstly trained with templates, simulated spectra (inputs) of galaxy models with a priori known parameters (expected outputs). Then, the system can be applied repeatedly to classify unknown galaxies and to estimate their parameters from the observed spectra. The problem here is to provide simulations as close as possible to the really observed spectra. Unfortunately, there is no library of observed or modeled galaxy spectra suitable for such a task, so it is required to create proper spectral libraries for UGC.

Finally, implementation of such a system in a software has to follow the DPAC requirements for the ground-based pipeline, the Software Engineering Guidelines (O'Mullane et al., 2011). The UGC is being developed in Java[3] language under the Eclipse[4] Integrated Development Environment. The DPAC approach of 6-month cycles, is being followed in the UGC development, starting from a very simple implementation towards the final, complete software product. The Cycle10 software release UGCv10, as of June 2011, is presented here. It is mainly intended for offline science tests of an optimized algorithm implemented in UGC, as well as for online performance tests of UGC within the simulation of the DPAC pipeline in one of the DPCs.

2. Library of galaxy models synthetic spectra

Accordingly to the UGC requirements, we created libraries of synthetic spectra (Tsalmantza et al., 2007, Tsalmantza et al., 2009, Karampelas et al., 2011) using the evolutionary synthesis of galaxy spectra provided by "Pegase2" model (Fioc & Rocca-Volmerange, 1997). The model produces spectra based on a number of astrophysical parameters. Most of them, like galaxy age, have been fixed by adopting four taxonomic galaxy classes *galType*: gal_E (early), gal_S (spiral), gal_I (irregular), and gal_B (quenched star-formation galaxies).

The *galType* also fixes the star-formation rate, SFR, law used and the meaning of the parameters. The exponential law (1) has been used for gal_E type, whereas (2) has been applied for the other three galaxy types with the SFR quenching limit (3) additionally applicable only to the ga_lB class:

$$SFR(t) = (P_2/P_1) * e^{-t/P_1} \quad (1)$$

$$SFR(t) = (1/P_2) * M_{gas}(t)^{P_1} \quad (2)$$

$$SFR(t) = 0 \qquad if\ (Age - t) \leq P_3 \quad (3)$$

[3] http://java.com
[4] http://www.eclipse.org

These SFR parameters, for the corresponding types, and the gas infall time (excluding the early type), are found to be of primary importance in modeling the spectra (Tsalmantza et al., 2007) and are grouped in UGC under the *galPar* name (see Tab.1).

Table 1. Intrinsic galaxy model parameters *galPar* for each *galType* class.

	galType			
	gal_E	***gal_S***	***gal_I***	***gal_B***
galPar		*S_infall*	*I_infall*	*B_infall*
	E_sfrP1	*S_sfrP1*	*I_sfrP1*	*B_sfrP1*
	E_sfrP2	*S_sfrP2*	*I_sfrP2*	*B_sfrP2*
				B_sfrP3
SFR law	(1)	(2)	(2)	(2 and 3)

Fixing the classes and varying the corresponding parameters within proper ranges, we created a large set of synthetic galaxy spectra that showed good coverage of the two-color diagrams of observed galaxies (Tsalmantza et al., 2009). This base set of synthetic spectra has been extended reproducing it for different redshift values z within 0.0 - 0.2 range and for random values of the coefficient Ao (range 0.0 - 6.0, equally distributed in log-scale) of the total galaxy extinction (TGE) towards the source. The latter is based on the extinction law (Cardelli et al., 1989) where Ao is the first coefficient (close to A_V for early stars) and the second is fixed to Ro=3.1.

The extended set of synthetic spectra has then been used to produce the Gaia BP/RP simulated spectra library. The simulations are provided by CU2 group with their Gaia Object Generator package, GOG (Isasi et al., 2010). A representative set of synthetic spectra and their BP and RP simulations, in the very first library, are illustrated in Fig.2.

In the UGCv10 release, the GOG7 library of galaxy spectra has been used. It is simulated for three Gaia magnitudes (G_{mag}=15.0, 18.5 and 20.0). Instrumental and calibration noise is properly applied and the spectra are averaged for a multi-epoch observations simulation with the number of transits varying from 40 up to 80.

3. UGC algorithm and implementation

The UGC system shall provide for each galaxy spectrum observed during the Gaia mission a taxonomy classification *galType* class-probability and the astrophysical *APs_gal* parameters prediction. The latter include the galaxy model's star formation parameters, *galPars* (Tab.1), and the two external parameters influencing the intrinsic galaxy spectrum, the redshift z and the TGE coefficient Ao.

