

Classification of Spectra of Emission Line Stars Using Machine Learning Techniques

Pavla Bromová¹ Petr Škoda² Jaroslav Vážný²

¹Faculty of Information Technology, Brno University of Technology, Božetěchova 1/2, 612 66 Brno, Czech Republic

²Astronomical Institute of the Academy of Sciences of the Czech Republic, Fričova 298, 251 65 Ondřejov, Czech Republic

Abstract: Advances in the technology of astronomical spectra acquisition have resulted in an enormous amount of data available in world-wide telescope archives. It is no longer feasible to analyze them using classical approaches, so a new astronomical discipline, astroinformatics, has emerged. We describe the initial experiments in the investigation of spectral line profiles of emission line stars using machine learning with attempt to automatically identify Be and B[e] stars spectra in large archives and classify their types in an automatic manner. Due to the size of spectra collections, the dimension reduction techniques based on wavelet transformation are studied as well. The result clearly justifies that machine learning is able to distinguish different shapes of line profiles even after drastic dimension reduction.

Keywords: Be star, stellar spectrum, feature extraction, dimension reduction, discrete wavelet transform, classification, support vector machines (SVM), clustering.

1 Introduction

The research in almost all natural sciences is facing the “data avalanche” represented by exponential growth of information produced by big digital detectors and large-scale multi-dimensional computer simulations. The effective retrieval of scientific knowledge from petabyte-scale databases requires the qualitatively new kind of scientific discipline called e-science, allowing the global collaboration of virtual communities sharing the enormous resources and power of supercomputing grids^[1, 2].

The emerging new kind of research methodology of contemporary astronomy — astroinformatics — is based on systematic application of modern informatics and advanced statistics on huge astronomical data sets. Such an approach, involving machine learning, classification, clustering and data mining, yields new discoveries and better understanding of nature of astronomical objects. It is sometimes presented as a new way of doing astronomy^[3], representing the example of working e-Science in astronomy. The application of methods of working e-science in some common astronomical tasks may lead to new interesting results and a different view of the investigated problem. We present a project that focuses on using machine learning of large spectral data archives in the investigation of emission line profiles of Be and B[e] stars in order to find new such objects.

1.1 Emission line stars

There are a lot of stellar objects that may show some important spectral lines in emission. The physical parameters may differ considerably, however, there seems to be the common origin of their emission — the gaseous circumstellar envelope in the shape of sphere or rotating disk. Among the most common types, they are Be stars, B[e] stars, pre-main-sequence stars (e.g. T Tau and Herbig stars), stars with strong stellar winds (like P Cyg or eta Carinae), Wolf-Rayet stars, Novae and Symbiotic stars.

1.2 Be and B[e] stars

The classical Be stars^[4] are non-supergiant B type stars whose spectra have or had at some time, one or more emission lines in the Balmer series. In particular, the H_α emission is the dominant feature in the spectra of these objects. The emission lines are commonly understood to originate in the flattened circumstellar disk, probably to be of secretion origin (i.e., created from the material of central star), however the exact mechanism is still unsolved. The Be stars are not rare in the universe: They represent nearly one fifth of all B stars and almost one third of B1 stars^[5].

The emission and absorption profiles of Be stars vary with the time scales, from years to fraction of a day, and they seem to switch between emission state and the state of pure absorption spectrum indistinguishable from normal B stars. This variability may be caused by the evolution and disappearing of disk^[4].

Similar strong emission features in H_α show the B[e] stars^[6]. However, they present as well forbidden lines of low excitation elements (e.g., iron, carbon, oxygen, nitrogen) and infrared excess (pointing to the presence of dusty envelope). The B[e] stars are very rare, mostly unclassified, so the new yet unknown members of this interesting group are highly desirable.

Regular Paper
Special Issue on Recent Advances on Complex Systems Control, Modelling and Prediction

Manuscript received August 27, 2013; revised October 11, 2013
This work was supported by Czech Science Foundation (No. GACR 13-08195S), the project Central Register of Research Intentions CEZ MSM0021630528 Security-oriented Research in Information Technology, the specific research (No. FIT-S-11-2), the project RVO: 67985815, the Technological agency of the Czech Republic (TACR) project V3C (No. TE01020415), and Grant Agency of the Czech Republic - GACR P103/13/08195S.

1.3 Be stars spectra archives

The spectra of Be and B[e] stars are dispersed worldwide in many archives of individual telescopes and space missions, and most of them are still not yet made available for public (namely archives of smaller observatories). The largest publicly available collection of about ninety thousand spectra of more than 900 different stars represents the BeSS database¹. The disadvantage of such archives for machine learning purposes is its large inhomogeneity as it contains randomly uploaded spectra from different telescopes, various kinds of spectrographs (both single order and Echelle) with different spectral resolutions and various processings applied by both amateurs and professional observers.

That is why we use spectra from Ondřejov 2 m Perek telescope of the Astronomical Institute of the Academy of Sciences of the Czech Republic obtained by its 700 mm camera in coudé spectrograph, which uses the same optical setup for more than 15 years. It contains about ten thousand spectra of different stars in the same configuration processed and calibrated according to the same recipe. Most of them are spectra of about 300 Be and B[e] stars. So it represents the largest homogeneous sample of spectra.

1.4 Motivation

By defining a typical property of spectra (shape of continuum or presence of a type-specific spectral line), we will be able to classify the observed samples. The appropriate choice of classification criterion will give us a powerful tool for searching of new candidates of interesting kind.

As the Be stars show a number of different shapes of emission lines like double-peaked profiles with or without narrow absorption (called shell line) or single peak profiles with various wing deformations (e.g., “wine-bottle”^[7]), it is very difficult to construct a simple criterion to identify the Be lines in an automatic manner as required by the amount of spectra considered for processing. However, even a simple criterion of combination of three attributes (width, height of Gaussian fit through spectral line and the medium absolute deviation of noise) was sufficient to identify interesting emission line objects among nearly two hundred thousand of SDSS SEGUE spectra^[8].

