3rd International Workshop on Spectral Stellar Libraries
ASI Conference Series, 2017, Vol. 14, pp 73–82
Editors: P. Coelho, L. Martins & E. Griffin

Massively Parallel Machine Learning in the Virtual Observatory as a Key Technology in the Era of Multi-Million Spectral Surveys

P. Škoda*
Astronomical Institute of the Czech Academy of Sciences, Fričova 298, Ondřejov

Received July 1st, 2017 ; accepted Oct 15th, 2017

Abstract. The archives of multi-object spectral surveys such as SDSS or LAMOST currently contain millions of pipeline-reduced spectra of celestial objects. Most ca be identified as stars of recognised spectral types, according to quick comparisons with extensive lists of template spectra. To date, the dominant application of spectral libraries is for statistic estimates of similarity, measured in a sequential or simply parallel manner, by comparing all the survey spectra and their PCA components with a grid of templates.

In this paper we propose a new approach that uses modern machine-learning techniques as semi-supervised training, deep learning, or outlier detecting that helps to identify specific rare cases of unusual objects like stars with strong emission lines or P-Cyg profiles, or blazars, as well as to eliminate the instrumental and processing artefacts which cannot be handled correctly by a normal streaming pipeline. The amount of data and time-absorbing algorithms require a 'Big Data' approach, using massively parallel processing in the cloud by applying modern technologies such as GPGPUs, Hadoop and Spark.

An important stage towards verifying the results is an interactive visualisation and cross-matching with other data such as photometric surveys, spectra acquired by other surveys, space missions and multi-wavelength data of similar coverage, as well as comparisons with alternative models. All this can be easily achieved through correct exploitation of Virtual Observatory standards.

Keywords : stars: emission-line, Be; methods: data analysis; techniques: spectroscopic; virtual observatory; machine learning

*email: skoda@sunstel.asu.cas.cz

1. Introduction

Classifying the spectrum of a star is a basic procedure that assigns to it some likely value of its physical parameters such as colour temperature, gravity (which represents its size and density) and sometimes its metallicity (chemical composition). The usual way to classify a spectrum is to compare it with a grid of template spectra, which may be either synthetic or a library of carefully selected stellar spectra of known types. Problems arise in interesting physical cases if the spectrum in question is not known and the library is not rich enough to include the corresponding unusual type, for instance a cataclysmic variable or a Be, B[e], symbiotic or T Tau early-type star. To classify exotic types like those, machine learning to classify line profiles may be a reasonable alternative.

2. 'Big data' in astronomy

Astronomy, as with many other scientific disciplines like biology, genetics or climatology, is currently facing an avalanche of data that are too voluminous to be processed and exploited in full. For example, the future sky survey LSST (Jurić et al. 2015) will yield 15TB of raw data every night requiring the processing power of about 1.6 GFLOPS for the reduction of its data. The expected size of processed data after ten years of observation is about 500 PB, including 50 PB database tables. The final catalogue alone will be 15 PB.

Even bigger data volumes are produced by radio arrays. Currently the LOFAR long term archive infrastructure hosts 27 PB of data (Valentijn et al. 2016). The biggest world astronomical archive already in preparation will be created by the Square Kilometre Array (SKA). The archive is expected to grow by 3 PB per day, and its total size will increase by 1.1 exabyte per year (Barbosa et al. 2016).

In summary, the current growth of archived astronomical data has been rising exponentially with the doubling constant less than 6–9 months; that is much steeper than the famous Moore's law of technology advances, which predicts the doubling of computer resources every 18 month (Quinn, P., Lawrence, A., Hanisch, R. 2004). One promising solution of handling this data deluge is the implementation of service oriented architecture moving the burden of data processing, pre-analysis and searching towards the high-performance, well-equipped data centres. That solution has already identified by the astronomical community in 2002, and was a main driver for setting up astronomy's Virtual Observatory (VO), which has been under development of the International Virtual Observatory Alliance (IVOA).

