# Identification of Artifacts and Interesting Celestial Objects in LAMOST Spectral Survey


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

[1]Astronomical Institute of the Czech Academy of Sciences
Ondřejov, Czech Republic

[2]Faculty of Information Technology of the Czech Technical University
Prague, Czech Republic



## Abstract

The LAMOST DR1 survey contains about two million of spectra labelled by its pipeline as stellar objects of common spectral classes. There is, however, a lot of spectra corrupted in some way by both instrumental and processing artifacts, which may mimic spectral properties of interesting celestial objects, namely emission lines of Be stars and quasars.

We have tested several clustering methods as well as outliers analysis on a sample of one hundred thousand spectra using Spark scripts running on Hadoop cluster consisting of twenty-four sixteen-core nodes. This experiment was motivated by an attempt to find rare objects with interesting spectra as outliers most dissimilar from all common spectra.

The result of this time-consuming procedure is a list of several hundred candidates where different artifacts are prominent, but also tens of very interesting emission-line spectra requiring further detailed examination. Many of them may be quasars or even blazars as well as yet unknown Be-stars. It deserves mentioning that most of the work benefitted considerably from technologies of Virtual Observatory.


## 1 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 (see Fig. 1) positioned by micro-motors. Its publicly accessible Data Release 1, (see Luo et al., 2015)) contains altogether **2 204 696** spectra, **1 944 329** of them being classified by the LAMOST pipeline as stellar ones.

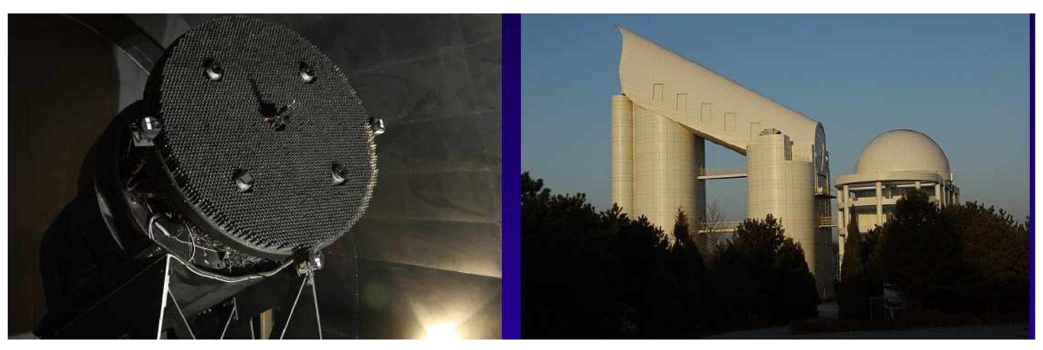

*Fig. 1. LAMOST telescope and its focal plane with fibres moved by micro-motors*

## 2 Emission Line Objects

There is a lot of objects in the Universe that may show interesting shapes of some important spectral lines. Very interesting are objects presenting emission lines, as are Be stars, cataclysmic variables or quasars, where a gaseous envelope in the shape of a sphere or a disk is expected. The emission lines may present under different physical conditions single peak, double peak with different ratios or even complicated combined emission and absorption profiles.

The unique source of such spectra is the archive of spectra obtained with 700mm camera of the coudè spectrograph of the 2m Perek Telescope at Ondřejov observatory — 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 in spectral range 6250–6700 Å with spectral resolving power about 13000. The examples of various emission line profiles of $H_\alpha$ line taken from CCD700 archive are given on Fig. 2.