To distinguish different types of emission line profiles (which is impossible using only Gaussian fit), we propose a completely new methodology, which seems to be not yet used (according to our knowledge) in astronomy, although it has been successfully applied in recent five years to many similar problems like detection of particular EEG activity. It is based on supervised machine learning of the set of positively identified objects. This will give some kind of classifier rules, which are then applied to a larger investigated sample set of unclassified objects. In fact, it is a kind of transformation of data from the basis of observed variables to another basis in a different parameter space, hoping that in this new space, the different classes will be easily distinguishable. As the number of independent input parameters has to be kept low, we cannot directly use all points of each spectrum but we have to find a concise description of the spectral features, which conserves most of

the original information content.

One of the quite common approaches is to make the principal components analysis (PCA) to get a small basis of input vectors for machine training. However, the most promising method is the wavelet decomposition (or multi-resolution analysis) using the pre-filtered set of largest coefficients or power spectrum of the wavelet transformation of input stellar spectra in the role of feature vectors. This method has been already successfully applied to many problems related to recognition of given patterns in input signal, as is the identification of epilepsy in EEG data^[9]. The wavelet transformation is often used for general knowledge mining^[10] or a number of other applications. A nice review was given by Li et al.^[11]. In astronomy, the wavelet transformation was used recently for estimating stellar physical parameters from Gaia radial velocity spectrometer (RVS) simulated spectra with low signal noise ratio (SNR)^[12]. However, they have classified stellar spectra of all ordinary types of stars, while we need to concentrate on different shapes of several emission lines which require extraction of feature vectors first. In the following chapters, we describe several different experiments with extraction of main features in an attempt to identify the best method as well as verification of the results using both unsupervised (clustering) and supervised (classification) learning of both extracted feature vectors and original data points.

2 Experiment 1: Comparison of different wavelet types on simulated spectra

In this experiment, we used the discrete wavelet transform (DWT) implemented in Matlab. One of the parameters of wavelet transform is the type of wavelet. The goal of this experiment was to compare the effect of using different types of wavelets on the results of clustering. An extensive literature exists on wavelets and their applications^[13–17].

We tried to find the wavelet best describing the character of our data, based on its similarity with the shape of emission lines. We were choosing from the set of wavelets available for DWT in Matlab, i.e., daubechies, symlets, coiflets, biorthogonal, and reverse biorthogonal wavelets family. We chose two types of different orders from each family:

- 1) Daubechies (db): orders 1, 4.
- 2) Symlets (sym): orders 6, 8.
- 3) Coiflets (coif): orders 2, 3.
- 4) Biorthogonal (bior): orders 2.6, 6.8.
- 5) Reverse biorthogonal (rbio): orders 2.6, 5.5.

2.1 Data

The experiment was performed on simulated spectra generated by computer. A collection of 1 000 spectra was created to cover as many emission line shapes as possible. Each spectrum was created using a combination of 3 Gaussian functions with parameters generated randomly within appropriately defined ranges, and complemented by a random noise. The length of a spectrum is 128 points, which approximately corresponds to the length of a spectrum seg-

¹<http://basebe.obspm.fr>

ment used for emission lines analysis. Each spectrum was then convolved with a Gaussian function, which simulates an appropriate resolution of the spectrograph.

2.2 Feature extraction

The DWT was performed in Matlab using the embedded functions. The feature vector is composed of the wavelet power spectrum computed from the wavelet coefficients.

Wavelet power spectrum. The power spectrum measures the power of the transformed signal at each scale of the employed wavelet transform. The bias of this power spectrum was further rectified^[18] by division by corresponding scale. The normalized power spectrum P_j for the scale j can be described as

$$P_j = 2^{-j} \sum_n |W_{j,n}|^2 \tag{1}$$

where $W_{j,n}$ is the wavelet coefficient of the j -th scale of the n -th spectrum.

2.3 Clustering

Clustering was performed using k -means algorithm to result in 3–6 clusters. The silhouette method^[19] was used for the evaluation. Clustering was performed with 50 iterations and the average silhouette values were presented as the results.

2.4 Results

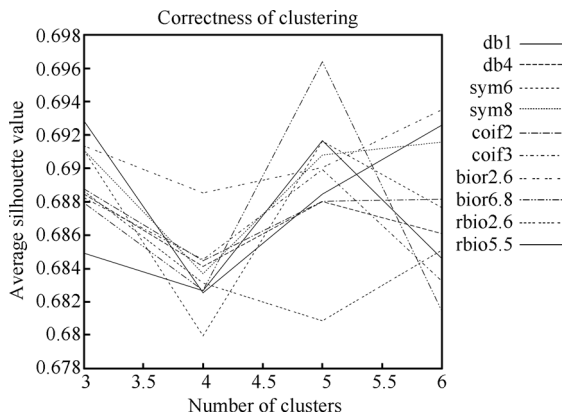


Fig. 1 Correctness of clustering for 3, 4, 5, and 6 clusters using different types of wavelets

In Fig. 1, we can see that there are minimal differences in the correctness of clustering using different types of wavelets (hundredths of unit), which suggests the type of wavelet has no big effect on the clustering results.

3 Experiment 2: Comparison of feature vectors using clustering

In this experiment, we presented a feature extraction method based on the wavelet transform and its power spectrum (WPS), and an additional value indicating the orientation of the spectral line. Both the discrete (DWT) and continuous (CWT) wavelet transforms are used. Different feature vectors were created and compared in terms of clustering of Be stars spectra from the Ondřejov archive. The

clustering was performed using the k -means algorithm.

