Identification of Interesting Objects in Large Spectral Surveys Using Highly Parallelized Machine Learning

Petr Škoda[1], Andrej Palička[2], Jakub Koza[2] and Ksenia Shakurova[2]

[1]Astronomical Institute of the Czech Academy of Sciences
Fričova 298, 251 65 Ondřejov, Czech Republic
email: `skoda@sunstel.asu.cas.cz`

[2]Faculty of Information Technology, Czech Technical University in Prague
Thákurova 9, 160 00 Prague 6, Czech Republic

Abstract. The current archives of LAMOST multi-object spectrograph contain millions of fully reduced spectra, from which the automatic pipelines have produced catalogues of many parameters of individual objects, including their approximate spectral classification. This is, however, mostly based on the global shape of the whole spectrum and on integral properties of spectra in given bandpasses, namely presence and equivalent width of prominent spectral lines, while for identification of some interesting object types (e.g. Be stars or quasars) the detailed shape of only a few lines is crucial. Here the machine learning is bringing a new methodology capable of improving the reliability of classification of such objects even in boundary cases.

We present results of Spark-based semi-supervised machine learning of LAMOST spectra attempting to automatically identify the single and double-peak emission of H_α line typical for Be and B[e] stars. The labelled sample was obtained from archive of 2m Perek telescope at Ondřejov observatory. A simple physical model of spectrograph resolution was used in domain adaptation to LAMOST training domain. The resulting list of candidates contains dozens of Be stars (some are likely yet unknown), but also a bunch of interesting objects resembling spectra of quasars and even blazars, as well as many instrumental artefacts. The verification of a nature of interesting candidates benefited considerably from cross-matching and visualisation in the Virtual Observatory environment.

Keywords. stars: emission-line, Be, surveys, methods: statistical, techniques: spectroscopic

1. Introduction

There are many objects in the Universe that may show interesting shapes of some important spectral lines. The objects presenting emission lines, as are Be stars, where a gaseous envelope in the shape of a sphere or a disk is expected (Porter & Rivinius 2003), and rare class of B[e] stars showing infrared excess (Zickgraf 2003), are especially interesting, thanks to their complicated physics. The emission lines in spectra of such objects may present under different physical conditions single peak, double peak with different ratios of components or even complicated combined emission and absorption profiles (Silaj *et al.* 2010). To find emission line objects with given shape of spectral line in a big survey, the automatic procedure must be used based on principles of supervised machine learning.

2. Machine Learning

Machine learning is the field of informatics, closely related to the advanced statistical inference, which tries to build models of data by learning from sample inputs and

make predictions based on such learned models. It is divided mainly into supervised and unsupervised methods with a number of subclasses.

Supervised: These algorithms have some prior knowledge about the data, that was supplied by some external means. This is usually done by a human domain expert. Main representatives are the *classification* tasks, where the models classify data into some predefined, finite set of classes, and *regression* tasks, which infer output of a real-valued function.

Unsupervised Algorithms in this group do not have any prior knowledge of the data. They attempt to discover corresponding relationships themselves. This is by definition a much harder task than supervised learning, however it can potentially lead to much more interesting results. Example tasks are *clustering*, *automated feature selection* or *outliers detection*.

A special category is a **Semi-Supervised Learning**. It is a group of algorithms designed to work on datasets, that have very few labelled data points compared to the amount of unlabelled points. It resembles a supervised learning in that we have labelled data, however with an addition of also having unlabelled data, which may help us estimate the distribution of the data set more precisely. There is, however, another possible variant of semi-supervised learning. Here, we are not using labelled data, but merely some constraints. These constraints may link some points, that share the same label, or they may reveal the actual number of classes. This resembles unsupervised learning, however with some a priori information about the data.

In big spectral archives, where it is almost impossible to investigate every spectrum visually, the yet unknown rare objects with strange features, or even sources with yet undiscovered physical mechanism may be in principle found using machine learning.

