Knowledge Discovery in Mega-Spectra Archives

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

[1]Astronomical Institute, Academy of Sciences, Ondřejov, Czech Republic
skoda@sunstel.asu.cas.cz

[2]Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic

[3]Faculty of Informatics, Czech Technical University, Prague, Czech Republic

[4]Masaryk University, Faculty of Science, Brno, Czech Republic

Abstract

The recent progress of astronomical instrumentation resulted in the construction of multi-object spectrographs with hundreds to thousands of micro-slits or optical fibres allowing the acquisition of tens of thousands of spectra of celestial objects per observing night. Currently there are several spectroscopic surveys (as SDSS or LAMOST) containing millions of spectra and much larger are in preparation.
These surveys are being processed by automatic pipelines, spectrum by spectrum, in order to estimate physical parameters of individual objects resulting in extensive catalogues, used typically 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 homogenised, pre-processed data for application of machine learning techniques common in Astroinformatics.
We present challenges of knowledge discovery in such surveys as well as practical examples of machine learning based on specific shapes of spectral features used in searching for new candidates of interesting astronomical objects, namely Be and B[e] stars and quasars.
Finally, the need for the new Big Data approach in such effort will be stressed, including the development of new massively parallel machine learning algorithms as well as better collaboration with non-astronomical communities sharing similar problems (e.g. Earth observation).

1 Spectral Mega-Surveys

The currently largest collections of millions of spectra ("mega-survey") comes from two projects:

- **Sloan Digital Sky Survey** (SDSS). In its DR10 there are **3.3 millions** of spectra. Two spectrographs were so far fed by 640 fibres placed in pre-drilled 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 contains **2.2 millions** spectra, The LAMOST 16 spectrographs are fed by 4000 fibres positioned by micro-motors. There are more than 1 million of stars with estimated parameters.

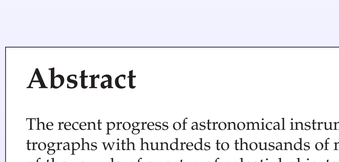

Fig. 1. The SDSS telescope and its focal plane with fibres in drilled holes

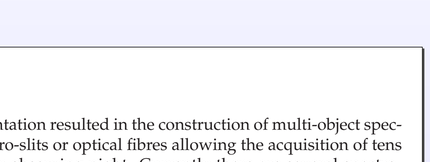

Fig. 2. LAMOST telescope and its focal plane with fibres moved by micro-motors

The processing of both surveys is done by several automatic pipelines which classify objects by best match templates and measure red shift. The result is a big catalogue with many parameters for every spectrum. **What objects are in the survey ?**
The stellar pipelines are mostly estimating the spectral type of a star by matching the global shape of spectra. The local features (e.g. line profiles) are ignored. Strong emissions are even rejected by pipeline as a possibly spoiled pixels.

2 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. One of this interesting group are the Be stars.

2.1 Be Stars

The classical Be stars 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. The shape of the emission line may be quite different and even variable on different time scales. Characteristic for Be stars are the double-peak profiles or even so called shell lines — deep absorptions in centre of the emission. 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.

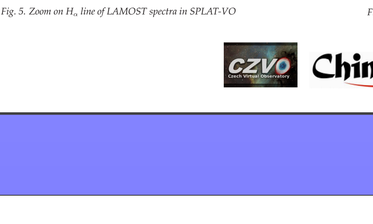

Fig. 3. Mechanism of creation of emission line in Be stars disk

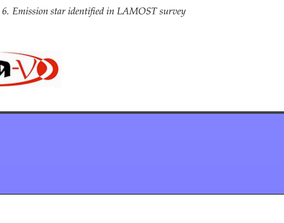

Fig. 4. Examples of H_α line profiles in Be stars

3 Identification of Be stars in Mega-surveys

The successful identification of a Be star (or another emission line object) requires visualisation of the zoomed profile of Balmer lines. The most prominent emission is shown in the H_α line at 6562.8Å. The spectral resolving power of about 2000 common to both surveys is satisfactory to distinguish even the double peak profile, although more details (e.g. shell lines) are not resolved.
The technology of Virtual Observatory allows the quick preview and visual marking of thousands of zoomed spectral lines in a SPLAT-VO viewer (see Fig. 5). To facilitate this, the VO-compatible archive of all spectra must be created combining Simple Spectral Access Protocol and DataLink to select the continuum normalised spectra and call on-line the spectral line cutout on the server. Such an archive was set up at Stellar Department of the Astronomical Institute of the Academy of Sciences of the Czech Republic in collaboration of Czech-VO with China-VO and LAMOST team .
The example of a emission star found in LAMOST survey is on Fig. 6.

