

# Searching of new Emission-line Stars using the Astroinformatics Approach Petr Škoda and Jaroslav Vážný

Astronomical Institute of the Academy of Sciences, Ondřejov, Czech Republic

Abstract: Using data mining techniques applied on emission line characteristics of Be stars spectra we attempted to find new Be stars candidates in SDSS SEGUE survey. The mid-resolution spectra of confirmed Be stars obtained from VO-compatible archive of Ondřejov observatory 2m telescope were transformed to the spectral resolution of SDSS and important characteristics of emission line profiles were estimated, to be used as a training base of supervised learning methods. The obtained knowledge base of the characteristic shapes and sizes of Be emission lines was finally used to identify new potential candidates in SDSS spectral survey. The several newly found Be stars candidates justify our approach and approve Astroinformatics as a viable research methodology.

## **Process Overview**

## **Degradation of Spectral Resolution**

## **Data Mining**

Schema of the process is on the Fig. 1. Using SSA protocol the spectra from Ondřejov 2m telescope archive server were acquired based on the list of justified Be stars obtained from other studies. Convolution of Ondřejov spectra with SDSS instrumental profile had to be performed to ensure the compatibility with the lower spectral resolution of SDSS. Then the desired features were extracted automatically from the spectra after the continuum normalisation and H $\alpha$  line was fitted by appropriate function. The same was done for spectra from SDSS except the convolution process. Thus the vectors of parameters characterising the typical Be star H $_{\alpha}$  emission line were obtained and subjected to data mining process. Spectra from Ondřejov Observatory have higher spectral resolution than SDSS, therefore the degradation of spectral resolution was applied on them followed by re-binning to the same The decision tree based classification was performed using Weka software with algorithm J48, which is the free implementation of algorithm C4.5. The training set had 173 and testing set



number of pixels as the SDSS. So we obtained the training set of Ondřejov Be stars spectra looking similar to SDSS spectra.

For that purpose convolution in discrete form was used

$$(f * g)[n] \stackrel{\text{def}}{=} \sum_{m=-\infty}^{\infty} f[m] g[n-m]$$

An example of this process applied on spectra of Be star 4 Her is on the Fig. 2. The top figure shows Gaussian function used for convolution with the spectrum, followed by the original spectrum, then there is a spectrum after convolution with the Gaussian profile. The last is the final spectrum after re-binning



178314 items.

J48 pruned tree

| === Summary ===                    |           |          |
|------------------------------------|-----------|----------|
| Correctly Classified Instanc       | es 145    | 83.815 % |
| Incorrectly Classified Insta       | nces 28   | 16.185 % |
| Kappa statistic                    | 0.6529    |          |
| Mean <b>absolute</b> error         | 0.1849    |          |
| Root mean squared error            | 0.3652    |          |
| Relative absolute error            | 39.8819 % |          |
| Root <b>relative</b> squared error | 75.8919 % |          |
| Total <b>Number</b> of Instances   | 173       |          |

| 2  |                                       |
|----|---------------------------------------|
| 3  | <b>max</b> <= -0.18843                |
| 4  | <b>max</b> <= −0.324763: o (46.0/5.0) |
| 5  | max > -0.324763                       |
| 6  | <b>max</b> <= -0.255475               |
| 7  | mad <= 0.004133: o (2.0)              |
| 8  | mad > 0.004133: be (13.0/1.0)         |
| 9  | max > -0.255475                       |
| 10 | mad <= 0.009862: o (10.0)             |
| 11 | mad > 0.009862                        |
| 12 | width <= 7.621593: o (3.0/1.0)        |
| 13 | width > 7.621593: be (2.0)            |
| 14 | <b>max</b> > -0.18843                 |
| 15 | mad <= 0.030316                       |
| 16 | <b>max</b> <= -0.091726               |
| 17 | width <= 5.286489                     |
| 18 | <b>max</b> <= -0.170022: be (2.0)     |
| 19 | max > -0.170022: o (3.0)              |
| 20 | width > 5.286489: be (9.0)            |
| 21 | <b>max</b> > -0.091726: be (76.0)     |
| 22 | mad > 0.030316                        |
| 23 | <b>max</b> <= 6.917615: o (4.0)       |

Fig. 1. The data mining process overview

### **Data Sources**

The spectra obtained with coudè spectrograph of Ondřejov Observatory 2m telescope were used as a training sample. Files were downloaded using SSA protocol. The SSA server is not publicly accessible outside of the local network of Ondřejov observatory. That is why the SSH tunneling of HTTP protocol was used. Two scripts for this process were created. First to construct the list of SSA compliant addresses, the second to analyse acquired response in VOTable format.

As testing sample the spectra from project SEGUE of SDSS were selected. This contains 178314 spectra in DR7. A simple SQL query was used to generate the list of URL links for individual FITS files.

## **Spectral Line Parameters**

#### The height of the $H\alpha$ line

The maximum value in the region of 50 Å around H $\alpha$  above the

 $Wavelenght[{
m \AA}]$ 

Fig. 2. The convolution with SDSS instrumental profile and re-binning applied on Ondřejov spectra

**Spectral Lines Characteristics** 

As parameters for data mining process characteristic values of  $H\alpha$  line were extracted from the spectra. Three parameters were finally selected. The height and the width of the  $H\alpha$  emission line and median absolute deviation as a characterisation of the noise level in the spectrum.



Fig. 4. The classifier decision tree obtained from data mining process

## Results

From the 10-fold cross-validation of training set we estimate the overall fruitfulness of classification to about 84%, which is quite good taking into account specific double-peak profile of Be stars and the temporal nature of their emission episodes. The classifier has identified 1110 Be stars candidates in SEGUE, however most of them are probably of different nature (e.g. AGNs, young stellar objects or reduction artifacts). Nevertheless, there are as well several highly probable Be stars like the one on Fig. 5.



linear fit was extracted from the spectrum.

#### The noise level of the spectrum

The noise in the spectrum contributes to the characteristics of the spectral lines. As an estimator of the noise level the median absolute deviation was used. It is defined as:

 $mad = median_i \left( \left| X_i - median_j(X_j) \right| \right)$ 

#### The width of the $\mathbf{H}\alpha$ line

The Gaussian function:

 $f(\lambda) = 1 + e^{-\frac{(\lambda - \lambda_0)^2}{S^2}}$ 

was fitted to the profile of H $\alpha$  spectral line, where  $\lambda_0$  is the center of line and *S* represents its width.

Fig. 3. The normalised spectrum of Be star 60 Cyg. The top figure depicts the continuum fit. The bottom figure shows the region (width of the green line) used for extraction. The position of the line corresponds to the maximum value in the region of 50 Å. The Gaussian fit is in red. Although the fit is almost perfect, this approach fails to get characteristic double peak of the emission line

Fig. 5. The example of candidate Be star found in SDSS SEGUE

#### Acknowledgements

The Astronomical Institute of the Academy of Sciences of the Czech Republic is supported by project AV0Z10030501. The support of Virtual Observatory Alliance project EURO-VO ICE is acknowledged as well. This work is based on spectra obtained with 2m telescope of Ondřejov observatory and on archival data of Sloan Digital Sky Survey (www.sdss.org)

Poster presented at ADASS 2011, 5-9th November 2011, Paris, France