3.1 Data selection

The data set consists of 656 samples of stellar spectra of Be stars and also normal stars divided manually into 4 classes (66, 150, 164, and 276 samples) based on the shape of the emission line. From the input data, a segment with the H_α spectral line is analyzed. The segment length of 256 pixels is chosen with regard to the width of the emission line and to the dyadic decomposition used in DWT. Examples of selected data samples typical for each of the 4 classes are illustrated in Fig. 2.

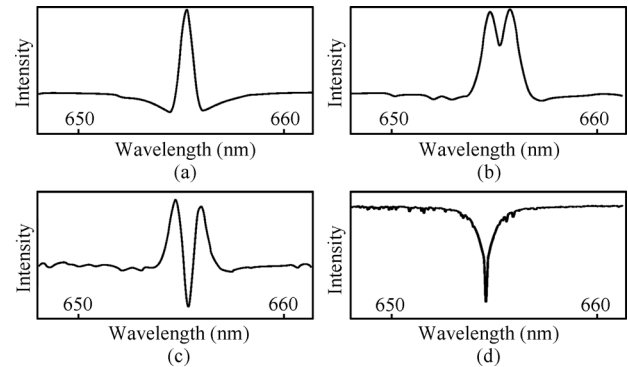


Fig. 2 Examples of selected data samples typical for each of the 4 classes: (a), (b) and (c) are spectra of Be stars; (d) is a normal star. In (a) there is a pure emission on H_α spectral line; (b) contains a small absorption part (less than $\frac{1}{3}$ of the height); (c) contains a larger absorption part (more than $\frac{1}{3}$ of the height). The spectrum of a normal star (d) consists of a pure absorption

3.2 Feature extraction

The feature vector is composed of two parts:

- 1) Set of features computed from wavelet coefficients.
- 2) Value indicating the orientation of the spectral line (this information is lost in the wavelet power spectrum).

In this experiment, the wavelet transform was performed in Matlab using the embedded functions, with the wavelet “symlet 4”.

Orientation of spectral line. The information about the orientation of a spectral line is lost in the wavelet power spectrum coefficients. Two data samples with the same shape but opposite orientations of the spectral line would yield an equal wavelet power spectrum. Therefore, this information must be added into the feature vector. We want to distinguish whether a spectral line is oriented up (emission line) or down (absorption line), so we use one positive value and one negative value. The question is which absolute value to choose. In this experiment, we tried three values: 1, 0.1, and the amplitude of a spectral line, measured from the continuum of value 1.

N largest coefficients. As we did not have any reference methods of feature extraction from Be stars for comparison, we compared our results with a common method of feature extraction from time series using wavelets, in which N largest coefficients of wavelet transform are kept and the rest of the coefficients are set to zero^[20]. In experiments, $N = 10$ was used. In this feature extraction technique, the orientation of a spectral line is not added to the feature

vector, as the wavelet coefficients do contain the information about the orientation and the amplitude of the spectral line.

Feature vectors. Different kinds of feature vectors were created from the resulting coefficients of the wavelet transform and used for comparison:

1) Spectrum: original spectrum values, normalized to range [0,1]. (In this case, the DWT coefficients are not used.)

2) Approximation: DWT approximation coefficients, normalized to range [0,1].

3) Approximation+detail: DWT approximation and detail coefficients of the last level, normalized to range [0, 1].

4) 10 largest coefs: 10 largest absolute values of coefficients, normalized to range [-1, 1].

5) 20 largest coefs: 20 largest absolute values of coefficients, normalized to range [-1, 1].

6) Discrete wavelet power spectrum (DWPS) + orientation 1: One part of the feature vector is the wavelet power spectrum of DWT, normalized such that its total energy equals 1. The second part of the feature vector is a value indicating the orientation of the spectral line—lines oriented up have the value of 1, lines oriented down have the value of -1.

7) DWPS+orientation 0.1: the same as the previous one except the absolute value of orientation 0.1.

8) DWPS+amplitude: One part of the feature vector is normalized wavelet power spectrum as in the previous case. The second part is the amplitude of the spectral line measured from the continuum of value 1.

9) Continuous wavelet power spectrum (CWPS) 16+orientation 1: Wavelet power spectrum (normalized) of CWT performed with 16 scales. The orientation is the same as in the previous cases with DWPS.

10) CWPS 8+orientation 1: Wavelet power spectrum (normalized) of CWT performed with 8 scales. The orientation is the same as in the previous case.

3.3 Clustering

The k -means algorithm in Matlab was used for clustering. Squared Euclidean distance was used as a distance measure. Clustering was repeated 30 times, each iteration with a new set of initial cluster centroid positions. k -means returns the solution with the lowest within-cluster sums of point-to-centroid distances.

3.4 Evaluation

We proposed an evaluation method utilizing our knowledge of ideal classification of spectra based on a manual categorizing.

The principle is to simply count the number of correctly classified samples. We have 4 target classes and 4 output classes, but the problem is we do not know which output class corresponds to which target class. So, at first we need to map the output classes to the target classes, i.e., to assign each output class a target class. This is achieved by creating the correspondence matrix, which is a square matrix of the size of the number of classes, and where the element on position (i, j) corresponds to the number of samples with an output class i and a target class j . In the case of a per-

fect clustering, all values besides the main diagonal would be equal to zero.

Now we find the mapping by searching for the maximum value in the matrix. The row and the column of the maximum element will constitute the corresponding pair of output and target class. We set this row and column to zero and again find the maximum element. By repeating this process we find all corresponding pairs of classes. The maximum values correspond to correctly classified samples. So now we simply count the number of correctly classified samples by summing all maximum values we used for mapping the classes. By dividing by the total number of samples we get the percentual match of clustering which is used as a final evaluation.

3.5 Results

Fig. 3 shows the percentual match of the clustering for different kinds of feature vectors. The numbers of feature vectors in the figure correspond to the numbers in the numbered list in Section 3.2.

