Knowledge discovery in big astronomical spectra archives

Petr Škoda[1], Pavla Bromová[2], Lukaš Lopatovský[3], Andrej Palička[3], Jaroslav Vážný[4]

[1]*Astronomical Institute, Academy of Sciences, Ondřejov, Czech*
[2]*Faculty of Information Technology, Brno University of Technology, Brno*
[3]*Faculty of Informatics, Czech Technical University, Prague*
[4]*Masaryk University, Faculty of Science, Brno*

Abstract

The recent progress of astronomical instrumentation resulted in the construction of multi-object spectrographs with hundreds to thousands of micro-slits or optical fibers allowing the acquisition of tens of thousands of spectra of celestial objects per observing night. Currently there are several spectroscopic surveys containing millions of spectra and much larger are in preparation. Most of the large-scale surveys are processed spectrum by spectrum in order to estimate physical parameters of individual objects. The parameters obtained are then used to construct the better models of space-kinematic structure and evolution of the Universe or its subsystems. Such surveys are, however, very good source of homogenized, preprocessed data for application of machine learning techniques and advanced statistical processing common in Astroinformatics. We present challenges of knowledge discovery process applied to large spectroscopic surveys as well as memory space and processing speed demands of current machine learning methods, requiring Big Data techniques.

Introduction

Although the current spectroscopic surveys have been producing much smaller object catalogues than photometric all-sky surveys, the amount of information collected is already escaping the capabilities of human preview and analysis so far common in classical spectroscopy. Millions of spectra (further called mega-surveys) are usually processed spectrum by spectrum by complex pipelines to yield redshift, spectral type or similar physical parameters. The more detailed knowledge hidden in this "Big Data" must, however, be discovered by methods of Artificial Intelligence and Machine Learning.

Spectral mega-surveys

The currently largest mega-surveys result from two long-term projects using multi-object fibre spectrographs:

- **Sloan Digital Sky Survey** (SDSS). In its DR10 [1] there are **3.3 million** spectra. Two spectrographs were so far fed by 640 fibres placed in predrilled holes of focal plate, recently a new spectrograph BOSS with 1000 fibres has been used. There are 1.8 millions identified as galaxies, 308 000 as quasars and more than 700 000 are stellar.
- **LAMOST survey**. Its DR1 [14] contains **2.2 million** spectra, The sixteen LAMOST spectrographs are fed by more than 4000 fibres positioned by micro-motors. In the survey there are more than 1 million of stars with estimated parameters.

The processing of both surveys is done 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.

Emission line objects

There is a lot of objects that may show some important spectral lines in emission. The physical parameters may differ considerable, however, there seems to be the common origin of their emission — the gaseous envelope in the shape of sphere or rotating disk. To this interesting group belong the Be and B[e] stars, but very similar shapes may be seen in some quasars and AGNs in general.

Be and B[e] stars

The classical Be stars [7] are non-supergiant B type stars whose spectra have or have had at some time, one or more emission lines in the Balmer series. In particular the H_α emission is the dominant feature in spectra of these objects. Characteristic for Be stars are the single or double-peak profiles and sometimes so called shell lines — deep absorptions in centre of the emission. They may be also variable on different time scales. The emission lines are commonly understood to originate in the flattened circumstellar disk, probably of decretion origin (i.e. created from material of central star), however the exact mechanism is still unsolved.

Similar strong emission features in H_α show the B[e] stars [15], however they present as well forbidden lines of low excitation elements (e.g. Iron, Carbon, Oxygen, Nitrogen) and infrared excess (pointing to the presence of dusty envelope). The B[e] stars are very rare, mostly unclassified.

Models of Be stars

The extensive effort of explaining the various shapes of Be stars as well as their variability led to sophisticated models of disks under different physical conditions.

The models [12] show the dependence of the shape of prominent emission lines on the spectral type of the star, the physical structure of the disk as the temperature and density distribution, and, namely, on the inclination angle to the observer's plane, which is clearly the main generator of changes from single peak to double peak and shell profiles. They succeeded also to find the good models fitting well the observed profiles of many well-known Be stars.

