1 2

International Journal of Automation and Computing X(X), X X, X-X
DOI: XXX

Classification of Spectra of Be Stars

Pavla BROMOVÁ[1] Petr ŠKODA[2] Jaroslav VÁŽNÝ[2]

[1]Faculty of Information Technology, Brno University of Technology, Božetěchova 1/2, 612 66 Brno, Czech Republic
[2]Astronomical Institute of the Academy of Sciences of the Czech Republic, Fričova 298, 251 65 Ondřejov

Abstract: An abstract should be a concise summary of the significant items in the paper, including the results and conclusions. It should be about 5% of the length of the article, but not more than about 500 words. Define all nonstandard symbols, abbreviations and acronyms used in the abstract. Do not cite references in the abstract. For a list of suggested keywords, please visit the website at http://www.ijac.net/download/keywrd98.txt.

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

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 a 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 (e.g. [?] and [?]).

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

1.1 Emission Line Stars

There is a lot of stellar objects that may show some important spectral lines in emission. The physical parameters may differ considerable, 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 belong 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 [?] are non-supergiant B type stars whose spectra have or have had at some time, one or more emission lines in the Balmer series. In particular the H_α emission is the dominant feature in spectra of these objects. The emission lines are commonly understood to originate in the flattened circumstellar disk, probably of decretion origin (i.e. created from 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 [?].

The emission and absorption profiles of Be stars vary on different time scales from years to fraction of a day and there 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 [?].

Similar strong emission features in H_α show the B[e] stars [?], 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.

1.3 Be Stars Spectra Archives

The spectra of Be and B[e] stars are dispersed world-wide and most of them are still not yet made available for public. The largest collection of about sixty thousand spectra of 675 different stars represents the BeSS database[1], which is as well accessible with VO protocols. Some individual spectra of Be stars are found in ESO archives, Multimission Archives of NASA (MAST) containing IUE spectra, HST spectra or in DAO archives. Sample of Be stars is also included in ELODIE and SOPHIE archives of OHP observatory. Most of these archives are or are expected to be soon included in VO infrastructure. The rich homogeneous sample of Be and B[e] stars spectra was collected as well by Ondřejov 2m telescope — using the 700mm camera of its coudè spectrograph. It contains about ten thousand spectra of more than 300 Be stars, including one thousand in RVS region. This archive was recently converted to VO-compatible format and is accessible through Simple Spectra Access (SSA) protocol of VO allowing the easy visualization

Manuscript received date; revised date
This work was supported by the specific research grant FIT-S-11-2.

[1]http://basebe.obspm.fr

and analysis of all spectra in VO tools. However access restriction to most of Be stars spectra are applied.

1.4 Motivation

The goal of this application is the exploitation of methods of Astroinformatics in the complex study of variability of emission observed in some types of emission line stars helping to reveal the physical nature of their behaviour and obtain rigorous constrains of their physical parameters useful for their modelling. We will use our long-term experience with research of this class of objects, namely Be and B[e] stars, in scientific analysis of data. We will use the power and efficiency of VO infrastructure to accomplish the extensive multi-wavelength data-mining.

By defining a typical property of a spectra (shape of continuum or presence of a type-specific spectral line), we will be able to classify the observed sample. The appropriate choice of classification criterion will give us a powerful tool for searching of new candidates of interesting kind. In case of data mining of new Be and B[e] stars candidates the differences in intensities in various photometric filters may be used as well as known shape of spectra (e.g. particular type of emission in certain spectral lines). The good source of massive set of stellar spectra seems to be e.g. the SEGUE survey, the extension of SDSS towards brighter galactic objects [?].

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 like e.g. "wine-bottle" (for detailed review see ?]), it is very difficult to construct a simple criteria to identify the Be lines in an automatic manner as required by the amount of spectra considered for processing. However, even simple criteria of combination of three attributes (width, height of Gaussian fit through spectral line and the medium absolute deviation of noise) were sufficient to identify interesting emission line objects in the 187 000 of SDSS SEGUE spectra as shown by ?]. An example of such an object is given in Fig. ??.