3. LAMOST Spectral Surveys

The LAMOST telescope (Cui *et al.* 2012) has been delivering one of currently largest mega-collections of spectra (similar to Sloan Digital Sky Survey). The sixteen LAMOST spectrographs are fed by 4000 fibres positioned by micro-motors. Its publicly accessible data release DR1 (Luo *et al.* 2015) contains altogether **2 204 696** spectra, with a spectral resolving power ~ 1800, covering the range 3690-9100Å. The LAMOST pipeline classified **1 944 329** of them as stellar ones.

4. Ondřejov CCD700 Archive

The unique source of the spectra of emission line stars (mostly Be and several B[e]) is the archive of spectra obtained with 700mm camera of the coudè spectrograph of the 2m Perek Telescope at Ondřejov observatory, a part of the Astronomical Institute of the Czech Academy of Sciences. The archive (named CCD700) contains about twenty thousand spectra of mainly Be stars and other emission-line objects exposed mostly in spectral range 6250–6700Å with spectral resolving power ~ 13000.

4.1. *Cross-matching and Labelling of LAMOST and CCD700 Archives*

Using the technology of Virtual Observatory, namely the combination of Table Access (Nandrekar-Heinis *et al.* 2014) and Simple Spectra Access (Tody *et al.* 2012) protocols we have identified less than ten objects (e.g. BT CMi or HD53416) observed both by Ondřejov 2m Perek Telescope and LAMOST. This limits the usage of normal supervised training due to the lack of labelled spectra in LAMOST DR1 selected for training set. So the semi-supervised method must be used. In CCD700 archive we have visually identified

1696 spectra with clearly defined classes of spectral shapes. The most important part consisted of spectra of selected Be stars with different spectral line profiles as double peak emission, emission with central absorption or deep absorption with small single or double emission peak. All such interesting cases were given label 1 (target class). They were complemented by absorption spectra of many classical stars labelled 0. This sample became a training set. The spectra, however, could not be used immediately due the the different spectral resolving power of CCD700 archive and LAMOST. We needed to perform a *Domain Adaptation* of training sample.

4.2. *Domain Adaptation of CCD700 Spectra*

Before starting the semi-supervised learning, all Ondřejov spectra with higher resolution must be converted to the same resolution as LAMOST data. In other words the samples from one domain (CCD700) are transformed to second domain (LAMOST), so that the CCD700 spectra of selected emission stars will look like exposed with LAMOST spectrograph. Here we use the model based on our (simplified) physical knowledge of the principles how the spectrograph works. The spectral resolving power degradation may be roughly approximated by the convolution with the Gaussian kernel with Full Width at Half Maximum (FWHM) proportional to the ratio of resolving powers. In our case we have used Gaussian with FWHM of 5 pixels. Finally, the spectrum has to be re-binned into larger pixels, as the LAMOST spectrum contains less pixels over the same spectral range. Such simulated spectra are treated as the original data for the input preprocessing. See Fig. 1 for an example of resolution degradation. The top spectrum is an original CCD700 spectrum of HD53416, the bottom one its convolved version (please notice the smearing of the double peak profile), while the middle one is the same star observed with LAMOST.

5. Input Spectra Preprocessing

An important part of data preparation before applying machine learning is the data pre-processing. In our case all the spectra had to be normalised to the continuum (rectified), cut to the same wavelength range 6250–6750Å (for CCD700 also convolved with Gaussian as explained above) and re-binned into the same grid of wavelength points. This gave us the number of so called Feature Vectors (FV). The result of the preprocessing is the Comma-separate values (CSV) file with all spectral intensities interpolated to the same wavelength grid mapped to the individual elements of the FV. Here the value of physical wavelength in Å is unknown and must be reconstructed for visualisation from the mapping metadata. This CSV file is uploaded to a computing cluster running the semi-supervised machine learning algorithms .