A supervised learning algorithm, Support Vector Machines (SVM), is used. It is applicable to both, the classification (Cortes & Vapnik, 1995) and

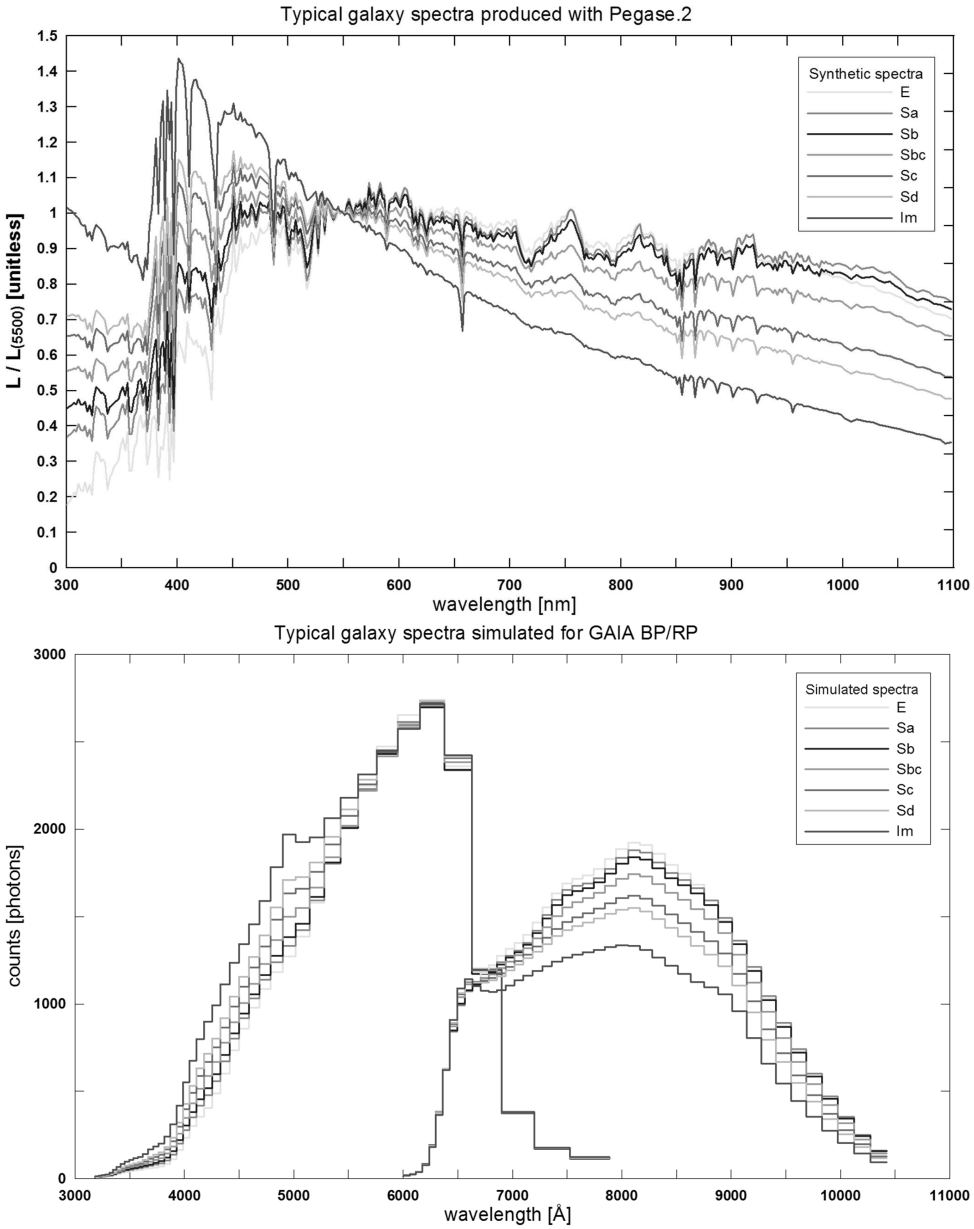

Fig. 2. Top: Pegase.2 synthetic spectra of representative galaxy types (emission lines, if any, not included); Bottom: The synthetic spectra (emission lines included) simulated for Gaia BP and RP instrument without applying redshift, extinction and instrumental noise.

the parameter regression problems (Drucker et al., 1997). A "labeled" data set is necessary for the learning procedure. In a parameter regression, the training set contains the input data (simulated galaxy spectra) and the desired outputs, the priori known parameter values used to create these spectra. The SVM maps the input/output vectors to a higher dimensional space applying a nonlinear kernel function 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. On the other hand, in the multiclass classification problem, the training output is a numerically coded taxonomic class. In this case, the trained SVM-model applied to an unknown spectrum estimates the probabilities the source to belong to the specific class. The classification mode of SVM is used to create the SVM-model for the *galType* parameter to classify the galaxy spectra whereas a model per parameter shall be constructed in regression mode for each of the APs_gal parameters value prediction.