3. Virtual Observatory

The important technology able to handle the astronomical 'Big Data' deluge is the concept of Virtual Observatory. Its goal is to provide global standards describing all astronomical resources worldwide and to enable the standardised discovery and access to these collections as well as powerful tools for scientific analysis and visualisation (Hanisch & De Young 2007).

As the VO mostly provides access to science-ready data, the data provider needs to make the calibrated data VO-compatible. That requires the creation of a set of metadata (for curation, provenance and characterisation), the conversion of data into a VOTable format (Ochsenbein et al. 2013), and the preparation of an access interface in accordance with appropriate VO standard protocols (Arviset, Gaudet & IVOA Technical Coordination Group 2012). In the case of spectra, the most important protocols are SSAP (Tody et al. 2012) and Spectral Data Model (McDowell et al. 2011).

Most of the highly acknowledged astronomical services like Vizier[1] and Simbad[2], or tools like Aladin[3], are practical examples of VO technology in everyday use. All the complex infrastructure of big servers, database engines and various data retrieval protocols is hidden under the hood of a simple web-based form and user-friendly graphical interfaces delivering complex tables, images and graphs at the click of a button. The VO also allows simple collaboration between multiple applications, thanks to the Simple Application Message Protocol — SAMP (Taylor, Boch & Taylor 2015).

For spectroscopic studies, it is important the collaboration among spectra analysis and visualisation tools such as SPLAT-VO (Škoda et al. 2014), TOPCAT[4] and Aladin, allowing e.g the investigation of spectra of multi-object spectrographs like LAMOST facilitated by the workflow consisting of (i) processing the metadata tables in TOPCAT, (ii) highlighting the individual object in Aladin sky image and (iii) sending the link to its spectrum to SPLAT-VO. The VO has already proved its power in solving problems too tedious for manual work, and in producing results that can hardly be achieved by classical methods (Chilingarian et al. 2009).

However, it is outside the scope of the VO to generate new knowledge, new models, and new scientific understanding from data. The effective retrieval of new and useful knowledge from these massive distributed databases requires automated and increasingly more effective approaches. Those are being addressed by progressive fields of knowledge discovery in databases (KDD) and by data mining (DM), based on machine-learning methods. Astronomy and astrophysics are entering the data-

[1] `http://vizier.u-strasbg.fr/viz-bin/VizieR`
[2] `http://simbad.u-strasbg.fr/simbad/`
[3] `http://aladin.u-strasbg.fr/aladin.gml`
[4] `http://www.star.bris.ac.uk/~mbt/topcat/`

intensive research paradigm represented by the newly emerging scientific discipline of Astroinformatics.

4. Astroinformatics

As it was mentioned earlier, the current science is commonly understood to be data-intensive or data-driven. Research in almost all the natural sciences is facing the 'data avalanche' represented by exponential growth of information. The effective retrieval of a scientific knowledge from petabyte-sized astronomical databases requires the qualitatively new kind of scientific discipline – Astroinformatics.

Astroinformatics is based on the systematic application of modern informatics and advanced statistics to huge astronomical data sets. Approaches involving techniques such as machine learning, classification, clustering and data mining yield new discoveries and better understanding of the nature of astronomical objects. Astroinformatics is an example of a scientific methodology in which new discoveries often result from searching for outliers in normal statistical patterns. It is sometimes presented as a new way of doing astronomy (Borne 2009; Ball & Brunner 2010).

Accomplishing an analysis in the VO infrastructure may benefit from the automatic aggregation of distributed archive resources (e.g. multi-spectral research), seamless on-the-fly data conversions, common interoperability of all tools, and powerful graphical visualisation of measured and derived quantities. Combining the VO's infrastructural power with easy and transparent high-performance computing will enable the use of advanced analysis of large spectral surveys to become feasible in a reasonable time. The crucial role in understanding the results of such an analysis plays the Astroinformatics as a methodology allowing the extraction of new physical knowledge from astronomical observations.