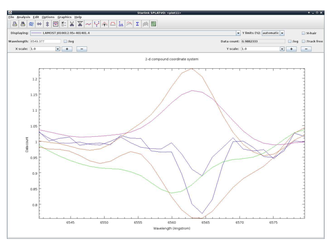

Fig. 5. Zoom on H_α line of LAMOST spectra in SPLAT-VO

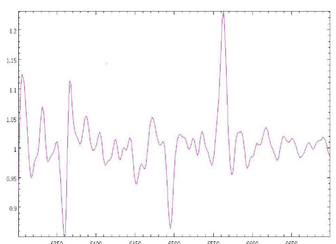

Fig. 6. Emission star identified in LAMOST survey

4 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-validation is applied until the system correctly recognises maximum of positive samples in any mixture of input vectors. Resulting classifier is applied on unknown spectra.

Fig. 7. Schema of supervised training using the Be stars sample

Fig. 8. Simple Gaussian fit of emission spectral line

We tried to use spectra of Be stars from Ondřejov 2m Perek telescope archive for training simple classifier based on two parameters (height, width) of a Gaussian line fit (after normalisation and convolution to spectral resolving power of SDSS) in order to find emission spectra in SDSS (see Fig. 7 and Fig. 8). The examples of Be star candidates found in SDSS SEGUE survey by this method is given on Fig. 9 and Fig. 10.

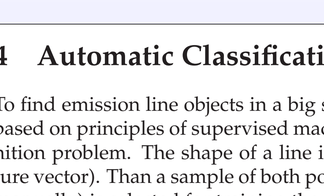

Fig. 9. Example Be star candidate found in SDSS

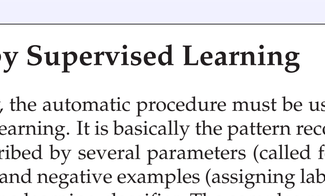

Fig. 10. Example Be star candidate found in SDSS

Various 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. Their advantage for application on big spectral archives is the possibility of their massive parallelisation, namely on GPUs.

5 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 identify outliers. So unknown rare objects with strange features hidden in the spectral archive, or even sources with yet undiscovered physical mechanism may be found using SOM.
The basic principle of SOM (see Fig. 12) is based on a set of artificial neurons (Fig. 11) connected in a special type of neuron network. Every input feature vector reinforces weights of interconnections leading to the neuron representing most similar features to the input ones. SOM is in fact a multi-dimensional topological map of such neurons projected in 2D space. 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).

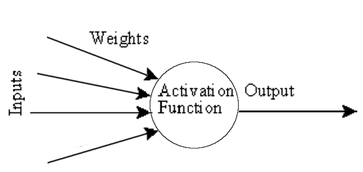

Fig. 11. Schema of artificial neuron

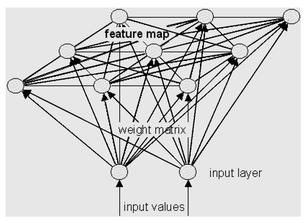

Fig. 12. Schema of construction of a SOM

In Fig. 13 we have fed the 30x30 neurons of SOM with almost 1700 spectra of both Be and ordinary stars from archive of 2m Ondřejov Perek Telescope, that were already visually classified into 4 classes with different shape of emission in H_α, including pure absorption. As it is seen, the most strange outliers (with manually assigned class 0, black colour) clearly clusters in four compact clusters. The characteristic shapes of each cluster are shown on Fig. 14. The advantage of SOMs for usage on big archives is their natural parallelizability and good scalability. However the big SOMs will hardly fit in memory of GPUs and so new specific algorithms for GPUs have to be created.