Quasars and AGNs

According to the commonly accepted unified model of AGNs [9] the shape of emission lines in quasars is also generated by different geometrical conditions and namely the inclination of disk to the observer's plane like for Be stars (but now the in-plane disk is thick and obscuring the underlying emission source)

As was shown by [10] and [11] the physics of central black hole in quasar may be estimated from the characteristic shape of emission line in so called Broad Line Regions and there seem to be two different populations of quasars with either Gaussian or Lorentzian components in complex profiles of prominent emission lines like H_β, Ly_α, CIV 1549Å or HeII 1640Å.

Identification of Be and B[e] stars in mega-surveys

As the most prominent emission of such stars is shown in the H_α line around the laboratory air wavelength 6562.8Å, the successful identification of a Be or B[e] star (or another low redshift emission line object) requires visualisation of the short wavelength interval (about 30–50Å) centered at this position. The spectral resolving power of about 2000 common to both surveys is satisfactory to distinguish the double peak profile, although more details (e.g. shell lines) are not resolved. The visual identification may be aided by over-plotting of large number of zoomed profiles on continuum normalized spectra with interactive point-and-click selection as it is realized in program SPLAT-VO [8], but still it presents and enormous amount of work. The more promising approach seems to be presented by application of machine learning methods.

Automatic classification by supervised learning

To find emission line objects in a big survey, the automatic procedure must be used based on principles of supervised machine learning. It is basically the pattern recognition problem. The shape of a line is described by several parameters (called feature vector). Than a sample of both positive and negative examples (assigning labels manually) is selected for training the machine learning classifier. The samples must be randomly mixed and the many-fold cross-validations are applied until the system correctly recognises maximum of positive samples in any mixture of input vectors. Resulting classifier is applied on unknown spectra.

After preliminary experiments with spectra of Be stars from Ondřejov 2m Perek telescope archive used for training simple classifier based on two parameters (height, width) of a Gaussian line fit [13], more advanced methods like Artificial Neuron Network, Support Vector Machines or Decision trees were tested as a kernel of the classifier, however the most promising are Random Decision Forests and Random Ferns [6]. Their advantage for application on big spectral archives is the possibility of their massive parallelisation, namely on GPUs.

Finding outliers with unsupervised learning

While the supervised training described above helps to classify the spectra archive and thus helps to find the objects of given class, that was already identified in a sample and labelled accordingly, the unsupervised learning tries to identify similar classes automatically without the human intervention. One of a very useful method is the Kohonen Self-Organising Map (SOM), which can even help to identify outliers, e.g. yet unknown or very rare objects with strange features hidden in the spectral archive. SOM is, in fact, a multi-dimensional topological map of artificial neurons projected in 2D space [4]. The measure of similarity is the distance between the neurons in such a space represented in a 2D by so Unified Distance Matrix (U-matrix). The outliers are situated in a places with most widely separated neurons (highest U-matrix values).

Experiments with almost 1700 spectra of Be, B[e] and ordinary stars from archive of 2m Ondřejov Perek Telescope, that were already visually classified into 4 classes with different shape of H_α line (pure absorption, single-peak emission, double peak emission and absorption combined with emission), have convincingly identified the B[e] stars as most exotic profiles in distant U-matrix clusters [5]. There are several GPU parallelizable implementations of SOMs with good scalability, however the big SOMs will hardly fit in memory of GPUs and so new specific algorithms for GPUs have to be created.

Dimensionality reduction

Even the quick massively parallel computer or GPU cluster will not be able to process millions of several thousand pixels long feature vector iteratively in a reasonable time. So the dimension of a feature vector must be reduced significantly, still conserving the most characteristic features of line shapes. The common method of dimensionality reduction, the Principal Component Analysis, unfortunately fails here as the bulk of data in every spectrum is similar, the difference is only the spectral line, which is localised in a small part of whole spectrum [3].

Thus another methods capable to emphasise the weight of the strictly localised features is needed. One of this is the Wavelet Transform. All spectra are converted in a vector of wavelet coefficients corresponding to different frequencies. Experiments show good performance of such procedure, which could still sep-

arate all classes of line shapes with high accuracy even if the 2000 pixels long vector was degraded into 10 numbers using Wavelet Power Spectrum [2].