5. Mega-spectra surveys

The largest current surveys, containing millions of spectra (called 'mega-spectral surveys') have resulted from two long-term projects that use multi-object fibre spectrographs:

The Sloan Digital Sky Survey (SDSS): In its 12^{th} data release (DR12 Alam et al. 2015) there are **4.3 million** spectra. Two spectrographs have been fed by 640 fibres placed in pre-drilled holes in a focal plate, but recently a new spectrograph (BOSS), having 1000 fibres, has been installed. The spectra span the range 3800–9200Å (SDSS spectrograph) and 3650–10400Å (BOSS) with a spectral resolving power of about $\sim$ 1800. In addition, there are also 0.6 million of H-band infrared spectra from the APOGEE spectrograph

LAMOST Spectral Survey: The LAMOST telescope (Cui et. al. 2012) has been delivering one of the largest mega-collections of spectra to date. The sixteen LAMOST spectrographs are fed by 4000 fibres positioned by micro-motors. Its publicly accessible Data Release 1 (DR1), (Luo et al. 2015) contains altogether **2.2 million** spectra, while the DR3 (He et al. 2016) provides already **5.7 million** spectra.

Processing the surveys data is carried out by automatic pipelines which classify individual objects using a set of templates by best matching the global shape of spectra. The local features (e.g. line profiles) are ignored. Strong narrow emissions may be even rejected by pipeline as a possibly spoiled pixels.

Massive amounts of spectra (90,000 channels per exposure in one arcmin FOV) are also produced by the integral field unit (IFU) spectrograph MUSE (Kelz et al. 2016). The planned HETDEX survey with the HET telescope is going to provide 33,600 individual spectra; it will have about 22 arcmin per exposure and will use an IFU (called VIRUS) that has special fibre attached (Adams et al. 2011).

6. Machine Learning

Machine learning is the field of informatics, closely related to the advanced statistical inference, which tries to (1) build models of data by learning from sample inputs, and (2) make predictions based on such learned models. It is divided mainly into supervised and unsupervised methods.

Supervised learning requires, in principle, manually assigned labels attached to the training sample of data. The method then will try to identify the same labels (same classes) in another sample of the data whose labels are unknown.

Unsupervised learning tries to identify similar classes automatically (based on some similarity metrics) without human intervention 'Outliers' are entities which cannot be assigned to any particular cluster, so they represent single-member clusters.

A special case of semi-supervised learning (Chapelle et al. 2006) combines both approaches by using the labels on a few samples of data with the knowledge of the local topology of the data set in order to label all the data automatically.

The yet unknown rare objects with strange features hidden in the spectral archive, or even sources with yet undiscovered physical mechanism may be in principle found using this method. In any case, numerous random instrumental artefacts will be found as well, since every one is unique and thus very rare. The artefacts caused by systematic errors of the same nature, which repeats very often, may be collected by clustering as well.

As one of the important cases of interesting objects, which may be found by Machine Learning, may be considered Be stars (Porter & Rivinius 2003; Rivinius, Carciofi & Martayan 2013) showing in Balmer lines single or double peaked line profiles or even more complicated profiles with emission components (Zickgraf 2003; Silaj et al. 2010).

7. Use case example: finding emission-line objects with unsupervised and semi-supervised machine learning

Using the machine learning technology described above, we tried to find objects looking as Be stars in a sample of LAMOST DR1 spectra. We used semi-supervised methods called Label Propagation and Label Spreading, which were trained on spectra of Be stars that had been observed with the 2-m Perek Telescope of Ondřejov observatory. Details are given in Palička (2016) and Škoda et al. (2016a).

We also tried to apply an outlier searching method called Local Outlier Factor (LOF) to find the most peculiar spectra in LAMOST DR1 and also in the spectra from the 2-m telescope. Details may be found in Shakurova (2016) and Škoda et al. (2016b).

As the amount of spectra investigated was of the order of 50,000, we needed to use the massively parallel processing of the Spark engine in the Hadoop environment. Apache Spark[5] is a cluster computing technology that allows fast parallel computation on a number of computing nodes. We used the academic cluster MetaCentrum, which consists of 24 sixteen-core nodes (the number of nodes assigned by the system is actually unknown, as it depends on availability and the load of the cluster). The data were distributed across all nodes by the Hadoop Distributed File System HDFS[6]. The search was run on more than 50,000 spectra that were randomly selected from those labelled as 'star by the LAMOST DR1 pipeline.