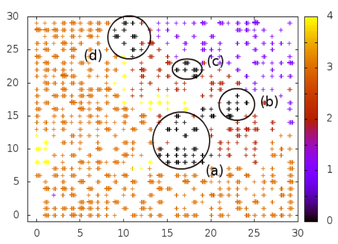

Fig. 13. Clusters of outliers in SOM

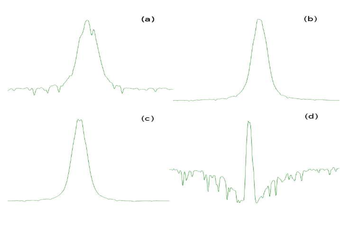

Fig. 14. Characteristic shapes of outliers in clusters marked black

6 Reduction of Dimensionality

Even the quick massively parallel computer or GPU cluster will not be able to process millions of several thousand pixels long feature vector in a necessary iterative way in a reasonable time. In order to be able to perform the unsupervised analyses of outliers in big spectra collections, the dimension of a feature vector must be reduced significantly, still conserving the most characteristic features of line shapes. Very common method of dimensionality reduction is the Principal Component Analysis (PCA) based on linear projection of a many dimensional feature space in a less dimensional space of Principal Components. On Fig. 15 is shown the separation of above mentioned emission and absorption shapes of H_α line. The Fig. 16 shows the emission outliers (yellow) mixed with other categories. Unfortunately PCA fails to distinguish them due to its linear nature. 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.

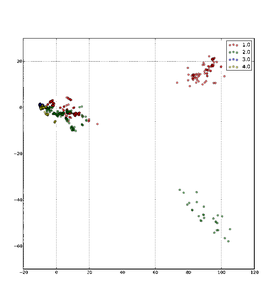

Fig. 15. Separation of line profiles shape in the PCA space

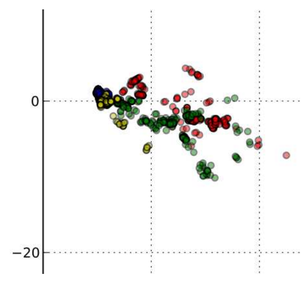

Fig. 16. Zoom of origin of PCA space

So another methods capable to emphasise the weight of the strictly localised features is needed. One of this is the Wavelet Transform (WT). All spectra are converted in a vector of wavelet coefficients (see Fig. 17) corresponding to different frequencies and next some aggregation function is applied (e.g. median, mean, maximum etc.). Experiments with clustering using such a combination shows high performance of this approach. We could still separate all classes of line shapes with accuracy better than 96% even if we degraded 2000 pixels long vector into 10 numbers using Wavelet Power Spectrum (see Fig. 18).

Fig. 17. Schema of construction of feature vectors using wavelet transform

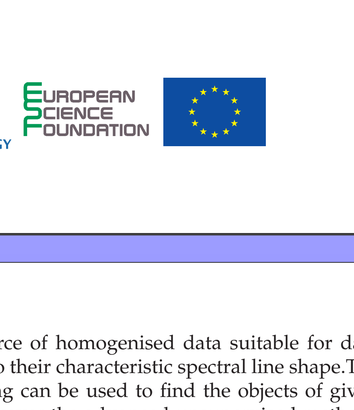

Fig. 18. Wavelet power spectra and first highest coefficients of selected line shapes

7 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

The four-year action starts in 2015 with kick-off meeting in December 2014. Everyone is welcome to join !

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 or quasars, however the advanced unsupervised methods like Self-Organising Maps help to identify outliers, even possibly yet unknown objects. The extremely long processing time of millions of spectra may become feasible with reduction of dimensionality of many point spectra to several elements of input feature vector, or by massively parallel processing, including GPUs. This, however, requires a change in so far commonly used algorithms in order to develop new massively parallelizable ones. As the astronomy is currently facing similar Big Data problems as other natural sciences, a wide collaboration between informaticians, astronomers and Earth scientists has been opened in the framework of European COST Action TD1403.

Acknowledgements

This work was supported by grants 13-08195S of Czech Science Foundation as well as the project RVO:67985815. For this research were used a number of spectra from Ondřejov 2m Perek telescope, public LAMOST DR1 survey and Sloan Digital Sky Survey.