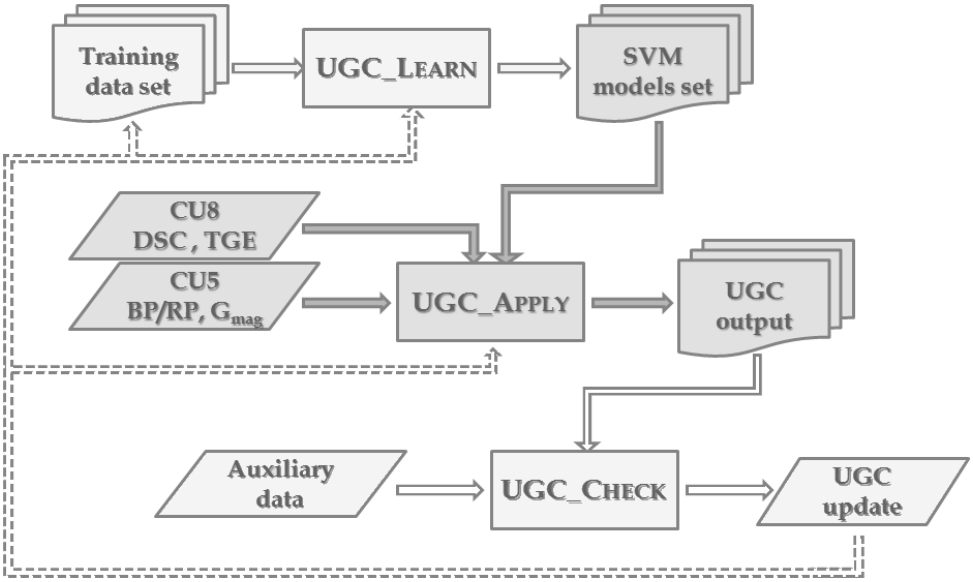

Fig. 3. The learning, application and checking functions of the Unresolved Galaxy Classifier (the dark-shaded elements are parts of the Gaia ground-based data processing pipeline).

The SVM algorithm implemented in LIBSVM library (Chang & Lin, 2007) is used in UGC. The current implementation, UGCv10 (Fig.3), consists of two modules, UGC_Learn and UGC_Apply, that provide the learning and the application function, correspondingly. One more module, not described here, will be provided for furthermore automated checking the UGC performance in order to analyse and, if necessary, to update the algorithm during the mission data processing cycles. The learning is used to prepare offline and to provide the "trained" SVM-models. On the other hand, the application

function is being implemented as part of the DPAC ground-based pipeline for Gaia science data processing.

3.1. Learning module

Three tasks are provided by *UGC_ Learn* module to prepare the SVM-models, one model at a time. Specific parameters of the SVM are necessary to be tuned to ensure the model's optimal performance. Then, the SVM-model is created in the training process and is saved. Both, the tuning and the training, use one and the same labeled data set (galaxy spectra with priori known class and parameters). The trained SVM is finally tested with another labeled data set for performance estimation. The sequence of these three tasks is applied offline to prepare SVMs for the classification and for each of the galaxy parameters resulting in a set of fifteen SVM-models. Moreover, such sets are obtained for particular source magnitudes and for different ranges of the external parameters influencing the galaxy spectra (see below).

In order to improve the UGC performance, following test results from a previous implementation (Bellas-Velidis et al., 2010), it has been decided to "split" the total range of the two external parameters that influence the intrinsic spectra of galaxy models. In addition to previously fixed three ranges in sources magnitude G_{mag}, four sub-ranges for the redshift z and five sub-ranges for the extinction coefficient Ao (see Tab.2) have been specified.

Table 2. Application ranges defined for SVM-models and their coding. The "total ranges" are defined only for the redshift z and for the extinction coefficient Ao initial estimation.

G_{mag}	z	Ao
G150 13.0-16.0	*Z0020* 0.00-0.20	*A0060* 0.0-6.0
G185 16.0-19.0	*Z0005* 0.00-0.05	*A0005* 0.0-0.5
G200 19.0-20.0	*Z0410* 0.04-0.10	*A0010* 0.0-1.0
	Z0915 0.09-0.15	*A0520* 0.5-2.0
	Z1420 0.14-0.20	*A1535* 1.5-3.5
		A3060 3.0-6.0

The total combination of the defined ranges required to tune, train and test 1350 SVM-models. The first tests showed that galaxy parameters regression is not effective for the faintest magnitudes. So well, it has been found that the smallest extinction range SVMs does not provide better accuracy than the next one. Finally, there is a redundant: the combination of the total range for z with the sub-ranges for Ao and vice versa. Following this, the total number of SVM-models, that are really usable and necessary for the UGC application, has been reduced to 534.

3.2. Application module

The *UGC_ Apply* module is intended to run online as part of the earth-based Gaia data processing pipeline. The application module will be activated every

time when a galaxy spectrum is identified. This identification is provided by another software package of the pipeline, the Discrete Source Classifier, DSC (Smith et al., 2011). The *UGC_ Apply* runs a sequence of two tasks: selection and processing. Firstly, it will select the spectrum if it is validated for its suitability as galaxy or galaxy-like, based on predefined criteria. Only when validated, the spectrum will be processed by the second task.

The *UGC_ Apply* processing function is illustrated in Fig.4. A specific pre-processing is applied to the spectrum, converting it to a proper form as required by the trained SVMs. Then, based on the source's magnitude G_{mag}, it is fixed the proper set of SVM-models to be used. Follows an initial estimation of the two external parameters z and Ao. The corresponding two SVMs trained for the total range of these parameters are applied to the source to provide their first estimate.

