Supervised Data Mining of Emission-Line Stars

Petr Škoda and Jaroslav Vážný

Astronomical Institute, Academy of Sciences, Ondřejov, Czech Republic

Abstract: Current data deluge in astronomy requires applying data mining techniques to extract new information about the physical nature of celestial objects. The possibility of cross-matching several surveys via Virtual Observatory protocols may play a key role in future discoveries. Data mining of large collections of spectra seems to be one of promising as well as challenging topics. We have focused on obtaining new candidates of Hα emission stars using supervised data mining method of Decision Trees on almost 200,000 spectra in SDSS SEGUE spectral survey.

Process Overview

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 Hα 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.

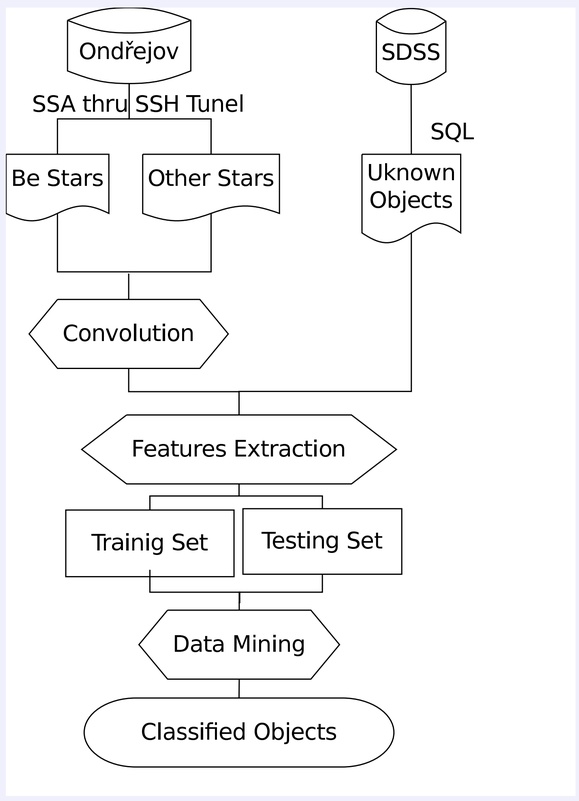

Fig. 1. The data mining process overview

Degradation of Spectral Resolution

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 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

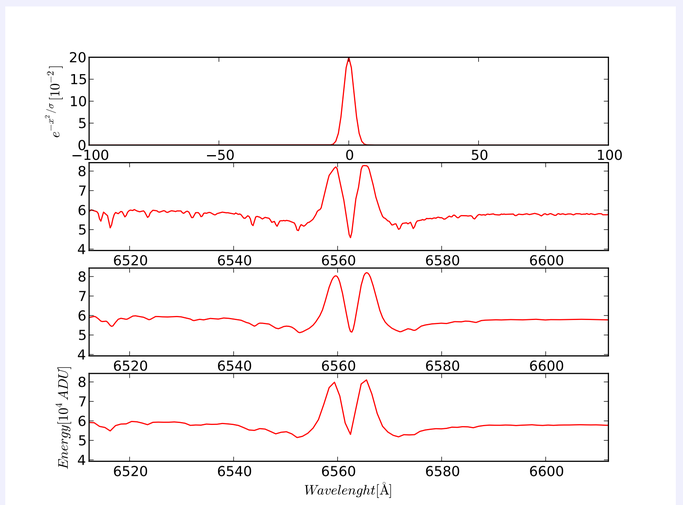

Fig. 2. The convolution with SDSS instrumental profile and re-binning applied on Ondřejov spectra