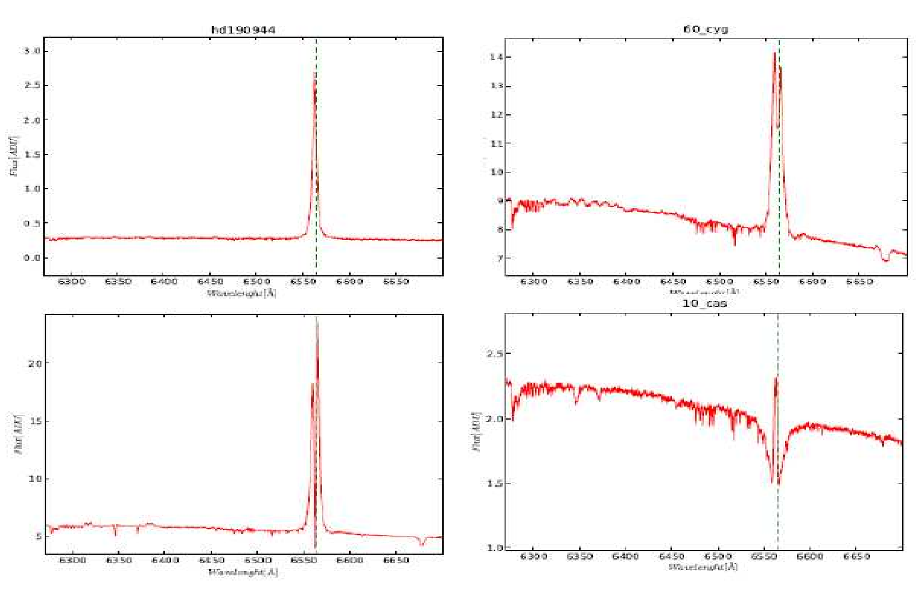

*Fig. 2. Examples of $H_\alpha$ line profiles Be stars in Ondřejov CCD700 archive*

## 3 Finding Outliers with Unsupervised 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.

Unsupervised learning (unlike supervised one requiring the labels assigned to part of data) tries to identify similar patterns (typically different clusters based on some similarity metrics) in data automatically without the human intervention. The outliers are entities which cannot be assigned to any of such cluster (so they represent the single member clusters).

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 a lot of random instrumental artifacts will be found as well as every one is unique and thus very rare. The artifacts caused by systematic errors of the same nature, which repeats very often, may be collected by clustering as well.

## 4 LOF Method for Finding Outliers

The Local Outlier Factor method (LOF) introduced by Breunig et al. (2000) is based on an idea to compare local density of an object to the local densities of its neighbours. The local density is estimated by the typical distance $\varepsilon$ at which a point can be "reached" from its neighbors (see Fig. 3).

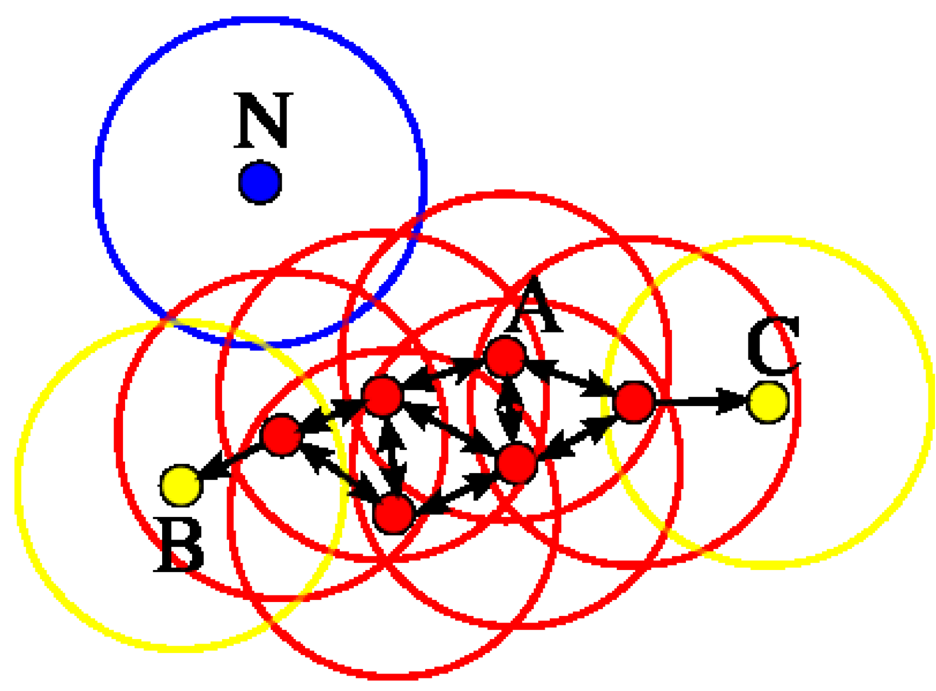


*Fig. 3. Concept of the local density for the minimal number of required points equal to three. Point A and the other red points are core points, because at least three points surrounding them are in an $\varepsilon$ radius. Points B and C are not core points but are reachable from A (via other core points) and thus belong to the cluster as well. Point N, the outlier, is more distant (typically a noise)*