To distinguish different types of emission line profiles (which was impossible using only Gaussian fit) we propose a completely new methodology, that 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 e.g. 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 on a larger investigated sample of unclassified objects. In fact it is 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 use directly all points of each spectrum but we have to find a concise description of the spectral features, however conserving most of the original information content.

One of the quite common approaches is to make the Principal Components Analysis (PCA) to get small basis of input vectors for machine training. However, the most promising method is the wavelet decomposition (or multi-resolution analysis) using the prefiltered 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 identification of epilepsy in EEG data [?]. The wavelet transformation is often used for general knowledge mining [?] or a number of other applications. A nice review was given by ?]. In astronomy the wavelet transformation was used recently for estimating stellar physical parameters from Gaia RVS simulated spectra with low SNR [?]. 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 requires the extraction of feature vectors first.

2 Experiment 1: Comparison of Wavelet Types

This section is based on [?].

In DWT, the type of wavelet must be determined. The goal of this experiment is to compare the effect of using different types of wavelet on the results of clustering. Extensive literature exists on wavelets and their applications, e.g. [? ? ? ? ?].

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 choosed two types of different order from each family:

- daubechies (db): order 1, 4
- symlets (sym): order 6, 8
- coiflets (coif): order 2, 3
- biorthogonal (bior): order 2.6, 6.8
- reverse biorthogonal (rbio): order 2.6, 5.5

2.1 Data

The experiment was performed on simulated spectra generated by computer. A collection of 1000 spectra was created trying to cover as many emission lines 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 segment 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 futher rectified [?] by division by corresponding scale. The spectrum P_j for the scale j can be described by (1).

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

2.3 Clustering

Clustering was performed using k-means algorithm into 3-6 clusters. The silhouette method [?] was used for the evaluation. Clustering was performed in 50 iterations and the average silhouette values are presented as the results.

2.4 Results

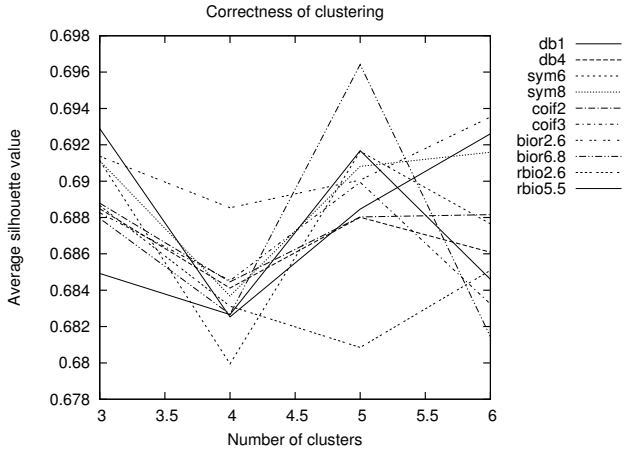

Figure 1 Correctness of clustering for 3, 4, 5, and 6 clusters using different types of wavelet

On Fig. 1 we can see that there are minimal differences in correctness of clustering using different types of wavelet (hundredths of units), which suggests that the type of wavelet has not a big effect on the clustering results.

3 Experiment 2: Comparison of Feature Vectors Using Clustering

This section is based on [?].

In this experiment, we present a feature extraction method based on the wavelet transform and its power spectrum, and an additional value indicating the orientation of the spectral line. Both the discrete and continuous wavelet transform are used. Different feature vectors are created and compared on clustering of Be stars spectra from the archive of the Astronomical Institute of the Academy of Sciences of the Czech Republic. The clustering is 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 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-taps 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 4 classes are illustrated in Fig. 2. The source of the data is the archive of the Astronomical Institute of the Academy of Sciences of the Czech Republic.