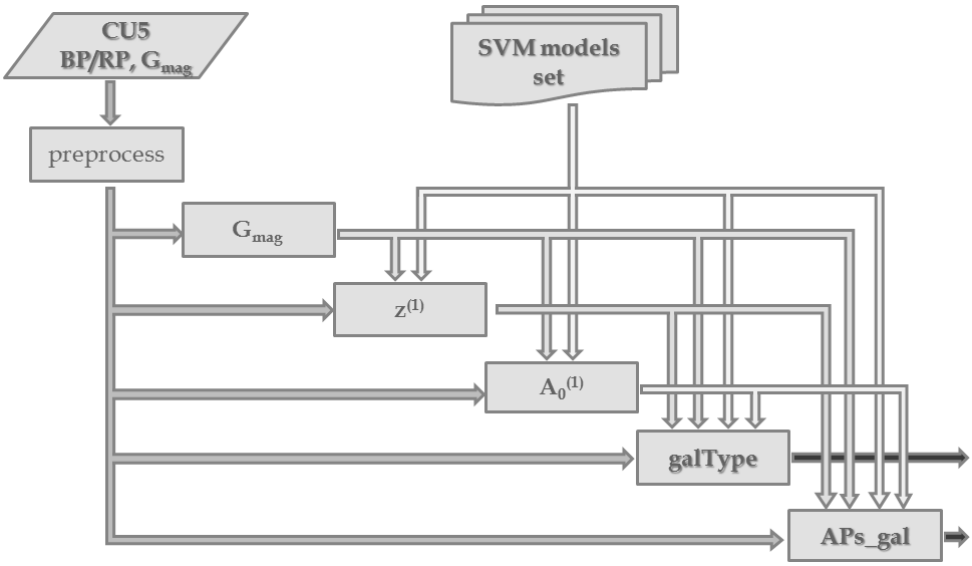

Fig. 4. The processing task of the Unresolved Galaxy Classifier application module *UGC_ Apply*. This task is applied only if the source has been already validated as galaxy or galaxy-like by the front-end selection task of the application module.

These initial values are now used to fix the applicable subset of specific-range SVM-models. Such a subset has been trained to provide classification and parameters regression for spectra influenced by the redshift and the extinction varying within a specific part of their total range (Tab.2). The corresponding SVMs from the subset are then used to perform *galType* classification and all the *APs_ gal* astrophysical parameters regression (including a final estimation of the two external parameters). Because of sub-ranges partially overlapping, there can be applicable one, two, or four subsets of SVM-models. In the latter two cases, a proper weighting of the estimates is provided. Finally, the results are output for saving in the Gaia Main Database.

4. UGCv10 performance

UGC performance for galaxy classification and for the two external parameters regression is tested with one and the same basic set of synthetic spectra simulated for three magnitudes (G_{mag}=15.0, 18.5, and 20.0) of the sources; the performance for the galaxy internal parameters has been tested for the first two magnitudes only. This basic set is created for four predefined galaxy types, varying only specific galaxy model parameters, and applying random redshift z and extinction Ao within predefined ranges (see Section 2).

The UGC application module has been tested on 70233 galaxy spectra. On a computing node of about 0.5GFlops/s the pure CPU time was 4600s, leading to about 0.066s or 34MFlops per source. About 1.5GBytes Java heap space is required for UGC application module to run.

4.1. Galaxy type classification performance

UGC provides for each source a class-probability vector with four elements, the normalized probabilities (sum to be unit) for the source to belong to each of the predefined galaxy types. The element corresponding to the source's real class contains the so-called True-Positive (TP) probability. If one of the four probabilities is above 0.5 the galaxy is considered positively classified. If all the probabilities are below 0.5 the source class is undefined. In the case of positive classification, if the probability corresponds to the real (a priori known) class of the galaxy, it is considered True-Positive classification, or TPC, else it is counted as False-Positive classification (FPC).

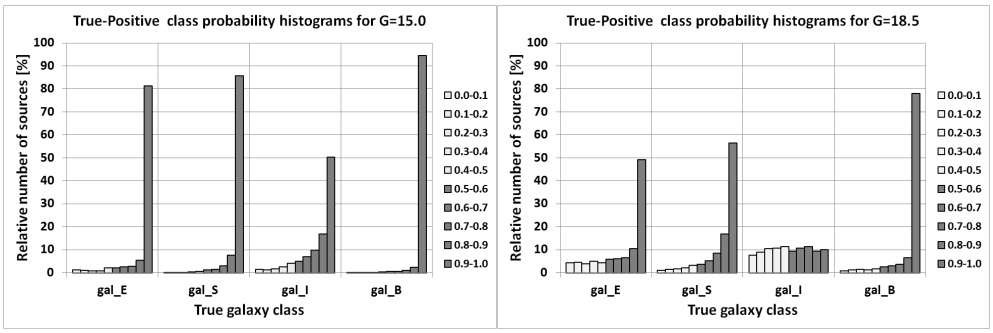

Fig. 5. The UGC performance for *galType* classification in the G_{mag}=15.0 and 18.5 sources tests. For each class, the percentage of the cases with True-Positive probability above 0.5 (dark-shaded bars) gives the TP classification performance for the class.

The distribution of the TP class-probability estimates in the two magnitude tests (G_{mag}=15.0 and 18.5), for each of the four subsets of spectra corresponding to a predefined galaxy class, is presented in Fig.5. The cases of TPCs (TP probabilities > 0.5) are shown dark-shaded. In the faintest magnitude test (G_{mag}=20.0), not presented here, almost all the counts for the different probability ranges are below 20%.

Table 3. UGC classification performance given by the percentage of True-Positive (bold) and False-Negative classifications for the three testing sets of spectra with different G_{mag}.