One of the key terms for LOF is the *k-distance* and *reachability distance* of $k$ nearest neighbours: for any $k > 0$ the k-distance of object $p$ is the distance $d(p, o)$ between $p$ and an object $o \in D$ such that:

- for at least $k$ objects $o' \in D \setminus p$ it holds that $d(p, o') \leq d(p, o)$;
- for at most $k - 1$ objects $o' \in D \setminus p$ it holds that $d(p, o') < d(p, o)$.

It means the distance of the object $p$ to the $k$-th nearest neighbor, but set of the k nearest neighbor ($N_k(p)$) includes all objects at this distance (it can contain more than $k$ objects). Using *k-distance* the *reachability distance* can be defined as

$$\textit{reach-distance}_k(p, o) = \max(\textit{k-distance}(o), d(p, o)) \tag{1}$$

The *local reachability density* of object p is defined as

$$lrd_k(p) = 1/\left(\frac{\sum_{o \in N_{k(p)}} \textit{reach-distance}_k(p, o)}{|N_k(p)|}\right) \tag{2}$$

The local outlier factor of p is defined as

$$LOF_k(p) = \frac{\sum_{o \in N_{k(p)}} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|} \tag{3}$$

If the LOF is considerably larger than 1, the object is an outlier, if it is about 1, the object is comparable with others.

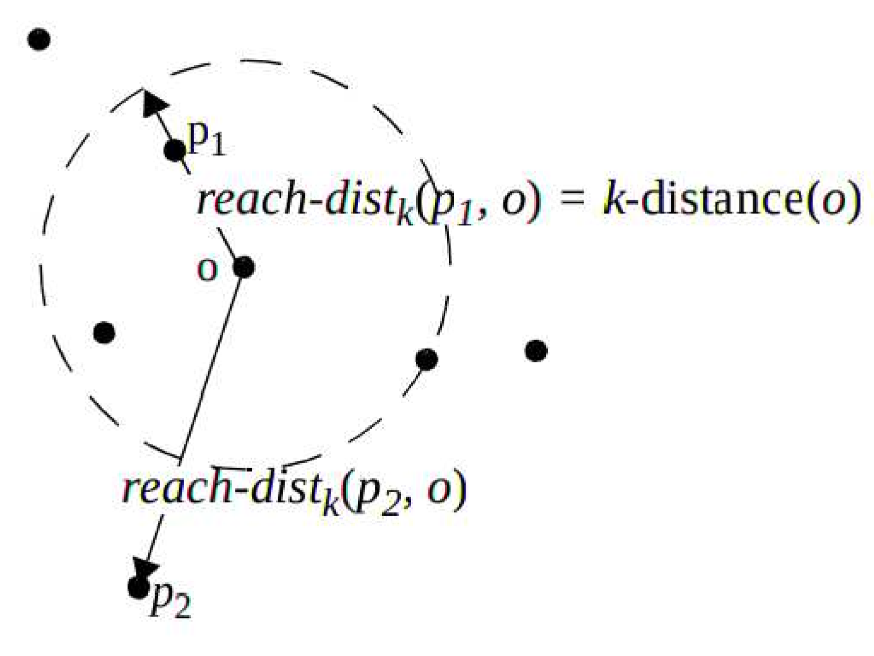


*Fig. 4. Concept of reachability distance. reach-dist($p_1$,o) and reach-dist($p_2$,o), for k=4*

## 5 Input Spectra and their Pre-processing

The important part of data preparation before applying machine learning is the data pre-processing. In our case the spectra have to be normalized to the continuum (rectified), cut to the same wavelength range and re-binned into the same grid of wavelength points. This gives us the number of so called Feature Vectors (FV). As we wanted compare the some algorithm on both Ondřejov CCD700 and LAMOST spectra, we have cut the LAMOST ones to the similar wavelength range (about 6250–6750 Å) as those from CCD700. The result of the preprocessing is the big CSV file with all spectral intensities interpolated to the same wavelength grid. This (big) CSV is loaded on a computing cluster.