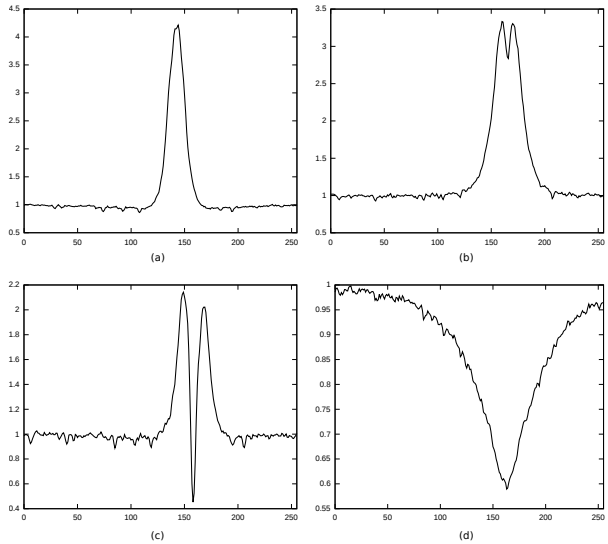

Figure 2 Examples of selected data samples typical for each of 4 classes. Figures (a), (b), and (c) are spectra of Be stars, Fig. (d) is a normal star. In (a) there is a pure emission on H_α spectral line, (b) contains a small absorption part (less than 1/3 of the height), (c) contains larger absorption part (more than 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. Due to the power of coefficients, two data samples with the same shape but opposite orientation of the spectral line would yield the 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 and one negative value. The question is which absolute value to choose. In this experiment, we have

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 do not have any reference method of feature extraction from Be stars for comparison, we compare our results with a common method of feature extraction from time series using wavelets, which keeps N largest coefficients of wavelet transform and the rest of the coefficients are set to zero [?]. 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 a spectral line.

Feature Vectors Different kinds of feature vectors were created from 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. **DWPS + orientation 1**: one part of a feature vector is the wavelet power spectrum of DWT, normalized so that its total energy equals to 1. Second part of a feature vector is a value indicating the orientation of a spectral line – lines oriented up have the value 1, lines oriented down have the value -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 a 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. **CWPS 16 + orientation 1**: wavelet power spectrum (normalized) of CWT performed with 16 scales. The same orientation as in the previous cases with DWPS.

10. **CWPS 8 + orientation 1**: wavelet power spectrum (normalized) of CWT performed with 8 scales. The same orientation 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 simply to 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 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 a size of a number of classes, and where the element on a position (i, j) corresponds to the number of samples with an output class i and a target class j. In a case of a perfect clustering, all values besides the main diagonal would be egual 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 3.2.

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

This section is based on [?].

In this experiment, we propose several feature extraction methods based on the discrete wavelet transform (DWT). The data set is the same as in the previous experiment. A small segment containing the H_α line is selected for feature extraction. Classification is performed using the support

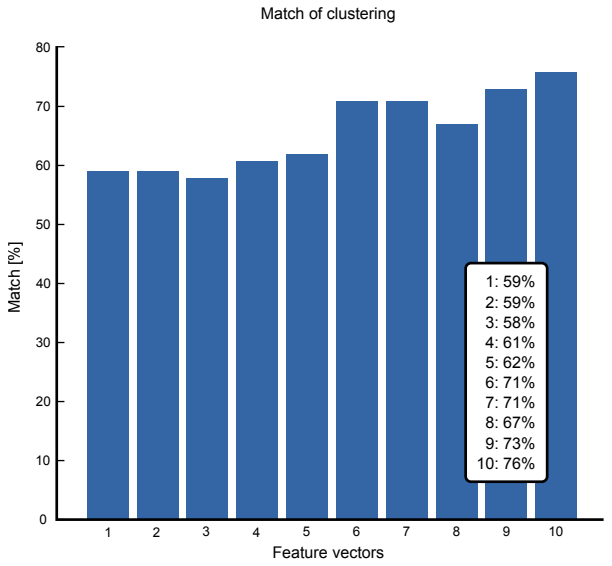

Figure 3 The match of the clustering using different feature vectors

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

4.1 Feature Extraction

In this experiment, the wavelet transform was performed using A Cross-platform Discrete Wavelet Transform Library [2]

The selected data samples were decomposed into J scales using the discrete wavelet transform with CDF 9/7 [?] 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 and the other with 9 coefficients.

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

On each obtained subband, the following descriptor was calculated forming the resulting feature vector as (3). The individual methods are further explained in detail.

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

Wavelet power spectrum Described in 2.2.

