ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Machine Learning and Data Analysis in Astroinformatics

Michael Biehl[1], Kerstin Bunte[1], Guiseppe Longo[2], and Peter Tino[3]

1- University of Groningen - Intelligent Systems Group
P.O. Box 407, 9700 AK Groningen - The Netherlands

2- Universitá degli Studi Federico II - Dipartimento di Fisica "E. Pancini"
via Cintia 6, I-80135 Napoli, Italia

3- University of Birmingham - School of Computer Science
Birmingham B15 2TT, UK

Abstract. Astroinformatics is a new discipline at the cross-road of astronomy, advanced statistics and computer science. With next generation sky surveys, space missions and modern instrumentation astronomy will enter the Petascale regime raising the demand for advanced computer science techniques with hard- and software solutions for data management, analysis, efficient automation and knowledge discovery. This tutorial reviews important developments in astroinformatics over the past years and discusses some relevant research questions and concrete problems. The contribution ends with a short review of the special session papers in these proceedings, as well as perspectives and challenges for the near future.

1 Introduction

The ever-growing amount of data which becomes available in many domains clearly requires the development of efficient methods for data mining and analysis. These challenges concern a variety of areas including fundamental scientific research, business and societal issues. Astronomy continues to be at the forefront of this development with archives in the Multi-Tera and even Petabyte domain and, thanks to the Virtual Observatory Infrastructure (VO)[1], data discovery and access in astronomy have never been easier [2, 3, 4]. However, while the established data centers can at least in principle expand and scale up to the next generation of sky surveys, space missions, etc. there is still the lack of a powerful arsenal of tools capable to effectively extract knowledge from these new data sets and data streams.

Modern observational techniques provide in fact enormous amounts of complex, heterogeneous data, which cannot be processed with traditional methods and require the introduction of novel techniques, largely based on machine learning at almost every step of the process: data acquisition and storage with automatic assessment of data quality, data fusion in order to exploit the legacy value of pre-existing data, data analysis and visualization and, finally data interpretation. These needs led to the birth of Astroinformatics [5, 6, 7, 8]: a new discipline at the cross-road of computer science, astronomy and advanced statistics.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Data, no matter how large and complex they are, are just incidental to the real task of scientists, *knowledge discovery*. Since large and complex data tend to preclude direct human examination, an automation of these processes is needed, requiring use of Machine Learning (ML) techniques. In spite of the fact that astronomical applications of ML are relatively recent and restricted to a small but rapidly growing number of problems, Astroinformatics still provides superb examples for challenging tasks like big data mining, image processing, filtering for streams of data, outlier and novelty detection, classification and clustering, statistical inference, modelling and simulation, to name only a few. Astronomical problems can therefore serve as an excellent workshop in which to put further advanced methods of machine learning and data science in general. The idea behind the discipline of Astroinformatics is therefore to provide an informal, open environment for the exchange of ideas, software, etc., and to act as a connecting cultural tissue between researchers active in the scientific exploitation of astronomical big data.

In the broader scientific context, Astroinformatics has a significant *and ever growing* impact in a variety of areas ranging from hard- and software solutions for large scale data management and data analysis, efficient automation of otherwise tedious and costly tasks, as well as knowledge discovery.

2 Important developments

To quote just a few recent examples of problems where the use of ML tools on massive astronomical data sets is proving crucial, we can mention: automated classification of unresolved vs resolved sources detected in sky surveys [9, 10] (e.g. stars vs galaxies, or globular clusters vs stars[11]); classification or selection of objects of a given type in some parameter space (e.g. normal vs active galaxies or quasars; morphological classification of galaxies, lensed galaxies or lensed arcs); estimates of photometric redshifts [12, 13, 14] for large samples of galaxies; identification of star forming regions and classification of young stellar objects, etc.

The rapidly developing field of time-domain astronomy [15, 16] poses however some new challenges. A new generation of synoptic sky surveys produces data streams that correspond to the traditional, one-pass sky surveys many times repeatedly. In addition to the dramatic increase of data rates and the resulting data volumes, there is a need to identify, characterize, classify, and prioritize for the follow-up observations any transient events or highly variable sources that are found in the survey data streams. Since many such events are relatively short in duration (think for instance to the multi messenger approach to the detection of gravitational waves), this analysis must be performed as close to the real time as possible. This entails challenges that are not present in the traditional automated classification approaches, which are usually done in some feature vector space, with an abundance of self-contained data derived from homogeneous measurements.

In contrast, measurements generated in the synoptic sky surveys are gener-

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

ally sparse and heterogeneous: there are only a few initial measurements, their types differ from case to case, and the values have differing variances; the contextual information is often essential, and yet difficult to incorporate; many sources of noise, instrumental glitches, etc., can mimic as transient events; as new data arrive, the classification must be iterated dynamically. We also require a high completeness (capture all interesting events) and a low contamination (minimize the number of false alarms). Finally; as in virtually all application areas, the *blind* use of complex algorithms in a *black-box* fashion should be avoided **wherever possible**. The use of interpretable models and systems is instrumental for both the integration of expert knowledge and the acquisition of novel insights.