Galaxy model		UGCv10 classification performance											
Real	Sources	G_{mag}=15.0				G_{mag}=18.5				G_{mag}=20.0			
galType	number	**E**	**S**	**I**	**B**	**E**	**S**	**I**	**B**	**E**	**S**	**I**	**B**
gal_E	1620	**94.1**	4.9	0.0	0.0	**78.1**	19.1	0.0	0.3	**44.8**	34.4	0.0	7.0
gal_S	6851	0.6	**98.7**	0.2	0.0	3.9	**90.6**	0.1	2.7	5.9	**67.4**	0.0	15.9
gal_I	1107	0.0	5.7	**88.9**	2.2	0.0	8.4	**50.8**	26.5	0.0	11.6	**30.2**	42.6
gal_B	8817	0.0	0.2	0.5	**99.1**	0.1	2.0	2.5	**93.7**	0.4	7.3	1.9	**83.4**

The classification performance is usually presented by so-called confusion matrix where the elements show the percentage of objects of a given class that have been positively classified. The diagonal elements show the percentage of the objects with True-Positive classifications for the corresponding real class, whereas the other elements on each row count the FPC's. The UGC performance in the three magnitude tests is presented in Tab.3.

As this can be seen from Fig.5 and Tab.3, the UGC classification performance is very good for the bright sources (G_{mag}=15.0). The percentage of TP classifications is very close to 100% for the two best classified types, *gal_S* and *gal_B*, it is about 95% for the *gal_E* and a little below 90% for *gal_I* type. The classification is acceptable for faint sources (G_{mag}=18.5), providing TPC above 90% for *gal_S* and *gal_B*, around 80% for *gal_E* and falling to around 50% for *gal_I*. Finally, for the faintest sources (G_{mag}=20.0), it is still acceptable for the best classified *S* and *B*-type, but there is a problem for the other two types. It seems that in this case we can still provide a two-class classification combining *E* with *S* in an "early" galaxy type, and *I* with *B* in a "late" type. On the other hand, the relatively worse performance for *gal_I* could be caused by the small number of simulated spectra of this type.

4.2. External parameters regression performance

The two parameters, z and Ao, are estimated by UGC in a two-stage process (see Section 3.2). Initially, they are estimated applying the corresponding magnitude SVM-models that has been trained for the total range of these parameters. The initial estimate is used for fixing which SVM-models, trained for specific sub-ranges of these parameters (Tab.2) shall be used for final estimation. The UGC performance in these two cases is presented in Tab.4 by the Root-Mean Square Deviation (RMSD) of the estimated from the real (a priory known) value in the testing set of spectra. The corresponding normalized value, NRMSD, in percents of the parameter range is also given for the final estimation. The performance improvement is evident in the final versus the initial estimation for both parameters and that this improvement is greater towards the brighter sources. This general performance is even better illustrated in Fig.6 plotting the finally estimated versus the real values for these parameters for the three magnitude tests. In each figure there is an inset showing the histogram of the differences (estimated minus real value).

Table 4. UGC regression performance for the external parameters initial and final estimation given by the Root-Mean Square Deviation for the three sets of spectra with different G magnitude. For the final estimation, the range-normalized, NRMSD, is also shown.

Galaxy model		UGCv10 regression performance								
External	Parameter	G_{mag}=15.0			G_{mag}=18.5			G_{mag}=20.0		
APs_gal	Range	initial	final		initial	final		initial	final	
z	0.0 - 0.2	0.005	**0.002**	**(1.0%)**	0.015	**0.011**	**(5.5%)**	0.030	**0.026**	**(13.0%)**
Ao	0.0 - 6.0	0.08	**0.04**	**(0.7%)**	0.18	**0.15**	**(2.5%)**	0.30	**0.29**	**(4.8%)**