## 6 Massively parallelized processing using Spark

The Apache Spark is a set of libraries written in SCALA language, adapted for calling from PYTHON, running 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 HDFS filesystem of Hadoop infrastructure. The special Spark-based version of LOF method was developed by K.S. (Shakurova, 2016) for this task. The experiments were run on all, almost 20 000 Ondřejov CCD700 spectra and then on about 120 000 spectra randomly selected from those labelled as star in LAMOST DR1.

## 7 Resulting Outliers

As it is seen on Fig. 5, the LOF method is able to find in the CCD700 archive all interesting cases of spectra like sharp emissions, asymmetric double peak emission or even noisy late type stars spectra. On the LAMOST data contaminated by a lot of spoiled spectra, it can identify those with random instrumental artifacts (see Fig. 6). However some spurious ones may be still valuable as they may represent interesting objects deserving further investigation. An example on Fig. 7 shows the noisy spectrum classified by the LAMOST pipeline as late type M7 class star, however clearly showing the combination of absorption and emission profile seen typically in Be stars in both observed Oxygen OI lines (see Fig. 8).

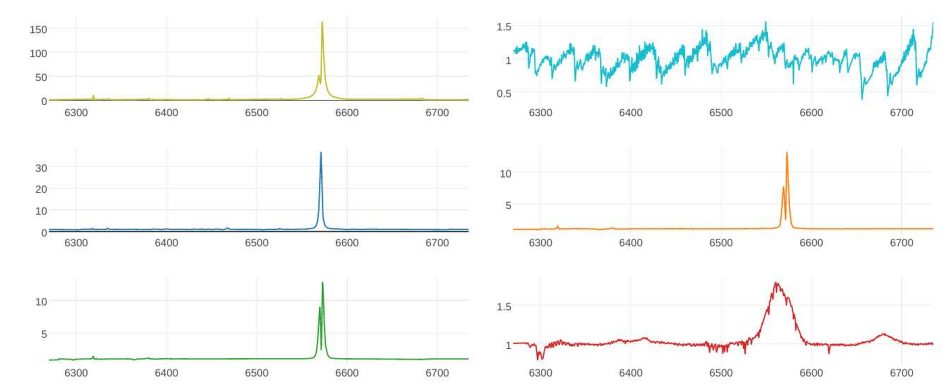

*Fig. 5. Examples of LOF outliers found in CCD700 archive.*

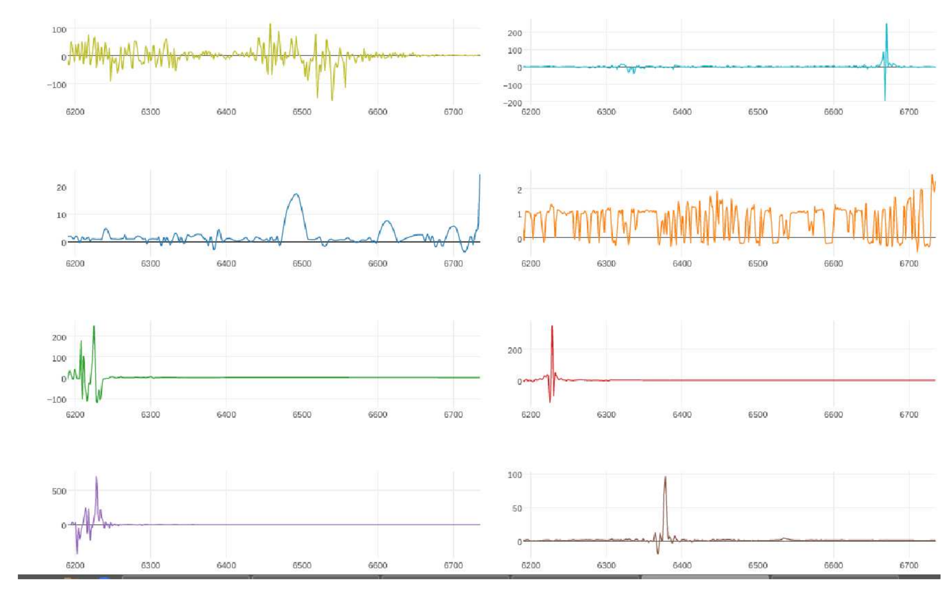

*Fig. 6. Examples of LAMOST outliers found by LOF*