3 Relevant research questions and concrete problems

This Special Session brings together researchers who develop, investigate or apply methods of machine learning and data analysis in the context of astronomical data. Obviously, the contributions can only address a small subset of the many practical challenges and research topics which are relevant in this context:

- Efficient handling and analysis of truly Big Data
- Data mining and knowledge discovery in astronomical data
- Processing and analysis of astronomical image data
- Filtering techniques for large volume streams of data
- Outlier and novelty detection in observational data
- Classification and Clustering of celestial objects
- Development of transparent, interpretable models and algorithms
- Simulation of astrophysical models and related inference problems
- Analysis of heterogeneous data sets obtained from various sources or technical platforms
- Methods of *transfer learning* and *learning from privileged information*
- Development of *human (expert) in the loop* approaches
- Systematic incorporation of prior information and domain knowledge in machine learning

The organizers hope to increase awareness of the timely and highly relevant challenges that Astroinformatics provides. They should attract considerable attention among the participants of ESANN 2018 in particular and the machine learning and data analysis community in general.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

4 Contributions to the ESANN 2018 special session on Astroinformatics: Machine Learning and Data Mining in Astronomy

The five accepted contributions to this special session present a nice cross section through the relevant research questions discussed in the previous section:

In the contribution of Chen et *al.*, the problem of unsupervised anomaly detection in star light curves is addressed [17]. The latter represent the time dependent brightness of celestial objects. The authors apply Hierarchical Gaussian Processes in order to establish a system for anomaly detection. It is shown to outperform established baseline methods in several example data sets.

Huijse et *al.* explore data from the Chilean Automatic Supernovae Search (CHASE) [18]. Deep unsupervised autoencoders are employed to obtain compressed latent space representations for a large candidate database. The authors argue that their method preserves more information and provides better classification of stellar transients than classical methods.

The contribution of Veneri et *al.* discusses an application of supervised machine learning and feature selection methods [19]. The aim is to estimate global Stellar Formation Rates, which play a crucial role in the theory of galaxy formation and evolution. The authors investigate and compare the performance of Random Forest models and Multi Layer Perceptrons in combination with two feature selection strategies, predicting average total star formation rates based on data from the Sloan Digital Sky Survey.

The detection of globular clusters is addressed by Mohammadi et *al.* [20]. The authors consider two different basic strategies for this task which are based on nearest neighbor retrieval and anomaly detection, respectively. Both techniques are illustrated and tested in terms of GAIA satellite survey data, where they facilitate the identification of previously known clusters and suggest novel candidate structures.

Nolte et *al.* analyse catalogue data from the *Galaxy and Mass Assembly* (GAMA) survey by use of prototype-based machine learning [21]. Unsupervised learning by means of Self-Organizing Maps and supervised Generalized Matrix Relevance Learning Vector Quantization are used to illustrate and demonstrate the conceptual difficulties of the popular visual-inspection based galaxy classification scheme.

5 Perspectives and future work

The coming years will see astronomy entering the Petascale regime in terms of data volumes, with an ever increasing growth in data size and data complexity pushing at least some parts of the astronomical research beyond the Exascale threshold. In the past the astronomical community has responded well and in a timely manner to the new challenges embracing Internet-accessible archives, databases, interoperability, standard formats and protocols, and realising a virtual scientific organization, the so called Virtual Observatory, that is now effec-

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

tively a global data grid of astronomy. It is not clear whether this virtuous trend will continue, unless the needed amount of economic and human resources will be invested.

Many good statistical and data mining tools and methods are in place and are gradually permeating the astronomical community, although their uptake has been slower than what could be hoped for. However, as it always happen, entering a new field, opens new problems and poses new challenges. One tangible technical problem which needs to be solved asap is the scalability of DM tools: most of the readily available ones do not scale well to the massive data sets foreseeable for the near future. The key problem is not so much in the data volume (expressible, e.g., as a number of feature vectors in some data set), but rather in their dimensionality: most data mining algorithms in fact scale very badly with increasing number of dimensions and become unusable when the intrinsic dimensionality of the data sets is measured in tens, hundred,or thousands. Effective, scalable software and a methodology needed for knowledge discovery in modern, large and complex data sets typically do not exist yet, at least in the public Domain, and will pose serious challenges to the astronomical as well to the scientific community at large. An even more problematic challenge is the fact that the most exciting problems in modern astronomy and cosmology require a panchromatic, multi-messenger approach. Large and distributed data sets (data usually stay close to the production centers) will need to be merged and mined without being transferred over the network. The only possible way seems to modify and optimize the codes in order to run them where the data are.