Euclidean norm The Euclidean or ℓ^2 norm is the intuitive notion of length of a vector. The norm for the specific subband j can be calculated as $\|W_j\|_2$ by (4).

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

Maximum norm Similarly, the maximum or infinity norm can be defined as maximal value of DWT magnitudes (5).

[2] A Cross-platform Discrete Wavelet Transform Library: http://www.fit.vutbr.cz/research/view_product.php?id=211

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

Arithmetic mean The mean (6) is the sum of a 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 bellow.

$$\mu_j = \mathrm{E}\left[W_j\right] \tag{6}$$

Standard deviation The standard deviation (7) is the square root of the variance of the specific wavelet subband at the scale j. It indicates how much variation exists from the arithmetic mean.

$$\sigma_j = \left(\mathrm{E}\left[(W_j - \mu_j)^2\right]\right)^{1/2} \tag{7}$$

4.2 Classification

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

There are two parameters for a RBF kernel: C and γ. It is not known beforehand which C and γ are 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 the 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. We have tried exponentially growing sequences of $C = 2^{-5}, 2^{-3}, \ldots, 2^{15}$ and $\gamma = 2^{-15}, 2^{-13}, \ldots, 2^{3}$). 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 to the interval $[0, 1]$.

4.3 Results

The results are obtained for different feature extraction techniques by the 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.

The results of all feature extraction methods are comparable with satisfying accuracy approaching the accuracy of a feature vector consisting of the original values of the stellar spectrum without the feature extraction. Moreover, the results are significantly better than in the case 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 in the case of the original data without the feature extraction.

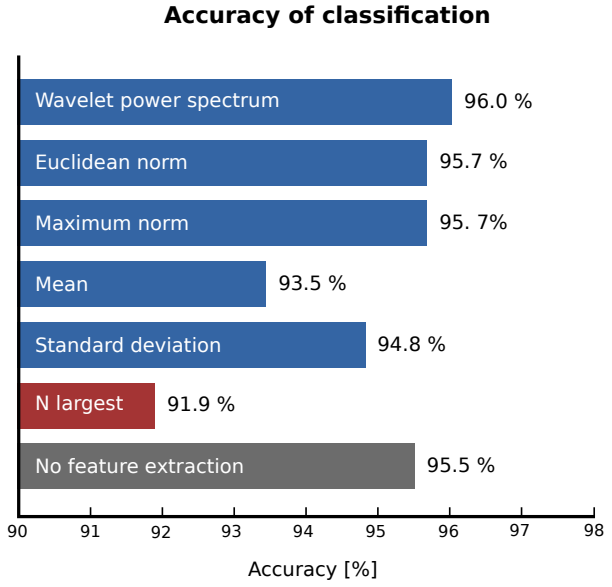

Figure 4 Accuracy of classification for different feature extraction methods

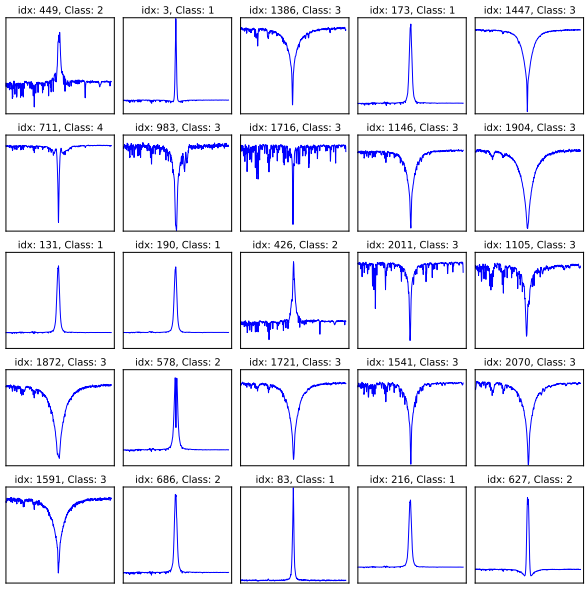

Figure 5 25 random samples from all categories

5 Experiment 4: Classification Without Feature Extraction

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

5.1 Data Selection

Training set consists of 2164 spectra from Ondrejov archive [3] divided into 4 distinct categories based on the region around Balmer H-alpha line (which is interesting region for that type of stars). The spectra were normalized and trimmed to 100Åaround H-alpha. Numbers of spectra in individual categories are following:

category	count
1	408
2	289
3	1366
4	129

For better understanding of the categories characteristics there is a plot of 25 random samples in the Fig. 5 and characteristics spectrum of individual categories created as a sum of all spectra in corresponding category in the Fig. 6.