6. Semi-Supervised Machine Learning

We have used two graph-based methods, described in Chapelle *et al.* (2006), which were adopted for parallelisation on Spark engine. The full implementation details are explained in Palička (2016).

Label Propagation Label propagation is an algorithm that leverages the graph representation of data to fit a model. Labels are encoded as an one-hot variable, so that we may support multi-class classification. The algorithm basically computes weights for labels for each data point based on the distance to its neighbours. Note that in our version of the algorithm, the initial labels do not change and are reset to their original value in each step.
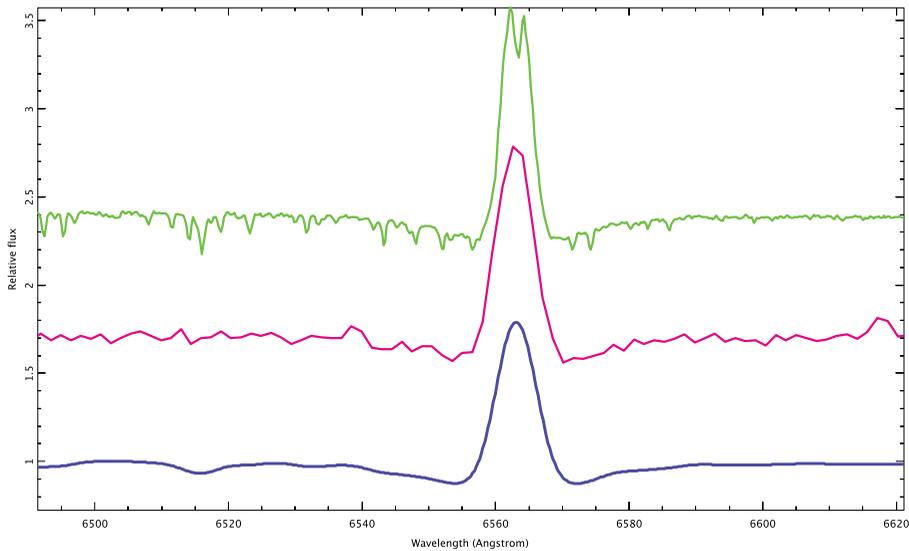
Figure 1. Comparison of the original (top) and convolved (bottom) CCD700 spectrum of HD53416 with one observed by LAMOST (middle)

Label Spreading is a similar algorithm to label propagation. It uses the normalised graph Laplacian to propagate the label information across the graph. It also allows the labels to retain some partial information from the initial labelling.

7. Massively Parallelised Processing Using Spark

The Apache Spark (`http://spark.apache.org`) is a cluster computing technology allowing the fast computation on number of computing nodes in parallel. We have used the academical cluster MetaCentrum consisting of twenty-four sixteen-core nodes (the number of nodes assigned by the system is however unknown, dependent on a availability and load of the cluster). The data were distributed across all nodes by Hadoop Distributed File System — HDFS (`http://hadoop.apache.org`). The search was run on more than fifty thousand spectra randomly selected from those labelled as star by LAMOST DR1 pipeline.