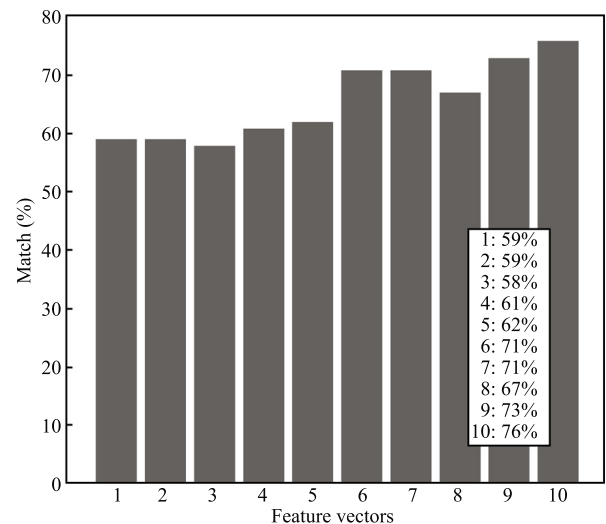


Fig. 3 The match of the clustering using different feature vectors

The best results are given by the last feature vector consisting of the continuous wavelet power spectrum calculated from 8 scales of CWT coefficients, and the value representing the orientation of the H_α line with absolute value of 1. The match is 14% higher than the best result of a feature vector without WPS. Also the results of all other feature vectors containing WPS are better than the feature vectors without WPS.

4 Experiment 3: Comparison of feature vectors using classification

In this experiment, we proposed several feature extraction methods based on the discrete wavelet transform (DWT). The data set was the same as in the previous experiment, but in addition the known target classes of spectra (manually assigned) were used for training. A small segment containing the H_α line was selected for feature extraction. Classification was performed using the support

vector machines (SVM). The results were given by the accuracy of classification.

4.1 Feature extraction

In this experiment, the wavelet transform was performed using the cross-platform discrete wavelet transform library^[21]. The selected data samples were decomposed into J scales using the discrete wavelet transform with CDF $9/7^{[22]}$ wavelet as in (2). This wavelet is employed for lossy compression in JPEG 2000 and Dirac compression standards. Responses of this wavelet can be computed by a convolution with two FIR filters, one with 7 coefficients and the other with 9 coefficients.

$$W_{j,n} = \langle x, \psi_{j,n} \rangle \tag{2}$$

where x is data sample, and ψ is the wavelet function.

On each obtained sub-band, the following descriptor was calculated to form the resulting feature vector as (3). The individual methods are further explained in detail.

$$v = \{v_j\}_{1 \leq j < J}. \tag{3}$$

Wavelet power spectrum. It is described in Section 2.2.

Euclidean norm. The Euclidean or ℓ^2 norm is the intuitive notion of length of a vector. The norm for the specific sub-band j can be calculated as $\|W_j\|_2$ by

$$\|W_j\|_2 = \left(\sum_n |W_{j,n}|^2 \right)^{\frac{1}{2}}. \tag{4}$$

Maximum norm. Similarly, the maximum or infinity norm can be defined as the maximal value of DWT magnitudes:

$$\|W_j\|_\infty = \max_n |W_{j,n}|. \tag{5}$$

Arithmetic mean. The mean (6) is the sum of wavelet coefficients W_j at the specific scale j divided by the number of coefficients there. In this paper, the mean is defined as the expected value with respect to the method below.

$$\mu_j = E[W_j]. \tag{6}$$

Standard deviation. The standard deviation (7) is the square root of the variance of the specific wavelet sub-band at the scale of j . It indicates how much variation exists with respect to the arithmetic mean.

$$\sigma_j = (E[(W_j - \mu_j)^2])^{\frac{1}{2}}. \tag{7}$$

4.2 Classification

Classification of resulting feature vectors was performed using the support vector machines (SVM)^[23]. The library LIBSVM^[24] was employed. The radial basis function (RBF) was used as the kernel function.

There are two parameters for an RBF kernel: C and γ . It is not known beforehand which C and γ are the best for a given problem, therefore some kind of model selection (parameter search) must be done. A strategy known as “grid-search” was used to find parameters C and γ for each feature

extraction method. In grid-search, various pairs of C and γ values are tried, each combination of parameter choices is checked using cross-validation, and the parameters with the best cross-validation accuracy are picked up. We tried exponentially growing sequences of $C = 2^{-5}, 2^{-3}, \dots, 2^{15}$ and $\gamma = 2^{-15}, 2^{-13}, \dots, 2^3$. Finally, values $C = 32$ and $\gamma = 2$ had the best accuracy. For cross-validation, 5 folds were used.

Before classification, scaling of feature vectors (before adding the orientation) was performed in the interval $[0, 1]$.

4.3 Results

The results were obtained for different feature extraction techniques in terms of accuracy of classification. For comparison, a feature vector consisting of the original values of the stellar spectrum without the feature extraction was also used for classification. The results are given in Fig. 4.

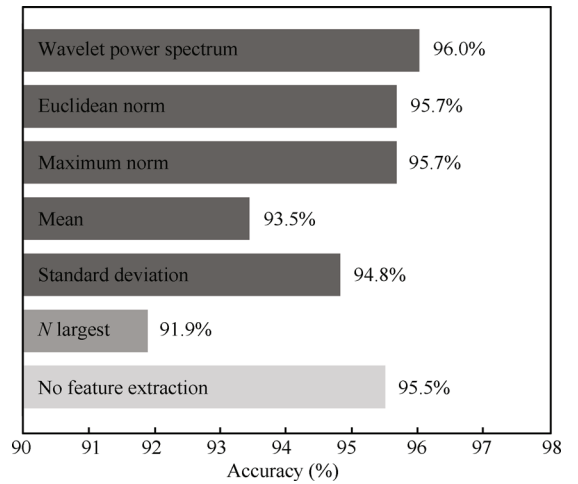


Fig. 4 Accuracies of classification for different feature extraction methods

The results of all the feature extraction methods are comparable with the satisfying accuracy which approaches the accuracy of a feature vector consisting of the original values of the stellar spectrum without feature extraction. Moreover, the results are significantly better than that of the common method of feature extraction from time series using wavelets — keeping N largest coefficients of the wavelet transform, which has been chosen for comparison.

