Statistical Challenges in 21st Century Cosmology
Proceedings IAU Symposium No. 306, 2014
A. F. Heavens, J.-L. Starck & A. Krone-Martins, eds.

doi:10.1017/S1743921314013453

Data-mining Based Expert Platform for the Spectral Inspection

Haijun Tian[1,2], Yang Xu[1], Yang Tu[2], Yanxia Zhang[1], Yongheng Zhao[1], Guohong Lei[2], Boliang He[1], Chenzhou Cui[1] and Xuelei Chen[1]

[1]National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012.

[2]China Three Gorges University, Yichang, 443002. Email: hjtian@lamost.org

Abstract. We propose and preliminarily implement a data-mining based platform to assist experts to inspect the increasing amount of spectra with low signal to noise ratio (SNR) generated by large sky surveys. The platform includes three layers: data-mining layer, data-node layer and expert layer. It is similar to the GalaxyZoo project and it is VO-compatible. The preliminary experiment suggests that this platform can play an effective role in managing the spectra and assisting the experts to inspect a large number of spectra with low SNR.

Keywords. techniques: photometric, methods: data analysis, catalogs

1. Introduction

With the telescopes established and the surveys ongoing, such as the Guoshoujing telescope (LAMOST, Zhao 1999), more and more spectra are collected and released. Unfortunately, except the qualified data, there still exist many spectra unclassified by the automated pipeline. Usually most of these spectra have too low SNR to be classified because of the limiting magnitude of the instruments or other reasons. Some important new discoveries may probably be hidden in these unknown spectra. Therefore, we should not throw away these seemingly useless data, even though they are quite defective. How to handle these unknown spectra is one of the biggest challenges to the modern statistics, data mining techniques as well as eyeball check. In order to ensure the accuracy of the results, we have to motivate users to check these spectra by visual inspection. Owing to huge amount of such spectra generated continuously by the large sky surveys, much time and efforts is needed to check spectra one by one. Consequently, a platform to efficiently manage and coarsely classify these unqualified data is in great requirement for large surveys.

2. Architecture

We propose such a platform, which aims to effectively process and manage a large volume of spectra with low SNR, and mine as much as possible knowledge from the spectra for the scientific research through the integration of machine learning and human inspection. The platform is a three-tier structure including data-node layer, data mining layer and expert layer, as shown in Fig. 1. The core part is data mining. Before the visual inspection, the dataset will be preprocessed with various up-to-date statistical and data-mining techniques. For instance, the dataset can be classified roughly as stars, galaxies or quasars according to the limited information contained in the low SNR spectra. All results from this layer could efficiently help the expert to inspect the spectra. The expert layer provides an interface by which anyone may query and inspect the spectra and save

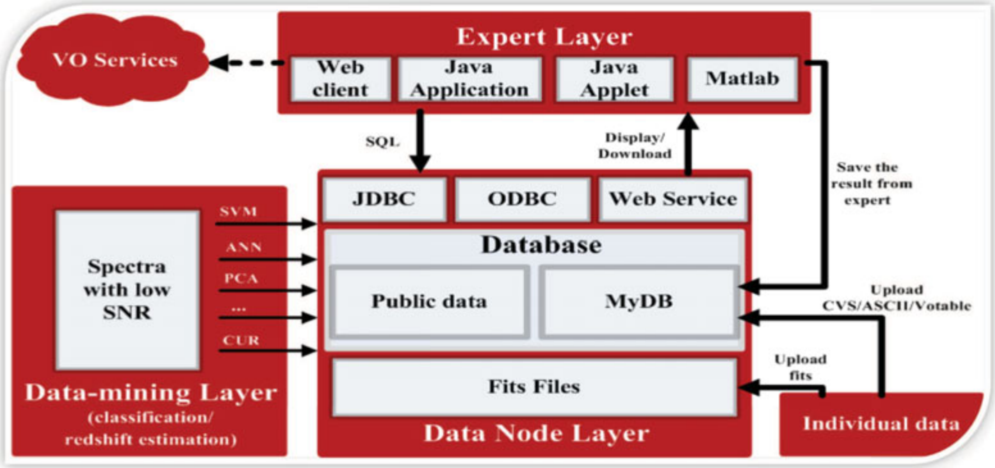

Figure 1. Architecture of expert platform for the spectral inspection

the results in the data node layer. Finally the system can give a more reliable result by efficiently utilizing the experts' input (one spectrum may be inspected by more than one expert). The platform will be VO-compatible and integrated seamlessly with other surveys. Its client will be of diverse forms, such as Matlab-GUI, Java-application, Java-applet, etc.

Several kinds of data-mining algorithms to pre-process the LAMOST spectra have been investigated. For example, Support Vector Machines (SVMs) are first applied to classify the LAMOST spectra, and the classified spectra are saved into the data-node layer as a public dataset. Then the spectra may be queried with a web client and inspected by a Java applet tool, finally the analysis results are returned back into MyDB in the data-node layer.