8. Results

The result of running the semi-supervised machine learning on a set of feature vectors is a confidence score of each — the number expressing the weight of the label assigned by the algorithm to the unlabelled data. By thresholding the highest confidences we get the most probable candidates of our target class (in our case the emission line spectra). Our experiments run many times of subsamples of LAMOST DR1 gave us a short list of interesting spectra which were previewed to eliminate artifacts. The final list contains tens of very interesting candidates of emission line stars deserving further investigation. On Fig. 2 and Fig. 3 are given examples of two different profiles of Be stars found by semi-supervised training (the whole spectrum and zoomed one centered around H_α line). The Fig. 4 shows another interesting object with emission both in Hydrogen and Helium 6678Å lines. This configuration of emission line profiles closely resembles the well-known
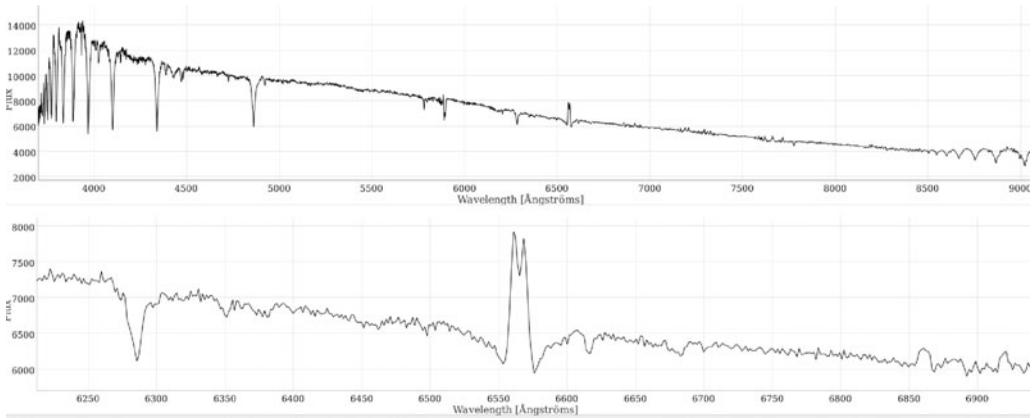
Figure 2. Example of Be star identified in LAMOST DR1 by machine learning: Strong double peak emission
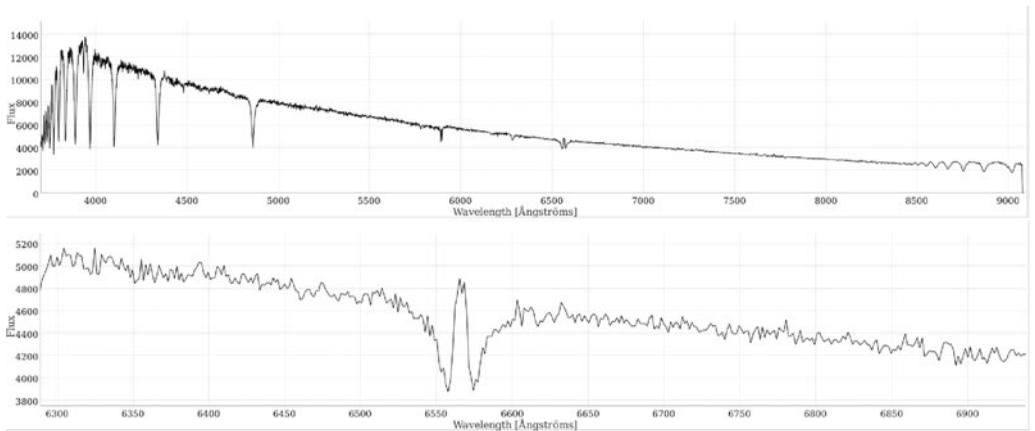
Figure 3. Example of Be star identified in LAMOST DR1 by machine learning: Absorption line with weak central emission peak
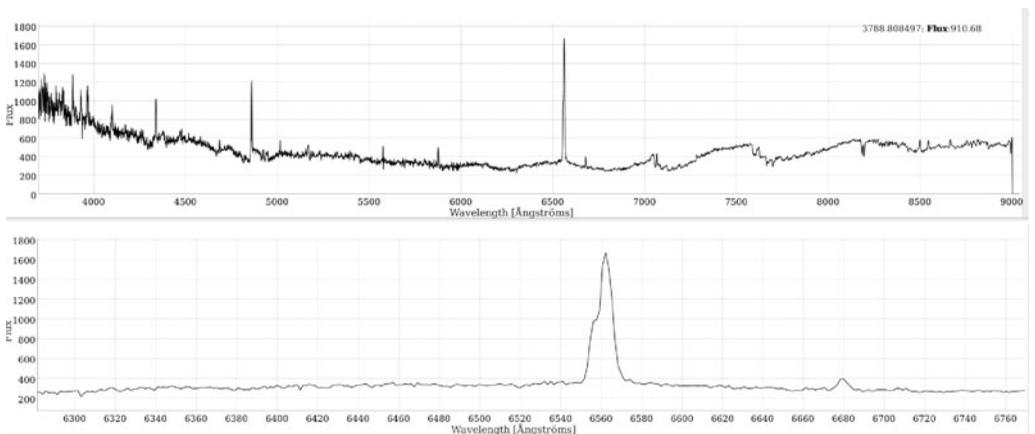
Figure 4. LAMOST star with line profiles similar to Beta Lyrae
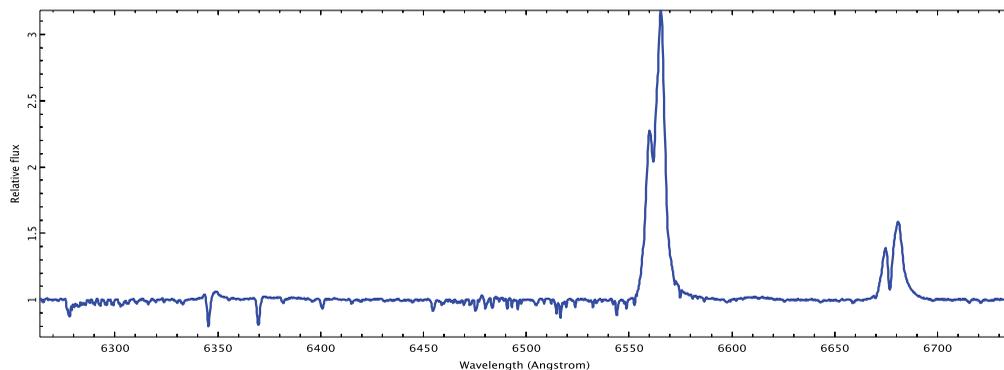
Figure 5. Beta Lyrae from Ondřejov CCD700 archive