Each trial returned the archive IDs of hundreds of candidates (for which a specified statistical measure was above the threshold), but not every candidate was a Be star. Flexible visualisation of each spectrum with VO technology played an important role in the final evaluation and identification of the nature of the candidates. It was particularly helpful to use a combination of cross-matching of the list of candidates with other catalogues and all-sky surveys in TOPCAT and Aladin, followed by displays of their spectra in SPLAT-VO (including the interactive zooming). All this was interactively orchestrated thanks to the SAMP interoperability protocol. We finally obtained a shortlist of very interesting-looking candidates for emission line stars that deserve further investigation.

[5] `http://spark.apache.org`
[6] `http://hadoop.apache.org`

Fig. 1 gives an example of a Be star that was found by semi-supervised training (the whole spectrum is shown in the upper panel, and in the panel below it is zoomed centred around H_α line). Fig. 2 shows another interesting object that has emission

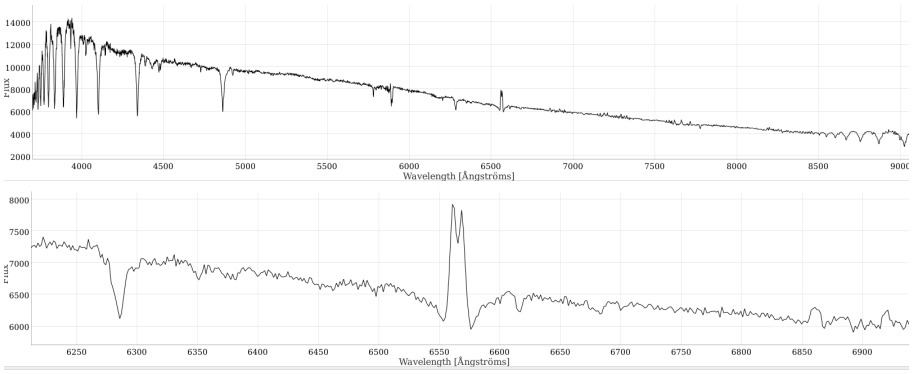

Figure 1. Example of a Be star identified in LAMOST DR1 by machine learning. It has strong double-peaked emission.

both in hydrogen and helium $\lambda 6678$Å lines. It was identified as the well-known cataclysmic variable star AM Her, so called Polar, with a strong magnetic field and X-ray activity (Terada et al. 2010). Fig. 3 gives an example of an outlier found by

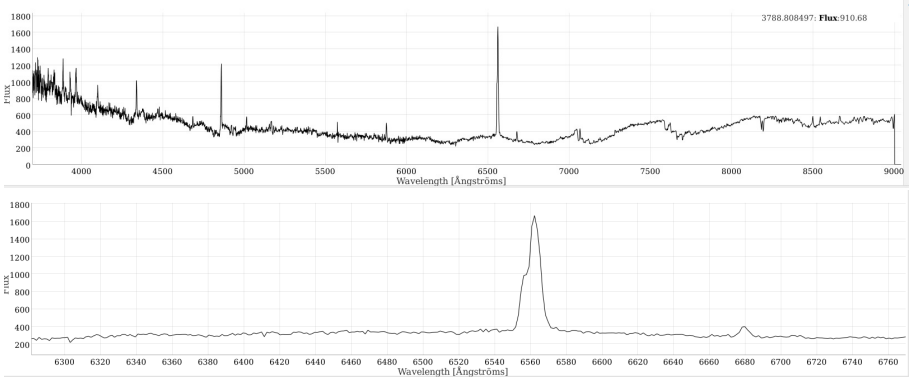

Figure 2. LAMOST spectrum of AM Her Polar

the LOF method. It was classified by the LAMOST pipeline as a late-type M7 star; it presents clearly a combination of absorption and emission in O I, which is seen in some Be stars. Even more interesting is the star found by the semi-supervised machine learning and shown in Fig. 4: it shows emission line profiles in several Balmer lines and even in other elements (probably nitrogen). The star could not be cross-matched