Data Mining

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 178314 items.

```
 === Summary ===
Correctly Classified Instances 145          83.815 %
Incorrectly Classified Instances 28         16.185 %
Kappa statistic                  0.6529
Mean absolute error              0.1849
Root mean squared error          0.3652
Relative absolute error         39.8819 %
Root relative squared error     75.8919 %
Total Number of Instances     173
```

```
 J48 pruned tree
------------------
max <= -0.18843
|  max <= -0.324763: o (46.0/5.0)
|  max > -0.324763
|  |  max <= -0.255475
|  |  |  mad <= 0.004133: o (2.0)
|  |  |  mad > 0.004133: be (13.0/1.0)
|  |  max > -0.255475
|  |  |  mad <= 0.009862: o (10.0)
|  |  |  mad > 0.009862
|  |  |  |  width <= 7.621593: o (3.0/1.0)
|  |  |  |  width > 7.621593: be (2.0)
max > -0.18843
|  mad <= 0.030316
|  |  max <= -0.091726
|  |  |  width <= 5.286489
|  |  |  |  max <= -0.170022: be (2.0)
|  |  |  |  max > -0.170022: o (3.0)
|  |  |  width > 5.286489: be (9.0)
|  |  max > -0.091726: be (76.0)
|  mad > 0.030316
|  |  max <= 6.917615: o (4.0)
|  |  max > 6.917615: be (3.0)
```

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

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 Hα line

The maximum value in the region of 50 Å around Hα above the 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:

$$\text{mad} = \text{median}_i \left(\left| X_i - \text{median}_j(X_j) \right| \right)$$

The width of the Hα line

The Gaussian function:

$$f(x) = 1 + e^{-\frac{(x-x_0)^2}{S^2}}$$

was fitted to the profile of Hα spectral line.

Spectral Lines Characteristics

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

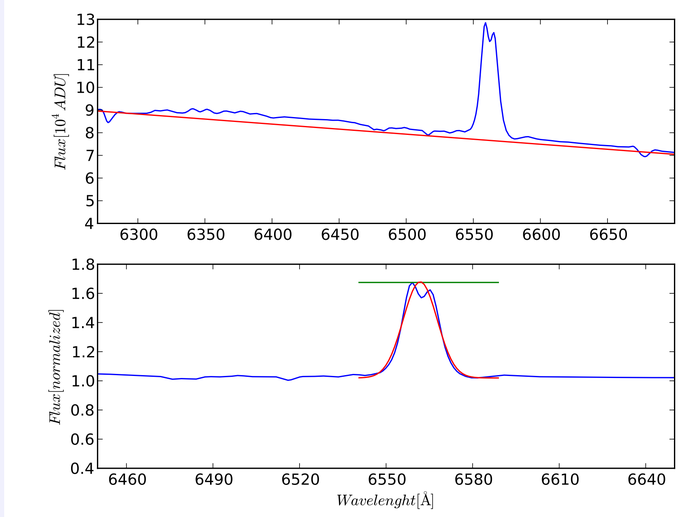

Fig. 3. The normalised spectrum of Be star 60 Cyg. The top figure depicts the continuum fit. The bottom figure shows the region (width of the green line) used for extraction. The position of the line corresponds to the maximum value in the region of 50 Å. The Gaussian fit is in red. Although the fit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

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.

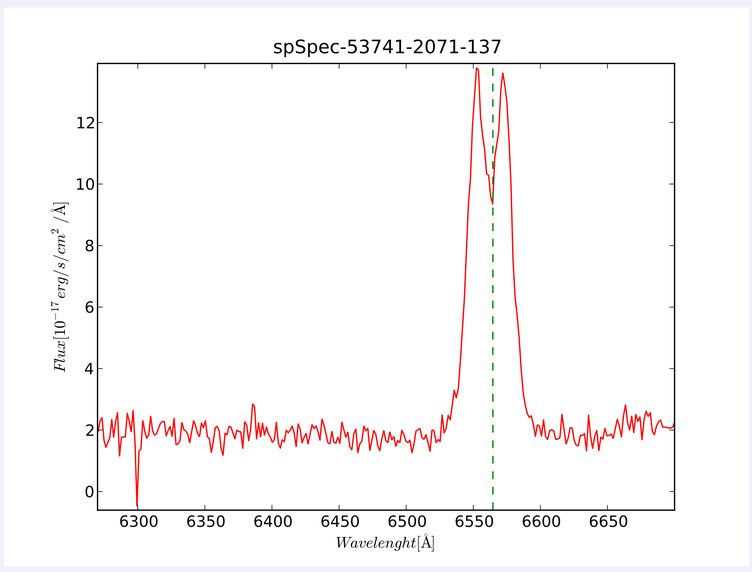

Fig. 5. The example of candidate Be star found in SDSS SEGUE survey using the decision tree shown above.

Poster presented at Astroinformatics 2011, 25-29th September 2011, Sorento, Italy