The best results are given by the feature extraction using the wavelet power spectrum, where the accuracy is even higher than that of the original data without the feature extraction.

5 Experiment 4: Classification without feature extraction

The aim of this experiment is to test if it is possible to train machine learning algorithm (SVM in this case) to discriminate between manually selected groups of Be stars spectra.

5.1 Data selection

The training set consists of 2 164 spectra from Ondřejov archive, which were divided into 4 distinct categories based on the region around Balmer H_α line (which is the interesting region for that type of stars). The spectra were normalized and trimmed to 100 Å around H_α . So we got samples with about eight hundred points. The numbers of spectra in individual categories are shown in Table 1.

Table 1 Numbers of spectrain individual categories

Category	1	2	3	4
Count	408	289	1366	129

For better understanding of the categories characteristics, there are a plot of 25 random samples in Fig. 5 and characteristics spectrum of individual categories created as a sum of all spectra in corresponding category in Fig. 3.

Principal component analysis (PCA) was also performed to visually check if there is a separation (and therefore a chance) to discriminate between individual classes, see Fig. 7. As seen, the cluster on the left (around 0,0) seems to contain the mixture of several classes (see the detail in Fig. 8). It is the proof that simple linear feature extraction like PCA cannot distinguish between some visible shapes and it confirms that a more appropriate method has to take into account both global large-scale shape and small details of certain features. This is exactly what the multi-resolution scalable method like DWT is doing.

5.2 Classification

Classification was performed using the support vector machines (SVM)^[23] with the library scikit-learn^[25] and IPython interactive shell. The radial basis function (RBF) was used as the kernel function.