COST Action BIG-SKY-EARTH

As was shown, the extraction of new information from the spectral mega-surveys requires a sophisticated Artificial Intelligence techniques, new highly scalable and massively parallelizable algorithms, namely for GPUs and handling of Big Data in a efficient manner (e.g. on-the-spot post-processing and distributed queries as in VO technology). The information discovery in a big databases is a subject of a new astronomical discipline, the Astroinformatics, emerging today.

Similar problems with Big Data have other natural sciences as well. The most similar to astronomical problems seem to be the Earth sciences like geophysics, remote sensing, oceanography etc. Therefore wide collaboration was set up in a framework of European COST Action TD1403 called BIG-SKY-EARTH. The main goals are:

- Optimisation of database tools in astro- and geophysics contexts
- Data mining and machine learning in petabyte era as frontiers in astronomy and Earth observations
- Education of new generation of experts in the knowledge extraction from massive datasets
- Visualisation of high dimensional database

This European networking action is planned for years 2015 to 2018.

Conclusions

The big spectral archives are good source of homogenised data suitable for data mining of interesting objects according to their characteristic spectral line shape. The standard methods of supervised learning can be used to find the objects of given class, e.g. emission stars, however the advanced unsupervised methods as SOM help to identify outliers. The long processing time of millions of spectra may become feasible with reduction of their dimensionality to several elements of input feature vector, or by massively parallel processing, including GPUs. This, however, requires a change in commonly used algorithms in order to develop new massively parallelizable ones.

Acknowledgements

This work was supported by grant GAČR 13-08195S of Czech Science Foundation, specific research grants FIT-S-14-2299, IT4Innovations Centre of Excellence

ED1.1.00/02.0070 and project RVO:67985815 as well as European COST action "Big Data Era in Sky-Earth Observations" in trans-domain research TD1403. The work is based on spectra from Ondřejov 2m Perek telescope, Sloan Digital Sky Survey and public LAMOST DR1 survey.

References

[1] Ahn & et al. 2014, ApJS, 211, 17

[2] Bromová, P., Bařina, D., Škoda, P., Vážný, J., & Zendulka, J. 2014a, in Astronomical Data Analysis Software and Systems XXIII, edited by N. Manset, & P. Forshay, vol. 485 of ASP Conference Series, 177

[3] Bromová, P., Škoda, P., & Vážný, J. 2014b, Int. J. Autom. Comput., 11, 265

[4] Kohonen, T. 1982, Biological Cybernetics, 43, 59

[5] Lopatovský, L. 2014, Bachelor's thesis, Czech Technical University in Prague, Faculty of Information Technology.
`https://dip.felk.cvut.cz/browse/pdfcache/lopatluk_2014bach.pdf`

[6] Palička, A. 2014, Bachelor's thesis, Czech Technical University in Prague, Faculty of Information Technology.
`https://dip.felk.cvut.cz/browse/pdfcache/palicand_2014bach.pdf`

[7] Porter, J. M., & Rivinius, T. 2003, PASP, 115, 1153

[8] Škoda, P., Draper, P. W., Castro Neves, M., Andrešič, D., & Jenness, T. 2014, Astronomy and Computing, Vol. 7–8, pp. 108–120 `arXiv 1407.1765`

[9] Urry, C. M., & Padovani, P. 1995, PASP, 107, 803

[10] Sulentic, J. W., Marziani, P., Zamanov, R., et al. 2002, ApJL, 566, L71

[11] Gaskell, C. M. 2009, New Astronomy Reviews, 53, 140

[12] Silaj, J., Jones, C. E., Tycner, C., Sigut, T. A. A., & Smith, A. D. 2010, ApJS, 187, 228

[13] Škoda, P., & Vážný, J. 2012, in Astronomical Data Analysis Software and Systems XXI, edited by P. Ballester, D. Egret, & N. P. F. Lorente, vol. 461 of Astronomical Society of the Pacific Conference Series, 573

[14] Zhao, G., Zhao, Y.-H., Chu, Y.-Q., Jing, Y.-P., & Deng, L.-C. 2012, Research in Astronomy and Astrophysics, 12, 723

[15] Zickgraf, F.-J. 2003, A&A, 408, 257