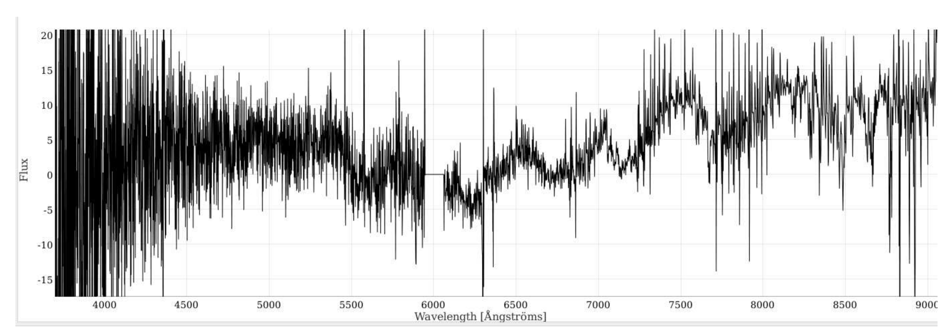

*Fig. 7. LAMOST outlier classified as the M7 star*

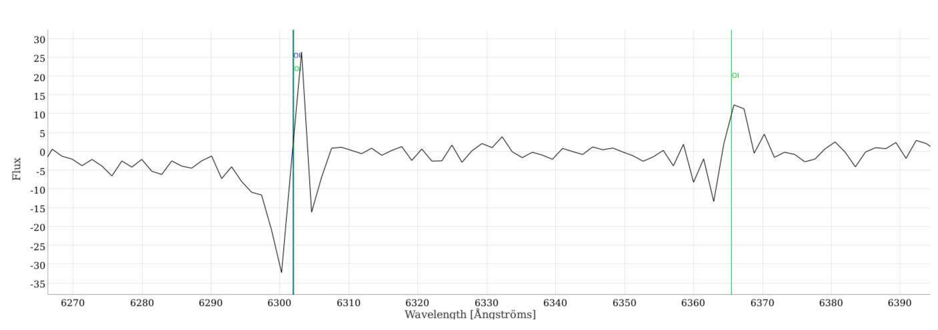

*Fig. 8. Detailed view of emission in OI lines of the LAMOST outlier*

## 8 Conclusions

Big spectral archives are good source of data suitable for machine learning of interesting objects according to their characteristic spectral line shape. The outlier finding methods as LOF may be successfully used for searching instrumental artifacts but also the results need further detailed examination as they may hide interesting scientific objects. The application of the method may benefit considerably from massive parallelization using Spark on Hadoop cluster.


## Acknowledgements

This work was supported by grant LD-15113 of Ministry of Education, Youth and Sports of the Czech Republic. This research 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 greatly appreciated.



## References

M. Breunig, M., H. P. Kriegel, R. T. Ng, and J. Sander. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.

X.-Q. Cui et al. *Research in Astronomy and Astrophysics*, 12:1197–1242, 2012.

A.-L. Luo et al. *Research in Astronomy and Astrophysics*, 15:1095, 2015.

K. Shakurova. Unsupervised learning and outlier detection in large archives of astronomical spectra. Master thesis, Czech Technical University in Prague, Faculty of Information Technology, 2016.



Poster presented at ADASS 2016, 16 – 20th October 2016, Trieste, Italy