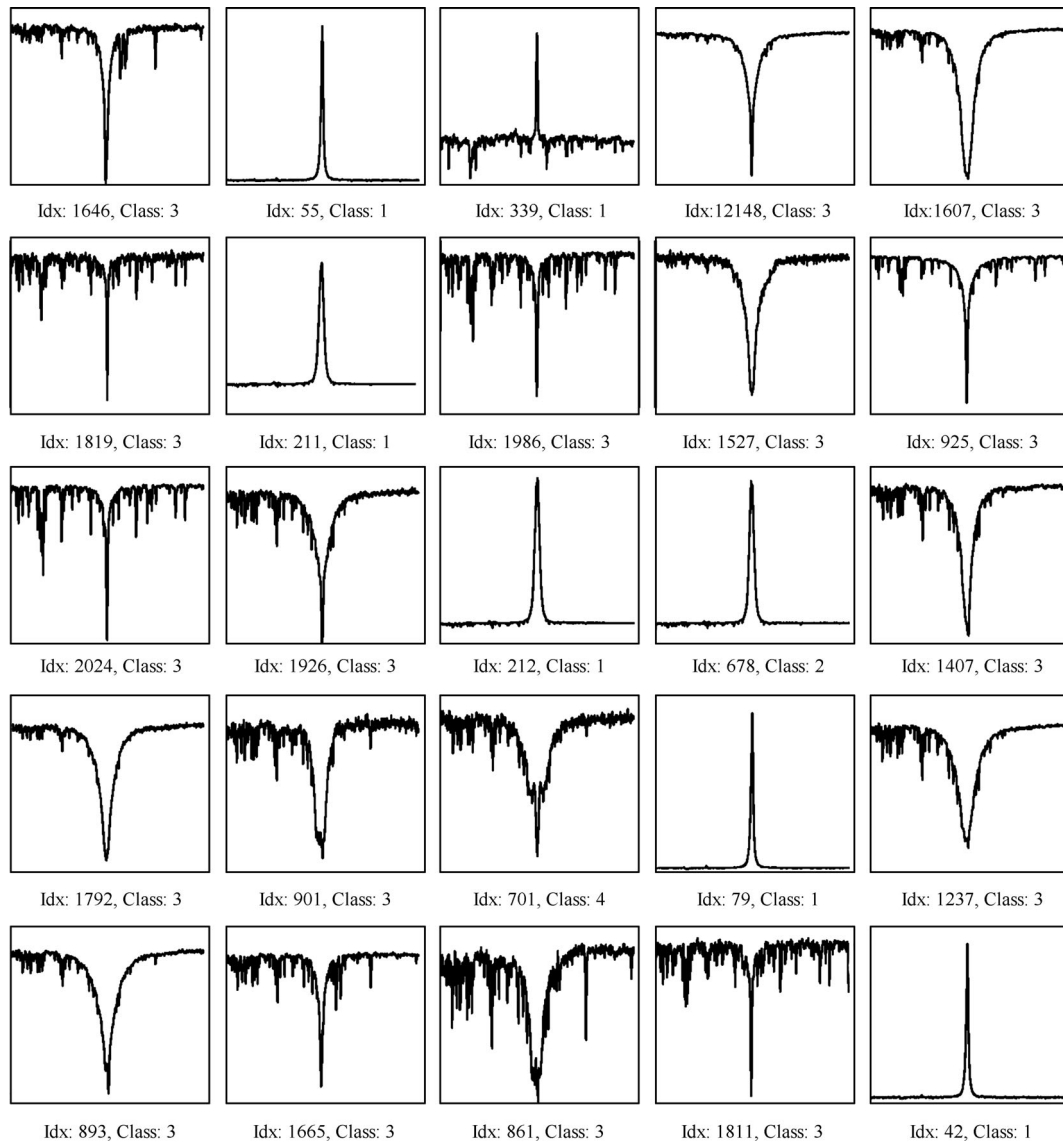


Fig. 5 Random samples from all categories

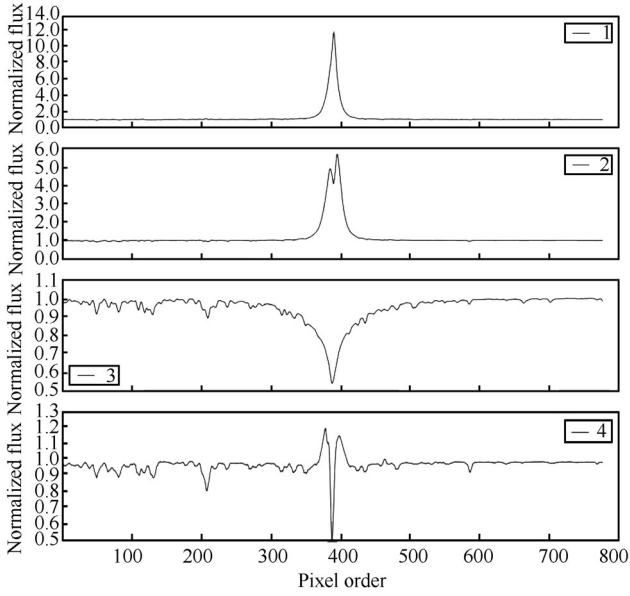


Fig. 6 Characteristic spectrum of individual categories created as a sum of all spectra in corresponding categories

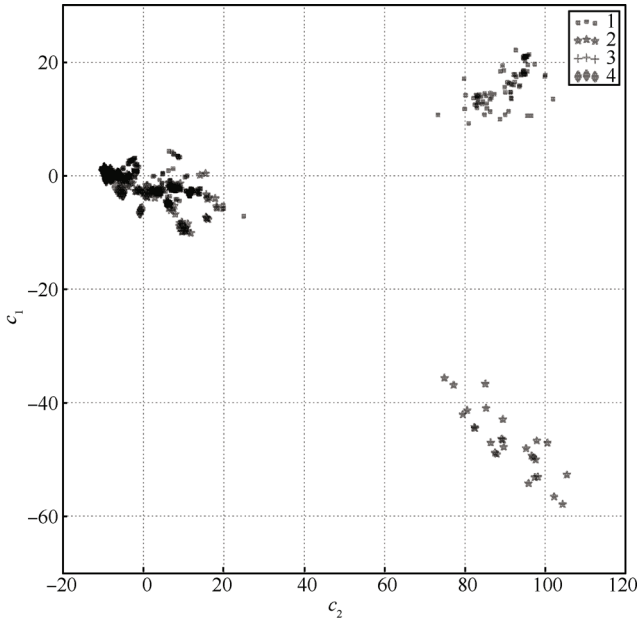


Fig. 7 PCA separation of individual classes. The symbols in the legend represent different categories of the line shapes from Fig. 6

To find optimum values of parameters C and γ , the grid-search was performed with 10-fold cross-validation with samples size = 0.1. The results are in Table 2. Based on this result, values $C = 100.0$ and $\gamma = 0.01$ were used in the following experiments.

5.3 Results

The mean score was $0.988 (\pm 0.002)$. There is a detailed report (now based on test sample = 0.25) in Table 3, where f1-score is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

Learning curve. It is an important tool which helps us understand the behaviour of the selected model. As you can see in Fig. 9 from about 1000 samples there is not big improvement and there is probably not necessary to have more than 1300 samples. Of course, this is valid only for this model and data.

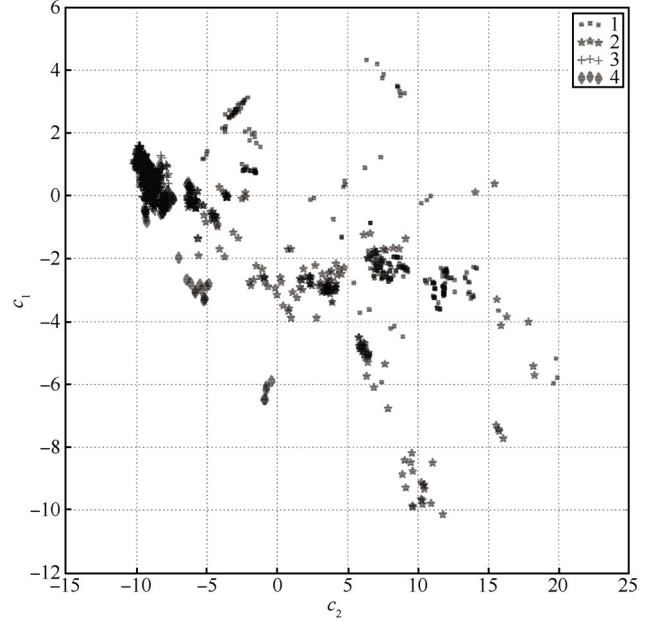


Fig. 8 Detailed view of PCA cluster around (0,0)