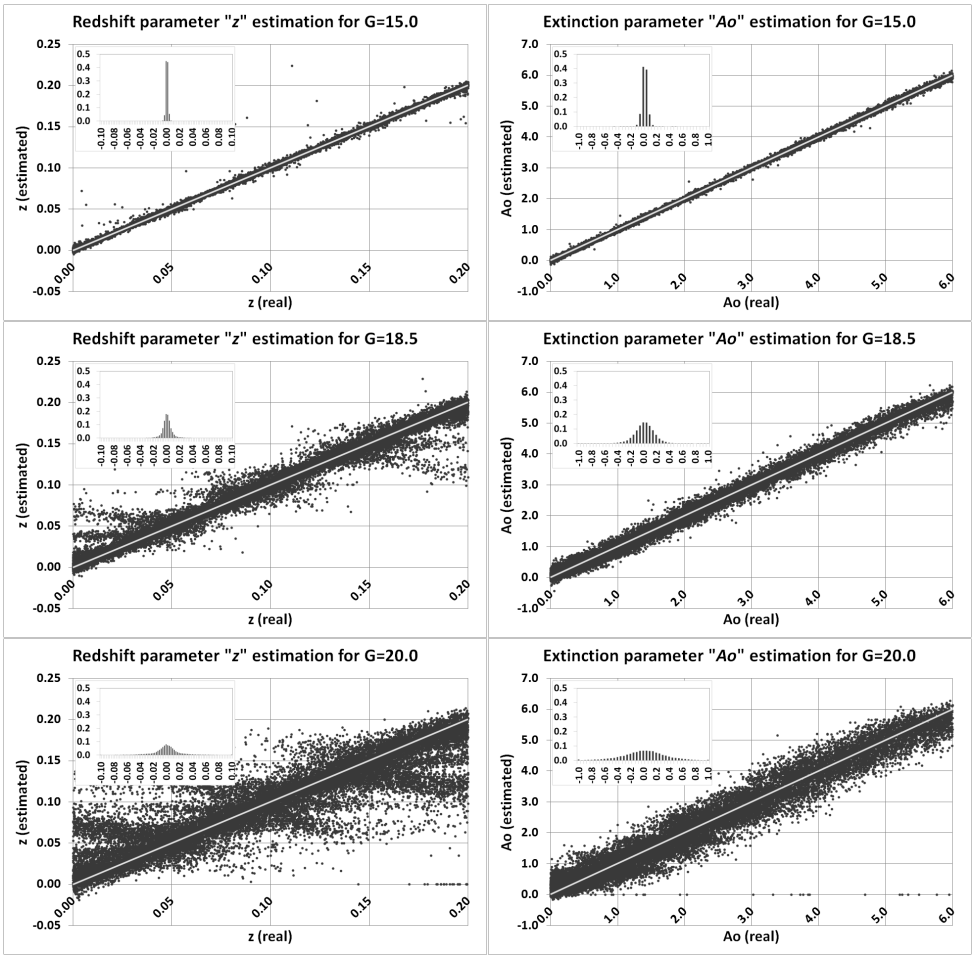

Fig. 6. Redshift "z" (left panel) and extinction "Ao" (right panel) estimation performance for G_{mag}=15.0, 18.5 and 20.0 source sets. Each inset shows the corresponding histogram of the differences between the UGC estimated and the real values.

The UGC very good performance for extinction coefficient Ao estimation, for all the magnitudes up to the faintest sources is clearly shown in Fig.6 (right panel). It is also very good for the redshift z (left panel) for the bright sources and it is relatively good for G_{mag}=18.5, but in the latter case there is a number of systematically deviated estimations. for the faintest sources this pattern of deviations isF quite extended, causing the estimation to be doubtful here. An investigation is necessary to find the reasons leading to the performance degradation.

4.3. Galaxy model parameters estimation performance

The UGC performance for the galaxy model parameters regression has been found unacceptable for the faintest magnitudes in earlier tests, so SVM-models have not been trained for the case. The performance, the RMSD and the corresponding NRMSD, is presented in the Tab.5 for the other two G_{mag} tests (15.0 and 18.5). For the parameters of each particular *galType* model, only the corresponding spectra subset has been tested. Normalized RMSD value above 20% indicates rather doubtful performance for the parameter regression.

Table 5. UGC regression performance for the *galPar* parameters estimation given by the Root-Mean Square Deviation and the range-normalized NRMSD for two testing sets of spectra with different G magnitudes (unacceptable performance is marked by an asterisks).

Galaxy model				UGCv10 performance	
galType	*galPar*	Range	Units	G_{mag}=15.0	G_{mag}=18.5
gal_E	***E_sfrP1***	10 - 14400	My	696 (4.8%)	1459 (10.1%)
	E_sfrP2	0.20 - 1.45	$\mathcal{M}_\odot$	0.06 (4.8%)	0.16 (12.8%)
gal_S	***S_sfrP1***	0.30 - 2.40		0.56 *(26.7%)	0.60 *(28.6%)
	S_sfrP2	5 - 30000	My/$\mathcal{M}_\odot$	3521 (11.7%)	4979 (16.6%)
	S_infall	5 - 16000	My	1336 (8.4%)	2714 (17.0%)
gal_I	***I_sfrP1***	0.60 - 3.90		0.61 (18.5%)	0.68 (20.6%)
	I_sfrP2	4000 - 70000	My/$\mathcal{M}_\odot$	14400 (21.8%)	18048 *(27.3%)
	I_infall	5000 - 30000	My	5759 *(23.0%)	6143 *(24.6%)
gal_B	***B_sfrP1***	0.60 - 3.90		0.56 (17.0%)	0.74 (22.4%)
	B_sfrP2	4000 - 70000	My/$\mathcal{M}_\odot$	11881 (18.0%)	17250 *(26.1%)
	B_sfrP3	1 - 150	My	15.03 (10.1%)	29.6 (19.9%)
	B_infall	5000 - 30000	My	5828 *(23.3%)	6113 *(24.5%)

The regression performance for estimation of the intrinsic parameters of the four galaxy model types is illustrated in the panels of Fig.7 and Fig.8. Each panel shows the estimated versus the real value of the particular parameter in the two magnitude tests. The UGC shows a good performance for the two parameters of the *gal_E* model in both magnitude tests, with the exception of *E_sfrP1* real values very close to zero or near their upper limit.