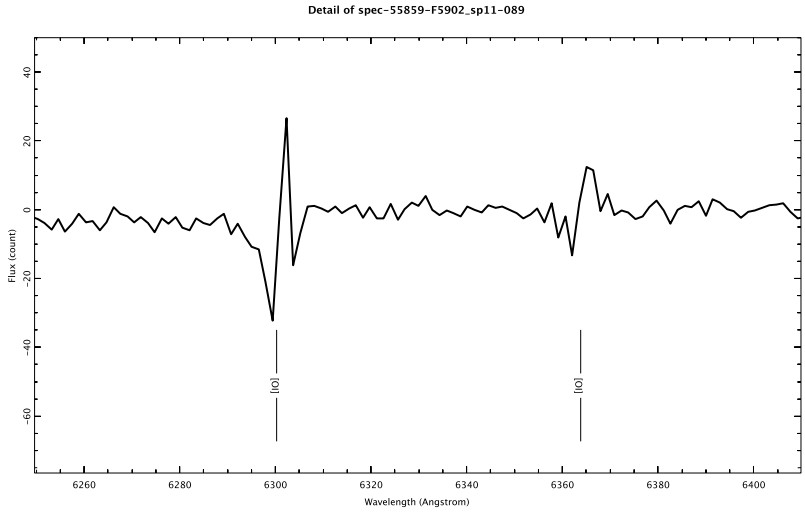

Figure 3. LAMOST outlier, classified as an M7 star

with any known stars in SIMBAD with sufficient accuracy, and so it deserves further investigation.

8. Conclusions

The massive amounts of spectra produced by current instrumentation may be analysed using the methodology of astroinformatics, namely the machine learning of specific spectral lines profiles, in order to identify candidates for interesting (e.g. emission line) objects. It may be a viable alternative to classical spectral classification – which uses the best fitting in a grid of stellar spectral libraries – focused on less common cases. The massive parallelisation of machine learning run in Spark environment can speed-up the task considerably. The environment of the Virtual Observatory and its specific technology may be very helpful thanks to its powerful analytic and visualisation capabilities.

Acknowledgements

This research was supported by the grant COST LD-15113 of the Ministry of Education Youth and Sports of the Czech Republic. Part of the work is based on spectra from Ondřejov 2m Perek telescope and public LAMOST DR1 survey. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is