Table 2 Results of the grid-search

Parameters	Score
$C = 100.0, \gamma = 0.01$	$0.985 (\pm 0.003)^*$
$C = 10.0, \gamma = 0.1$	$0.978 (\pm 0.003)^*$
$C = 100.0, \gamma = 0.1$	$0.977 (\pm 0.004)^*$
$C = 10.0, \gamma = 0.01$	$0.973 (\pm 0.002)$
$C = 1.0, \gamma = 0.1$	$0.970 (\pm 0.003)$
$C = 100.0, \gamma = 0.001$	$0.969 (\pm 0.002)$
$C = 1.0, \gamma = 1.0$	$0.966 (\pm 0.003)$
$C = 10.0, \gamma = 1.0$	$0.965 (\pm 0.004)$
$C = 100.0, \gamma = 1.0$	$0.965 (\pm 0.004)$
$C = 1.0, \gamma = 0.01$	$0.958 (\pm 0.002)$
$C = 10.0, \gamma = 0.001$	$0.956 (\pm 0.003)$
$C = 100.0, \gamma = 0.0001$	$0.953 (\pm 0.003)$
$C = 0.1, \gamma = 0.1$	$0.929 (\pm 0.005)$
$C = 10.0, \gamma = 0.0001$	$0.915 (\pm 0.004)$
$C = 1.0, \gamma = 0.001$	$0.914 (\pm 0.003)$
$C = 0.1, \gamma = 0.01$	$0.908 (\pm 0.003)$
$C = 0.1, \gamma = 1.0$	$0.885 (\pm 0.004)$
$C = 1.0, \gamma = 0.0001$	$0.811 (\pm 0.003)$
$C = 0.1, \gamma = 0.001$	$0.811 (\pm 0.003)$
$C = 0.1, \gamma = 0.0001$	$0.785 (\pm 0.003)$

*denotes the best values

Miss-classification. There were only seven miss-classified cases (based on test_{size} = 0.25). Fig. 10 shows that spectra. It is seen that the boundary between classes is sometimes very thin — e.g., the small distortion of peak is considered as a double peak profile or deeper absorption masks the tiny emission inside it. But these are cases where

the human would have same troubles to decide the about right class. In general, the performance of machine learning is very good, and prove that it is a viable approach to find desired objects in current petabyte-scaled astronomical collections.

Table 3 Results of classification

Class	Precision	Recall	f1-score	Support
1	0.98	0.96	0.97	100
2	0.95	0.97	0.96	72
3	1.00	1.00	1.00	341
4	0.96	0.96	0.96	28
Average/total	0.99	0.99	0.99	541

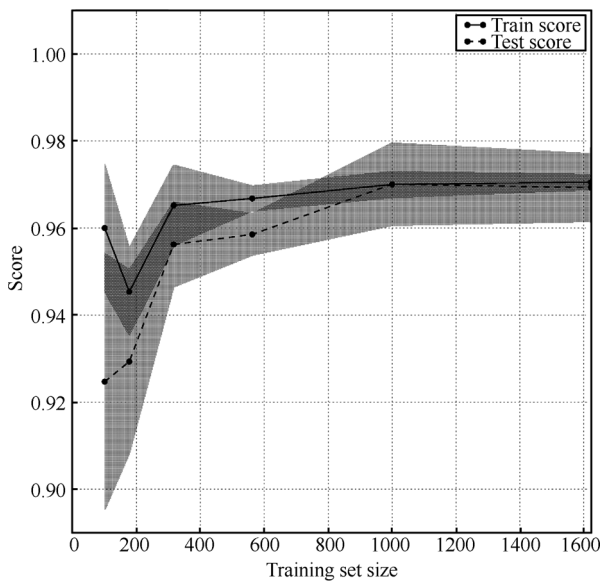
Main train and test scores ± 2 standard errors

Fig. 9 Learning curve

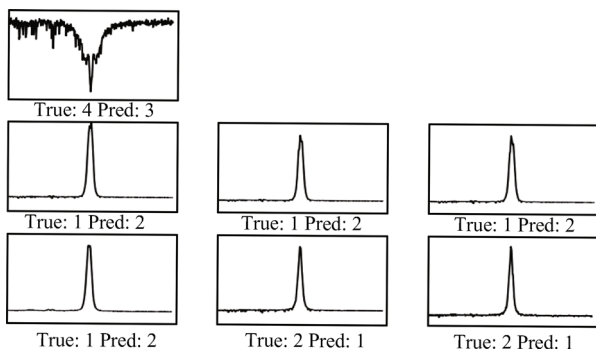


Fig. 10 The miss-classified samples

6 Conclusion

This paper describes the initial experiments in the field of investigation of spectral line profiles of emission line stars using machine learning with an attempt to automatically identify Be and B[e] stars spectra in large archives and classify their types in an automatic manner. Due to the huge size of spectra collections, dimension reduction techniques based on wavelet transformation are studied as well.

The results clearly justify that the machine learning is

able to distinguish different shapes of line profiles even after drastic dimension reduction.

In the future work, we will compare different classification methods and use the results for comparison with the clustering results. Based on this, we will try to find the best clustering model and its parameters, which will then be possible to be used for clustering all spectra in Ondřejov archive, and possibly to find new interesting candidates.

Acknowledgement

We would like to acknowledge the help of Omar Laurino from Harvard-Smithsonian Center for Astrophysics and Massimo Brescia and Giuseppe Longo from the DAME team of University of Naples for the original ideas this research is based on, as well as patience of Jaroslav Zendulka, supervisor of P.B.