PCA (Principal component analysis) was also performed to visually check if there is a separation (and therefore a chance) to discriminate between individual classes. See the Fig. 7.

5.2 Classification

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

[3] Ondrejov archive: http://physics.muni.cz/ ssa/archive/

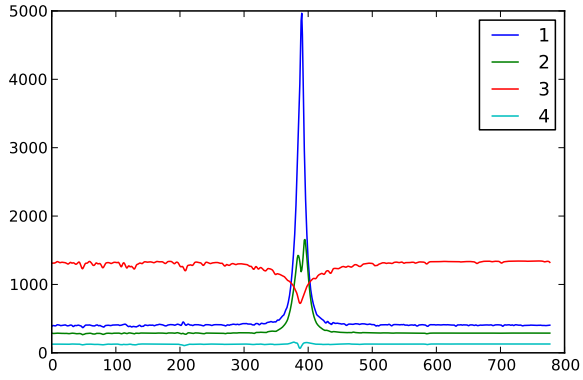

Figure 6 Characteristic spectrum of individual categories created as a sum of all spectra in corresponding category

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 1.

Based on this result, values $C = 100.0$ and $\gamma = 0.01$ were used in following experiments.

5.3 Results

Mean score was 0.988 (+/-0.002). There is a detailed report (now based on test sample=0.25) in Table 2.

Learning curve is an important tool which help us understand the behavior of the selected model. As you can see on Fig. 8 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

parameters	score
C=100.0, gamma=0.01:	0.985 (+/-0.003) *
C=10.0, gamma=0.1:	0.978 (+/-0.003) *
C=100.0, gamma=0.1:	0.977 (+/-0.004) *
C=10.0, gamma=0.01:	0.973 (+/-0.002)
C=1.0, gamma=0.1:	0.970 (+/-0.003)
C=100.0, gamma=0.001:	0.969 (+/-0.002)
C=1.0, gamma=1.0:	0.966 (+/-0.003)
C=10.0, gamma=1.0:	0.965 (+/-0.004)
C=100.0, gamma=1.0:	0.965 (+/-0.004)
C=1.0, gamma=0.01:	0.958 (+/-0.002)
C=10.0, gamma=0.001:	0.956 (+/-0.003)
C=100.0, gamma=0.0001:	0.953 (+/-0.003)
C=0.1, gamma=0.1:	0.929 (+/-0.005)
C=10.0, gamma=0.0001:	0.915 (+/-0.004)
C=1.0, gamma=0.001:	0.914 (+/-0.003)
C=0.1, gamma=0.01:	0.908 (+/-0.003)
C=0.1, gamma=1.0:	0.885 (+/-0.004)
C=1.0, gamma=0.0001:	0.811 (+/-0.003)
C=0.1, gamma=0.001:	0.811 (+/-0.003)
C=0.1, gamma=0.0001:	0.785 (+/-0.003)

Table 1 Results of the grid-search

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
avg/total	0.99	0.99	0.99	541

Table 2 Results of classification

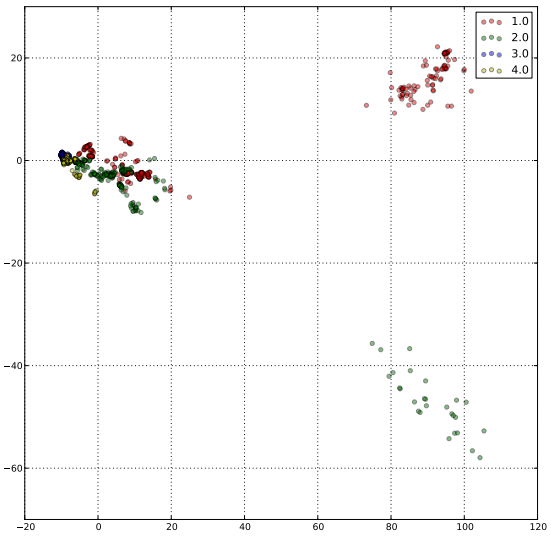

Figure 7 PCA separation of individual classes

this model and data.