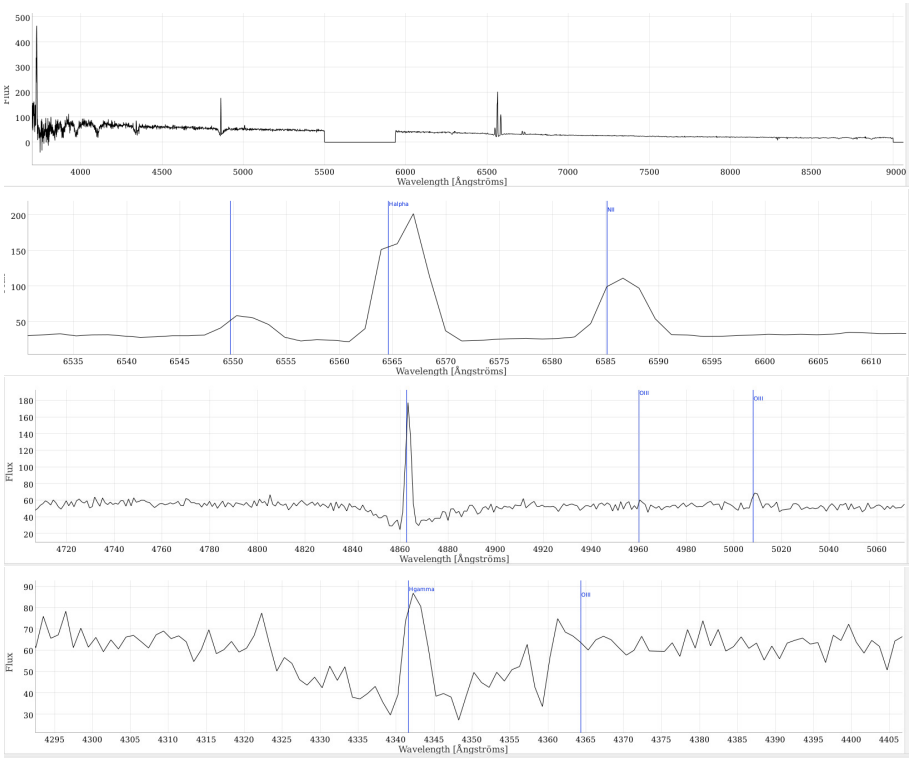

Figure 4. Interesting LAMOST star showing emission in several Balmer lines

greatly appreciated. The author's participation at the conference was made possible with FAPESP project 2016/13479-0.

References

Adams J. J., et al., 2011, ApJS, 192
Alam S., et al., 2015, ApJS, 219, 12
Arviset C., Gaudet S., IVOA Technical Coordination Group, 2012, The IVOA Architecture, EPSC2012-626, Arxiv:1106.0291
Ball N. M., Brunner R. J., 2010, IJMPD, 19, 1049
Barbosa D., Barraca J. P., Carvalho B., Maia D., Gupta Y., Natarajan S., Le Roux G., Swart P., 2016, SPIE, 9913, 99130K
Borne K., 2009, arXiv, arXiv:0911.0505
Chapelle, O., Schölkopf, B., & Zien, A. 2006, Semi-supervised learning, MIT press, Cambridge, Massachusetts
Chilingarian I., Cayatte V., Revaz Y., Dodonov S., Durand D., Durret F., Micol A., Slezak E., 2009, Sci, 326, 1379

Cui, X.Q. et al. 2012, Research in Astronomy and Astrophysics, 12, 1197
Hanisch R. J., De Young D., 2007, ASPC, 382, 1
He B., Fan D., Cui C., Li S., Li C., Mi L., 2016, arXiv, arXiv:1601.02334
Jurić M., et al., 2015, arXiv, arXiv:1512.07914
Kelz A., Kamann S., Urrutia T., Weilbacher P., Bacon R., 2016, ASPC, 507, 323
Luo, A.L. et al. 2015, Research in Astronomy and Astrophysics, 15, 1095
McDowell J., et al., 2011, IVOA Recommendation: IVOA Spectrum Data Model Version 1.1, arXiv:1204.3055
Ochsenbein F., et al., 2013, IVOA Recommendation: VOTable Format Definition Version 1.3, online: http://www.ivoa.net/documents/VOTable/
Quinn, P., Lawrence, A., Hanisch, R. 2004, IVOA Note: The Management, Storage and Utilization of Astronomical Data in the 21st Century, online: http://www.ivoa.net/about/OECD-QLH-Final.pdf
Palička, A. 2016, Master Thesis, Czech Technical University in Prague, Faculty of Information technology
Porter, J. M., & Rivinius, T. 2003, PASP, 115, 1153
Rivinius T., Carciofi A. C., Martayan C., 2013, A&ARv, 21, 69
Shakurova, K. 2016, Master Thesis, Czech Technical University in Prague, Faculty of Information technology
Silaj, J., Jones, C. E., Tycner, C., Sigut, T. A. A., & Smith, A. D. 2010, ApJS, 187, 228
Škoda P., Draper P. W., Neves M. C., Andrešič D., Jenness T., 2014, A&C, 7, 108
Škoda P., Palička A., Koza J., Shakurova K., 2016, arXiv, arXiv:1612.07536
Škoda P., Shakurova K., Koza J., Palička A., 2016, arXiv, arXiv:1612.07549
Taylor M. B., Boch T., Taylor J., 2015, A&C, 11, 81
Terada Y., Ishida M., Bamba A., Mukai K., Hayashi T., Harayama A., 2010, ApJ, 721, 1908
Tody, D. et al. 2012, IVOA Recommendation: Simple Spectral Access Protocol Version 1.1, ArXiv:1203.5725
Valentijn E. A., et al., 2016, arXiv, arXiv:1612.05996
Zickgraf, F.-J. 2003, A&A, 408, 257