Relatively good is the UGC performance for the *gal_S* parameters with exception of the *S_sfrP1*. Even worse, in the latter case it is clear that the

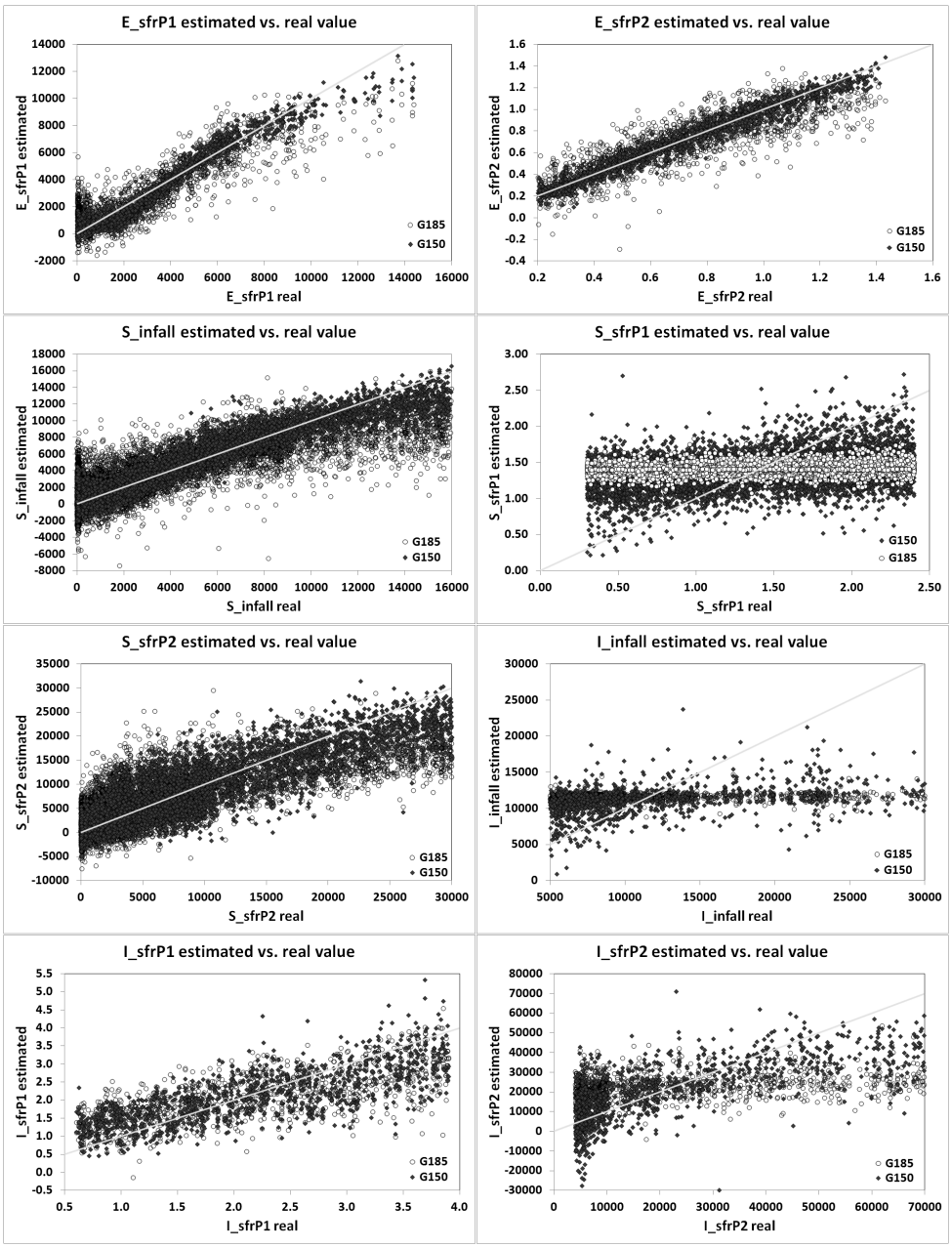

Fig. 7. UGC regression performance for the *gal_E*, *gal_S* and *gal_I* types parameters *galPar* for the G_{mag}=15.0 and 18.5 source sets.

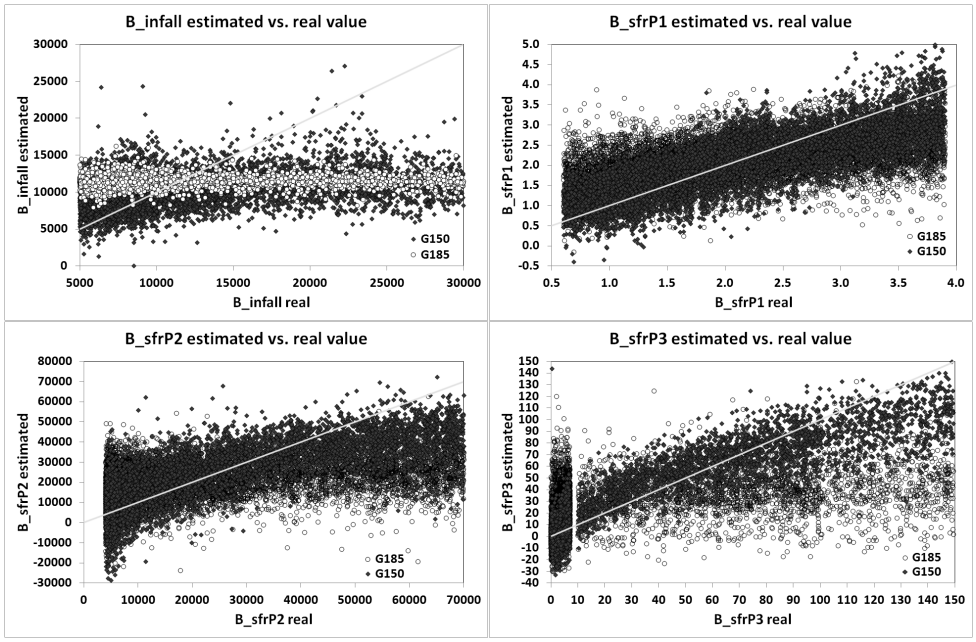

Fig. 8. UGC regression performance for the *gal_ B* type *galPar* parameters.