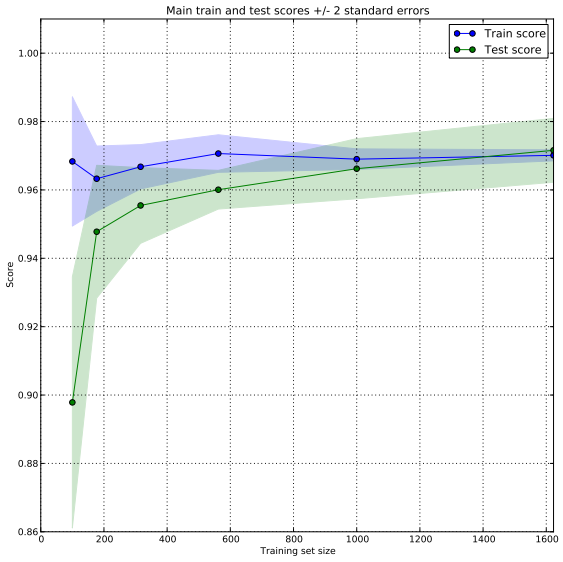

Figure 8 Learning curve

Miss-clasification There were 29 miss-classified cases (based on $\text{test}_{\text{size}}$= 0.25). The Fig. 9 shows that spectra.

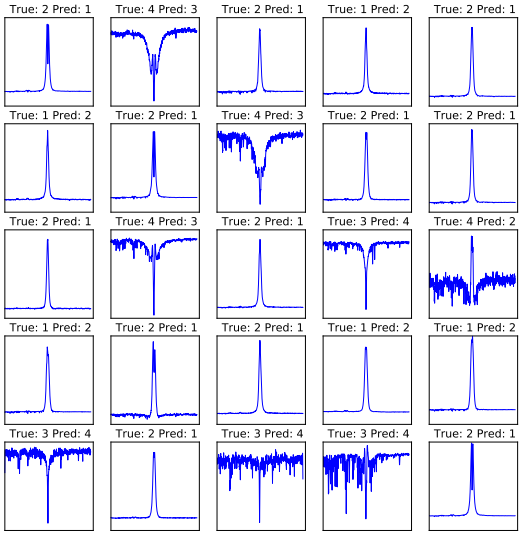

Figure 9 The miss-classified samples

6 Conclusion

This paper presents the results of classification of Be stars using different feature extraction methods based on the discrete wavelet transform. The support vector machines (SVM) is used for classification. The results are given by the accuracy of classification.

The results of all five tested feature extraction methods are comparable with satisfying accuracy approaching the accuracy of a feature vector consisting of the original values of the stellar spectrum without the feature extraction. Moreover, the results are significantly better than in the case of the common method of feature extraction from time series using wavelets, 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 in the case of the original data without the feature extraction.

Acknowledgement

In general, limit acknowledgments to those who helped directly in the research itself or during discussions on the subject of the research. Financial support of all kinds acknowledgments are placed in the unnumbered footnote on the first page.