Another important issue which needs to be addressed, is that of reproducibility. Already with the existing data sets and even more with the coming ones, data mining experiments have become extremely demanding in terms of computing time and it is of paramount relevance to be able to define proper standards capable to both preserve the results of old experiments and to make them available to the community for future re-analysis. Furthermore, we believe that given the plethora of ML methods -both existing and to be invented in the near future- it will become more and more difficult to compare the results obtained with different methods and or by different groups in an homogeneous and unbiased way. This problem, while well discussed and partly solved in the computer science community, is still in its infancy for what astronomical applications are concerned. A possible solution could be the implementation of a library of template data sets of suitable size and complexity to be used as benchmarks for specific applications.

Acknowledgement
The organization of this special session was inspired and supported by the European H2020 MSCA Innovative Training Network *Survey Network for Deep Imaging Analysis and Learning* (SUNDIAL), visit
`http://www.astro.rug.nl/sundial/` for further information.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

References

[1] S.G. Djorgovski and the NVO Science Definition Team. Towards the national virtual observatory. In R. Brunner, S.G. Djorgovski, and A. Szalay, editors, *Virtual Observatories of the Future*, volume 225 of *Astronomical Society of the Pacific Conference Series*, 2001.

[2] Brunner R.J., Djorgovski S.G., Prince T.A., and Szalay A.S. Massive datasets in astronomy. In Abello J., Pardalos P.M., and Resende M.G.C., editors, *Handbook of Massive Data Sets*, volume 4, pages 931–979. Springer, Boston, MA, 2002.

[3] J. Bloom and J. Richards. Data mining and machine-learning in time-domain discovery & classification. In *Advances in Machine Learning and Data Mining for Astronomy*. 2011.

[4] C. Donalek, M. Graham, A. Mahabal, S.G. Djorgovski, and R. Plante. Tools for data to knowledge. Technical report, VAO, 2011.

[5] N. M. Ball and R. J. Brunner. Data mining and machine learning in astronomy. *International Journal of Modern Physics D*, 19(07):1049–1106, 2010.

[6] J. Kremer, K. Stensbo-Smidt, F. Gieseke, K. Steenstrup Pedersen, and C. Igel. Big universe, big data: Machine learning and image analysis for astronomy. *ArXiv e-prints*, April 2017.

[7] M. Hobson, P. Graff, F. Feroz, and A. Lasenby. Machine-learning in astronomy. In A. Heavens, J.-L. Starck, and A. Krone-Martins, editors, *Statistical Challenges in 21st Century Cosmology*, volume 306 of *IAU Symposium*, pages 279–287, May 2014.

[8] G. Longo, M. Brescia, S. G. Djorgovski, S. Cavuoti, and C. Donalek. Data driven discovery in astrophysics. *ArXiv e-prints*, October 2014.

[9] C. Donalek et al. New approaches to object classification in synoptic sky surveys, 2008. AIP Conf. Ser., 1082, 252.

[10] R. D'Abrusco, G. Longo, and N.A. Walton. Quasar candidate selection in the virtual observatory era, 2009. Mon. Not. Royal Astron. Soc., 396, 223.

[11] M. Brescia, S. Cavuoti, M. Paolillo, G. Longo, and T. Puzia. The detection of globular clusters in galaxies as a data mining problem, 2012. Monthly Notices Royal Astron. Soc., 421, 1155.

[12] R. Tagliaferri, G. Longo, S. Andreon, S. Capozziello, C. Donalek, and G. Giordano. Neural networks and photometric redshifts, 2002. astro-ph 0203445.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

[13] A.E. Firth, O. Lahav, and R.S. Somerville. Estimating photometric redshifts with artificial neural networks, 2003. Mon. Not. Royal Astron. Soc. 339, 1195.

[14] M. Brescia, S. Cavuoti, R. D'Abrusco, G. Longo, and A. Mercurio. Photometric redshifts for quasars in multi band surveys, 2013. Astroph. J, 772, 2, 140.

[15] S.G. Djorgovski, A.A. Mahabal, A.J. Drake, M.J. Graham, C. Donalek, and C. R. Williams, 2002. in Proc. IAU Symp. 285, New Horizons in Time Domain Astronomy, eds. E. Griffin et al., p. 141. Cambridge: Cambridge Univ. Press.

[16] M.J. Graham et al. Data challenges of time domain astronomy, 2012. in Distrib. Parallel Databases, eds. Qiu,X., Gannon, D.,30 (5-6), 371.

[17] H. Chen, T. Diethe, N. Twomey, and P. Flach. Anomaly detection in star light curves using hierarchical gaussian processes. This volume.

[18] P. Huisje, N. Astorga, P. Estévez, and G. Pignata. Latent representations of transient candidates from an astronomical image difference pipeline using variational autoencoders. This volume.

[19] M. delli Veneri, S. Cavuoti, M. Brescia, G. Riccio, and G. Longo. Stellar formation rates in galaxies using machine learning models. This volume.

[20] M. Mohammadi, R.F. Peletier, F.M. Schleif, N. Petkov, and K. Bunte. Globular cluster detection in the gaia survey. This volume.

[21] A. Nolte, L. Wang, and M. Biehl. Prototype-based analysis of gama galaxy catalogue data. This volume.