Beta Lyrae star. For comparison the spectrum of Beta Lyrae from Ondřejov CCD700 archive is given on Fig. 5.

9. Conclusions

Big spectral archives such as LAMOST DR1 are good source of data suitable for machine learning of interesting objects according to their characteristic spectral line shape. Examples of objects with emission lines found by our method confirm, that the whole idea of machine learning applied to spectral line profiles is viable and the methods described above are able to identify object of given spectral properties. The identified candidates with emission profile need further detailed examination as they may hide interesting scientific objects. The semi-supervised methods may benefit considerably from massive parallelisation using Spark on Hadoop cluster.

Acknowledgements

References

Chapelle, O., Schölkopf, B., & Zien, A. 2006, *Semi-supervised learning*, MIT press, Cambridge, Massachusetts
Cui, X. Q., *et al.* 2012, *Research in Astronomy and Astrophysics*, 12, 1197
Luo, A. L., *et al.* 2015, *Research in Astronomy and Astrophysics*, 15, 1095
Nandrekar-Heinis, D., Michel, L., Louys, M., & Bonnarel, F. 2014, *Astronomy and Computing*, 7, 37
Palička, A. 2016, *Master Thesis*, Czech Technical University in Prague, Faculty of IT
Porter, J. M. & Rivinius, T. 2003, *PASP*, 115, 1153
Shakurova, K. 2016, *Master Thesis*, Czech Technical University in Prague, Faculty of IT
Silaj, J., Jones, C. E., Tycner, C., Sigut, T. A. A., & Smith, A. D. 2010, *ApJS*, 187, 228
Tody, D. *et al.* 2012, *IVOA Recommendation: Simple Spectral Access Protocol Version 1.1*, ArXiv:1203.5725
Zickgraf, F.-J. 2003, *A&A*, 408, 257