References

[1] K. Borne, A. Accomazzi, J. Bloom, R. Brunner, D. Burke, N. Butler, D. F. Chernoff, B. Connolly, A. Connolly, A. Connors, C. Cutler, S. Desai, G. Djorgovski, E. Feigelson, L. S. Finn, P. Freeman, M. Graham, N. Gray, C. Graziani, E. F. Guinan, J. Hakkila, S. Jacoby, W. Jefferys, Kashyap, B. Kelly, K. Knuth, D. Q. Lamb, H. Lee, T. Loredo, A. Mahabal, M. Mateo, B. McCollum, A. Muench, M. Pesenson, V. Petrosian, F. Primini, P. Protopapas, A. Ptak, J. Quashnock, M. J. Raddick, G. Rocha, N. Ross, L. Rottler, J. Scargle, A. Siemiginowska, I. Song, A. Szalay, J. A. Tyson, T. Vestrand, J. Wallin, B. Wandelt, I. M. Wasserman, M. Way, M. Weinberg, A. Zezas, E. Anderes, J. Babu, J. Becla, J. Berger, P. J. Bickel, M. Clyde, I. Davidson, D. van Dyk, T. Eastman, B. Efron, C. Genovese, A. Gray, W. Jang, E. D. Kolaczyk, J. Kubica, J. M. Loh, X.-L. Meng, A. Moore, R. Morris, T. Park, R. Pike, J. Rice, J. Richards, D. Ruppert, N. Saito, C. Schafer, P. B. Stark, M. Stein, J. Sun, D. Wang, Z. Wang, L. Wasserman, E. J. Wegman, R. Willett, R. Wolpert, and M. Woodroofe. Astroinformatics: A 21st Century Approach to Astronomy. In *astro2010: The Astronomy and Astrophysics Decadal Survey*, volume 2010 of *Astronomy*, page 6P, 2009.

[2] P. Bromová, D. Bařina, P. Škoda, and J. Zendulka. Classification of be stars using feature extraction based on discrete wavelet transform. In *Proceedings of the 12th annual conference Znalosti 2013*. MATFYZPRESS, 2013. submitted.

[3] P. Bromová, P. Škoda, and J. Zendulka. Wavelet based feature extraction for clustering of be stars. In *Nostradamus 2013*, 2013.

[4] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[5] A. Cohen, Ingrid Daubechies, and J.-C. Feauveau. Biorthogonal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 45(5):485–560, 1992.

[6] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[7] I. Daubechies. *Ten lectures on wavelets*. CBMS-NSF regional conference series in applied mathematics. Society for Industrial and Applied Mathematics, 1994.

[8] R. W. Hanuschik, W. Hummel, E. Sutorius, O. Dietle, and G. Thimm. Atlas of high-resolution emission and shell lines in Be stars. Line profiles and short-term variability. , 116:309–358, April 1996.

[9] Pari Jahankhani, Kenneth Revett, and Vassilis Kodogiannis. Data mining an EEG dataset with an emphasis on dimensionality reduction. In *2007 IEEE Symposium on Computational Intelligence and Data Mining, Vols 1 and 2*, pages 405–412. IEEE, 2007.

[10] G. Kaiser. *A friendly guide to wavelets*. Birkhäuser, 1994.

[11] T. Li, S. Ma, and M. Ogihara. Wavelet methods in data mining. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 553–571. Springer, 2010.

[12] Tao Li, Qi Li, Shenghuo Zhu, and Mitsunori Ogihara. A survey on wavelet applications in data mining. *SIGKDD Explor. Newsl.*, 4:49–68, December 2002.

[13] Y. Liu, X. San Liang, and R. H. Weisberg. Rectification of the bias in the wavelet power spectrum. *Journal of Atmospheric and Oceanic Technology*, 24(12):2093–2102, 2007.

[14] S. Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, 3rd edition, 2008.

[15] M. Manteiga, D. Ordóñez, C. Dafonte, and B. Arcay. ANNs and Wavelets: A Strategy for Gaia RVS Low S/N Stellar Spectra Parameterization. , 122:608–617, May 2010.

[16] Y. Meyer and D.H. Salinger. *Wavelets and Operators*. Number sv. 1 in Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1995.

[17] Murugappan Murugappan, Ramachandran Nagarajan, and Sazali Yaacob. Combining Spatial Filtering and Wavelet Transform for Classifying Human Emotions Using EEG Signals. *JOURNAL OF MEDICAL AND BIOLOGICAL ENGINEERING*, 31(1):45–51, 2011.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[19] J. M. Porter and T. Rivinius. Classical Be Stars. , 115:1153–1170, October 2003.

[20] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(0):53 – 65, 1987.

[21] P. Škoda and J. Vážný. Searching of new Emission-line Stars using the Astroinformatics Approach. In Ballester, P. and Egret, D., editor, *ADASS XXI*, Astronomical Society of the Pacific Conference Series, page *in print*, 2012.