Data-mining layer. The data-mining layer has two main tasks: classification and redshift or radial velocity estimation for low SNR spectra using various advanced data-mining techniques, such as Support Vector Machines (SVMs, Peng *et al.* 2012), Artificial Neural Networks (ANN, Zhang *et al.* 2009), K-Nearest Neighbors (KNN, Li *et al.* 2008), Principal Component Analysis (PCA, Zhang & Zhao 2003; Yip *et al.* 2004), Wavelet (Machado *et al.* 2014), CUR Matrix Decomposition (Yip *et al.* 2014).

DataNode layer. The data-node layer include two parts: one is based on the database, which manages the catalog generated by the data-mining layer (public dataset available for each user) and the private dataset input by the users (MyDB, visible only for the data owner); the other needs enough storage space to save the spectral fits files, which are also divided into the public and private. The results from the expert layer can also be saved in MyDB as the private file. The data-node layer provides two interfaces, e.g. the database interface (JDBC or ODBC) and the Web Service interface, by which expert layer can query catalogs and retrieve spectra.

Expert layer. The clients with diverse forms, such as Matlab-GUI, Java-application, Java-applet, etc., can be used to display the spectra and assist the experts to inspect the spectra in this layer. These tools provide all kinds of spectral templates and reference lines, which can be easily shaped at expert's will. The redshift can be calculated in real time as the template is moved. The templates will automatically be best-matched based on the least square method, and the optimal redshift can be figured out when the spectrum is best fitted with the template.

https://doi.org/10.1017/S1743921314013453 Published online by Cambridge University Press

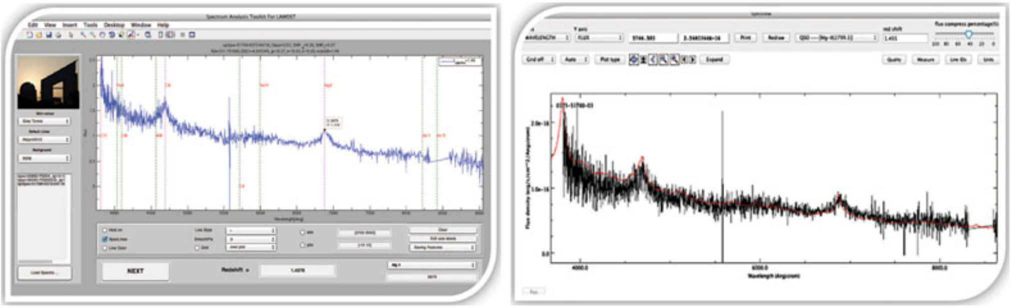

Figure 2. Matlab (left) and Java (right) clients for the spectral inspection.

The clients will support two kinds of queries: one creates simple SQLs through choosing the fields and filling the query criteria; the other one provides a input form, that allows users to compose much more complex SQL statements, and also provides the functions of checking query syntax and database privilege. Through the JDBC, web service or VO interfaces, the clients can access the local or remote databases. Experts with different levels have different weights to contribute the final results.

For the data-node layer, we have finished the database and file systems with public data and MyDB, and released the JDBC and Web Service interfaces for the query functions.

For the expert layer, a Matlab client and a Java client have been preliminary developed, as displayed in the Fig. 2. Actually, the Java client is modified SpecView, an interactive java tool for visualization and analysis of spectral data (Busko 2000), in which we extended some functions including the template fitting, automatic redshift estimation, the interaction with the data-node layer, etc.

3. Conclusions

In order to overcome the difficulty in processing the poor spectra, we propose and develop a data-mining based platform to encourage and assist experts to inspect the low SNR spectra. With the help of the platform, we tentatively process LAMOST spectra. The preliminary experiment indicates that this platform indeed can play an effective role in assisting experts to recognize a large number of spectra with low SNR.

4. Acknowledgments

The authors thank the grants (No. U1231123, U1331202, U1331113, U1231108, 11073024, 11103027, 11303020) from NSFC. THJ thanks the support from LAMOST Fellowship.

References

Busko, I. 2000, *ASPC*, 216, 79B

Li, L. L., Zhang, Y. X., & Zhao, Y. H. 2008, *Science in China*, G, 51(7), 916

Machado, D., Leonard, A., Starck, J. L., & Abdalla, F. 2014, *A&A*, in press

Peng, N., Zhang, Y. X., Zhao, Y. H., & Wu, X. B. 2012, *MNRAS*, 425, 2599

Yip, C. W., Connolly, A. J., Vanden Berk, D. E., *et al.* 2004, *AJ*, 128, 2603Y

Yip, C. W., Mahoney, M. W., Szalay, A. S., Csabai, I., & Budavári, T., *et al.* 2014, *AJ*, 147, 110Y

Zhao, Y. H. 1999, *PYunO*, S, 1Z

Zhang, Y. X., Li, L. L., & Zhao, Y. H. 2009, *MNRAS*, 392, 233

Zhang, Y. & Zhao, Y., 2003, *PASP*, 115, 1006

https://doi.org/10.1017/S1743921314013453 Published online by Cambridge University Press