UGC is unable to estimate the parameter even for bright sources. Same is the situation with the *infall* parameter for the *gal_ I* and the *gal_ B* models. For these two models, there is also a problem with the *sfrP2* parameter estimation for faint sources (G_{mag}=18.5). For such sources, there is also unsatisfactory performance for the *B_ sfrP3* parameter estimation. Each case of unacceptable performance is marked by an asterisk on Tab.5.

The small number of irregular galaxies spectra (Tab.3) can not be the reason of an unsuccessful training, as for the *gal_ B* models this number is quite large. Possibly, either the spectral changes caused by these parameters are really undetectable because of the instrumental S/N, the TGE and the redshift influence, or the parameter values range and sampling shall be optimized.

Conclusion

The SVM-based algorithm implemented in UGCv10 is promising to provide a very good performance for unresolved galaxies spectra classification with the future Gaia mission. The tests based on simulated spectra library show also that UGC can estimate with good accuracy the galactic extinction towards the observed sources as well as their redshift. Acceptable regression performance is shown for most of the galaxy model parameters for brighter sources, whereas there is a problem with some of the parameters, mostly in the fainter range of magnitudes.

Further investigation of the specific parameters influence on the simulated spectra is necessary to optimize the spectral library and to improve the SVM learning. Applying nonlinear pre-processing of particular parameters seems promising to help towards better performance. Development of similar algorithm, but based on Artificial Neural Networks (ANN) is worth to be considered. Unlikely the SVM, an ANN-model can be trained to estimate few parameters at once.

Acknowledgments

Simulated data provided by the Simulation Unit (CU2) of the Gaia Data Processing Analysis Consortium (DPAC) have been used to complete this work. The simulations have been done in the supercomputer MareNostrum at Barcelona Supercomputing Center - Centro Nacional de Supercomputacion (The Spanish National Supercomputing Center). They are gratefully acknowledged for this contribution. We thank Antonella Vallenari and Rosanna Sordo, members of the CU8 of DPAC, for their work in organization and preparation the data for simulation.

The South-Eastern Europe Grid infrastructure, part of the EGI[5], has been used to complete the training of the large number of SVM-models.

References

Bellas-Velidis I., Kontizas M., Livanou E., Tsalmantza P., 2010, Unresolved Galaxy Classifier for ESA's Gaia Mission, Proceedings of the 9th International Conference of the Hellenic Astronomical Society, held 20-24 September 2009 in Athens, Astronomical Society of the Pacific Conference Series, 424, 256-260

Cardelli J.A., Clayton G. C., Mathis J.S., 1989, Astrophys.J. 345, 245-256

Chang, C-C., & Lin, C-J., 2007, LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/ cjlin/libsvm

Cortes C., Vapnik V., 1995, Support-Vector Networks, in Machine Learning 20, 273-297

DPAC Proposal for the Gaia Data Processing, Data Processing & Analysis Consortium, 2007, Editors: Mignard F., Drimmel R., GAIA-CD-SP-DPAC-FM-030-2 (restricted access)

Drucker H., Burges C.J.C. , Kaufman L., Smola A., Vapnik V., 1997, Support Vector Regression Machines, Advances in Neural Information Processing Systems 9, NIPS 1996, MIT Press, 155-161

Fioc M., & Rocca-Volmerange B., 1997, Astron.Astrophys. 326, 950-962

Isasi Y., Borrachero R., Martinez O., Sartoretti P., Luri X., Babusiaux C., Zaldua I., 2010, GOG User Guide, GAIA-C2-UG-UB-YI-003-7 (restricted access)

Karampelas A., Kontizas M., Rocca-Volmerange B., Bellas-Velidis I., Kontizas E., Livanou E., Tsalmantza P., Dapergolas A., 2011, VizieR On-line Data Catalog: J/A+A/538/A38

O'Mullane W., Hoar J., Levoir T., De Angeli F., Nguyen A-T., Lammers U., Sadowski G., Sidiqqui H., 2011, Software Engineering Guidelines for DPAC, GAIA-C1-UG-ESAC-WOM-011-6 (restricted access)

Smith K., Bailer-Jones C.A.L., Tsalmantza P., 2011, Discrete Source Classifier performance and status report, GAIA-C8-TN-MPIA-KS-019-01 (restricted access)

Tsalmantza P., Kontizas M., Bailer-Jones C. A. L., Rocca-Volmerange B., Korakitis R., Kontizas E., Livanou E., Dapergolas A., Bellas-Velidis I., Vallenari A., Fioc M., 2007, Astron.Astrophys. 470, 761-770

Tsalmantza P., Kontizas M., Rocca-Volmerange B., Bailer-Jones C. A. L., Kontizas E., Bellas-Velidis I., Livanou E., Korakitis R., Dapergolas A., Vallenari A., Fioc M., 2009, Astron.Astrophys. 504, 1071-1084

[5] http://www.egi.eu