[22] G. Strang and T. Nguyen. *Wavelets and filter banks*. Wellesley-Cambridge Press, 1996.

[23] B. Yanny, C. Rockosi, H. J. Newberg, G. R. Knapp, J. K. Adelman-McCarthy, B. Alcorn, S. Allam, C. Allende Prieto, D. An, K. S. J. Anderson, S. Anderson, C. A. L. Bailer-Jones, S. Bastian, T. C. Beers, E. Bell, V. Belokurov, D. Bizyaev, N. Blythe, J. J. Bochanski, W. N. Boroski, J. Brinchmann, J. Brinkmann, H. Brewington, L. Carey, K. M. Cudworth, M. Evans, N. W. Evans, E. Gates, B. T. Gänsicke, B. Gillespie, G. Gilmore, A. N. Gomez-Moran, E. K. Grebel, J. Greenwell, J. E. Gunn, C. Jordan, W. Jordan, P. Harding, H. Harris, J. S. Hendry, D. Holder, I. I. Ivans, Ž. Ivezič, S. Jester, J. A. Johnson, S. M. Kent, S. Kleinman, A. Kniazev, J. Krzesinski, R. Kron, N. Kuropatkin, S. Lebedeva, Y. S. Lee, R. F. Leger, S. Lépine, S. Levine, H. Lin, D. C. Long, C. Loomis, R. Lupton, O. Malanushenko, V. Malanushenko, B. Margon, D. Martinez-Delgado, P. McGehee, D. Monet, H. L. Morrison, J. A. Munn, E. H. Neilsen, A. Nitta, J. E. Norris, D. Oravetz, R. Owen, N. Padmanabhan, K. Pan, R. S. Peterson, J. R. Pier, J. Platson, P. R. Fiorentin, G. T. Richards, H.-W. Rix, D. J. Schlegel, D. P. Schneider, M. R. Schreiber, A. Schwope, V. Sibley, A. Simmons, S. A. Snedden, J. A. Smith, L. Stark, F. Stauffer, M. Steinmetz, C. Stoughton, M. Subba Rao, A. Szalay, P. Szkody, A. R. Thakar, S. Thirupathi, D. Tucker, A. Uomoto, D. Vanden Berk, S. Vidrih, Y. Wadadekar, S. Watters, R. Wilhelm, R. F. G. Wyse, J. Yarger, and D. Zucker. SEGUE: A Spectroscopic Survey of 240,000 Stars with g = 14-20. , 137:4377–4399, May 2009.

[24] Y. Zhang, H. Zheng, and Y. Zhao. Knowledge discovery in astronomical data. In *SPIE Conference*, volume 7019, August 2008.

[25] Y. Zhao, I. Raicu, and I. Foster. Scientific Workflow Systems for 21st Century e-Science, New Bottle or New Wine? *arXiv:0808.3545*, August 2008.

[26] F.-J. Zickgraf. Kinematical structure of the circumstellar environments of galactic B[e] -type stars. , 408:257–285, September 2003.

[27] J. Zorec and D. Briot. Critical study of the frequency of Be stars taking into account their outstanding characteristics. , 318:443–460, February 1997.

[28] P. Škoda, P. Bromová, and J. Zendulka. Feature extraction using wavelet power spectrum for stellar spectra clustering. In *Proceedings of the 11th annual conference Znalosti 2012*, pages 31–40. MATFYZPRESS, 2012.

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.

She has published 4 conference papers. Her research interest covers data mining, dimension reduction, and their applications in astrophysics.

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

Second-Bb Author received his B. Sc. and M. Sc. degrees in mechanical engineering from the *** University, China, in 1977 and 1984, respectively, and the Ph. D. degree in computing from *** University, UK in 1992. In 1994, he was a faculty member at *** University, China and in 1996 at *** University, USA. Currently, he is a professor in the Department of Information System Engineering at *** University, PRC.

He has published about 100 refereed journal and conference papers. His research interest covers robotics, feedback control systems, and control theory.

Prof. Author received research award from Science Foundation, and the Best Paper Award of the IS International Conference in 2000 and 2006, respectively. He is a member of SICE, IEE and IEEE.

E-mail: ijac@ia.ac.cn