References

- [1] Y. Zhao, I. Raicu, I. Foster. Scientific workflow systems for 21st century, new bottle or new wine? In *Proceedings of the 2008 IEEE Congress on Services*, IEEE, Washington, USA, pp. 467–471, 2008.
- [2] Y. X. Zhang, H. W. Zheng, Y. H. Zhao. Knowledge discovery in astronomical data. In *Proceedings of SPIE Conference*, vol. 7019, 2008.
- [3] K. D. Borne. Astroinformatics: a 21st century approach to astronomy. In *Proceedings of ASTRO 2010: The Astronomy and Astrophysics Decadal Survey*, 2009.
- [4] J. M. Porter, T. Rivinius. Classical be stars. *The Publications of the Astronomical Society of the Pacific*, vol. 115, no. 812, pp. 1153–1170, 2003.
- [5] J. Zorec, D. Briot. Critical study of the frequency of Be stars taking into account their outstanding characteristics. *Astronomy and Astrophysics*, vol. 318, pp. 443–460, 1997.
- [6] F. J. Zickgraf. Kinematical structure of the circumstellar environments of galactic B[e]-type stars. *Astronomy and Astrophysics*, vol. 408, no. 1, pp. 257–285, 2003.
- [7] R. W. Hanuschik, W. Hummel, E. Sutorius, O. Dietle, G. Thimm. Atlas of high-resolution emission and shell lines in Be stars. Line profiles and short-term variability. *Astronomy and Astrophysics Supplement Series*, vol. 116, pp. 309–358, 1996.
- [8] P. Škoda, J. Vázný. Searching of new emission-line stars using the astroinformatics approach. *Astronomical Data Analysis Software and Systems XXI*, *Astronomical Society of the Pacific Conference Series*, vol. 461, pp. 573, 2012.
- [9] P. Jahankhani, K. Revett, V. Kodogiannis. Data mining an EEG dataset with an emphasis on dimensionality reduction. In *Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining*, IEEE, Honolulu, HI, USA, pp. 405–412, 2007.
- [10] M. Murugappan, R. Nagarajan, S. Yaacob. Combining spatial filtering and wavelet transform for classifying human emotions using EEG signals. *Journal of Medical and Biological Engineering*, vol. 31, no. 1, pp. 45–51, 2011.
- [11] T. Li, Q. Li, S. H. Zhu, M. Ogihara. A survey on wavelet applications in data mining. *SIGKDD Explorations Newsletter*, vol. 4, pp. 49–68, 2002.

- [12] M. Manteiga, D. Ordóñez, C. Dafonte, B. Arcay. ANNs and wavelets: A strategy for Gaia RVS Low S/N Stellar Spectra Parameterization. *Publications of the Astronomical Society of the Pacific*, vol. 122, no. 891, pp. 608–617, 2010.
- [13] S. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed., San Diego: Academic Press, 2008.
- [14] I. Daubechies. *Ten Lectures on Wavelets, CBMS-NSF Regional Conference Series in Applied Mathematics*, Philadelphia, PA: Society for Industrial and Applied Mathematics, 1994.
- [15] G. Kaiser. *A Friendly Guide to Wavelets*, Boston: Birkhäuser, 1994.
- [16] Y. Meyer, D. H. Salinger. *Wavelets and Operators, Number sv. 1 in Cambridge Studies in Advanced Mathematics*, Cambridge: Cambridge University Press, 1995.
- [17] G. Strang, T. Nguyen. *Wavelets and Filter Banks*, Cambridge: Wellesley-Cambridge Press, 1996.
- [18] Y. G. Liu, X. S. Liang, R. H. Weisberg. Rectification of the bias in the wavelet power spectrum. *Journal of Atmospheric and Oceanic Technology*, vol. 24, no. 12, pp. 2093–2102, 2007.
- [19] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.
- [20] T. Li, S. Ma, M. Ogihara. Wavelet methods in data mining. *Data Mining and Knowledge Discovery Handbook*, O. Maimon, L. Rokach, Eds., New York: Springer, pp. 553–571, 2010.
- [21] D. Bařina, P. Zemčık. *A cross-platform discrete wavelet transform library*. Authorised software, Brno University of Technology, 2010-2013, [Online], available: http://www.fit.vutbr.cz/research/view_product.php?id=211.
- [22] A. Cohen, I. Daubechies, J. C. Feauveau. Biorthogonal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, vol. 45, no. 5, pp. 485–560, 1992.
- [23] C. Cortes, V. Vapnik. Support-vector networks, *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [24] C. C. Chang, C. J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, Article 27, 2011.
- [25] F. Pedregosa. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.



Pavla Bromová received her B.Sc. degree in information technologies and M.Sc. degree in intelligent systems from the Faculty of Information Technologies, Brno University of Technology, Czech Republic in 2007 and 2010, respectively. Currently, she is a Ph.D. student in the Department of Information Systems at the Faculty of Information Technologies, Brno University of Technology, Czech Republic.

Her research interests include data mining, machine learning, dimension reduction, and their applications in astronomy.

E-mail: ibromova@fit.vutbr.cz (Corresponding author)



Petr Škoda graduated with honours in astrophysics from the Faculty of Mathematics and Physics, Charles University, Prague, Czechoslovakia in 1987, and received his Ph.D. equivalent title CSc. (Candidatus Scientiarum) in astrophysics at the same faculty in 1996. Since the end of 1987, he has been with the Stellar Department of the Astronomical Institute of the Czechoslovak Academy of Sciences in Ondřejov, Czech

Republic. He is the author and co-author of more than 100 publications and co-author of several standards of virtual observatory. He is a member of the International Astronomical Union and several working groups of International Virtual Observatory Alliance (IVOA) as well as an external supervisor and consultant at main Czech faculties of informatics.

His research interests include the stellar optical spectroscopy, modern astronomical instrumentation in general, however, his strongest interest is in big-data realm of astronomical data archives, the virtual observatory and astroinformatics.

E-mail: skoda@sunstel.asu.cas.cz



Jaroslav Vážný obtained a certified specialist degree (DiS) in computer science in 2000 at SPŠE Pardubice and used to work as a senior developer in Infineon Technologies in Germany, USA and Malaysia. Later he started the studies of astrophysics and received his B.Sc. and M.Sc. degrees in astrophysics at the Faculty of Science, Masaryk University, Brno, Czech Republic in 2008 and 2011, respectively. Now he is a

Ph.D. candidate of this faculty working on his thesis part-time at Stellar Department of Astronomical Institute of the Academy of Sciences of the Czech Republic in Ondřejov.

He has participated in several summer schools about virtual observatory, astronomical data mining and advanced astrostatistics and worked as the European Organization for Nuclear Research (CERN) Summer Student. In 2012 he obtained the certificate of successful finishing of Coursera internet course of machine learning given by Professor Ng. at Stanford University.

His research interests include big data mining and application of machine learning in astronomy as well as on quantum physics and cosmology.

E-mail: jaroslav.vazny@gmail.com