

# Classification of Be Stars Using Feature Extraction Based on Discrete Wavelet Transform

Pavla Bromová<sup>1</sup>, David Bařina<sup>1</sup>, Petr Škoda<sup>2</sup>, Jaroslav Vážný<sup>2</sup>, and Jaroslav Zendulka<sup>1</sup>

<sup>1</sup> Faculty of Information Technology, Brno University of Technology,  
Bořetěchova 1/2, 612 66 Brno, Czech Republic

<sup>2</sup> Astronomical Institute of the Academy of Sciences of the Czech Republic,  
Fričova 298, 251 65 Ondřejov

**Abstract.** We describe the initial experiments in the field of automated classification of spectral line profiles of emission line stars. We attempt to automatically identify Be and B[e] stars spectra in large archives and classify their types in an automatic manner. To distinguish different types of emission line profiles, we propose a completely new methodology, that seems to be not yet used in astronomy. Due to the expected size of spectra collections, the dimension reduction is required. We propose to perform the discrete wavelet transform (DWT) of the spectra, compute the wavelet power spectrum and other statistical metrics over the wavelet coefficients, and use them as feature vectors in classification. The results show that using proposed method of data transformation we can reduce the number of attributes and the processing time to a small fraction, and moreover increase the accuracy.

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

## 1 Introduction

Technological progress and growing computing power are causing data avalanche in almost all sciences, including astronomy. The full exploitation of these massive distributed data sets clearly requires automated methods. One of the difficulties is the inherent size and dimensionality of the data. The efficient classification requires that we reduce the dimensionality of the data in a way that preserves as many of the physical correlations as possible.

Be stars are hot, rapidly rotating B-type stars with equatorial gaseous disk producing prominent emission  $H_\alpha$  lines in their spectrum [17]. The emission lines are bright lines in a spectrum caused when the atoms and molecules in a hot gas emit extra light at certain wavelengths [7]. The distribution of these lines in a spectrum is unique for each chemical element.  $H_\alpha$  line is created by hydrogen with a wavelength of 656.28 nm. Be stars show a number of different shapes of the emission lines, as we can see in Fig. 1. These variations reflect underlying physical properties of a star.

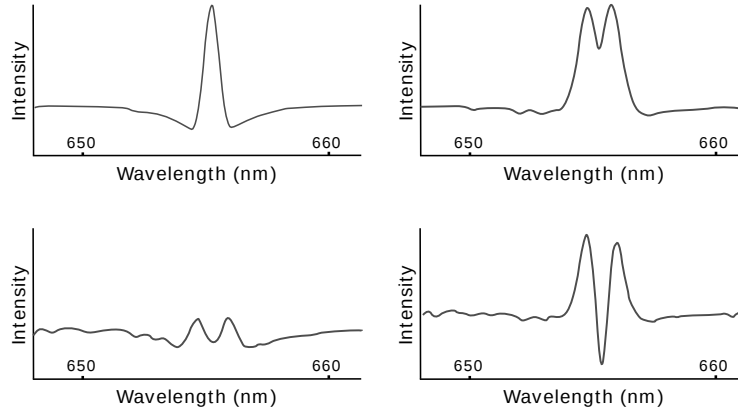


Fig. 1: Examples of typical shapes of emission lines in spectra of Be stars

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, 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 among nearly two hundred thousand of SDSS SEGUE spectra [18].

To distinguish different types of emission line profiles (which is 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 years to many similar problems like a detection of particular EEG activity. 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.

We propose to perform the discrete wavelet transform (DWT) of the spectra, compute the wavelet power spectrum and other statistical variables over the wavelet coefficients, and use them as feature vectors for classification. 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 [9]. Extensive literature exists on wavelets and their applications, e.g. [13, 6, 10, 15, 16, 11]. In astronomy the wavelet transform was used recently for estimating stellar physical parameters from Gaia RVS simulated spectra with low SNR [14]. 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.

In next chapters we describe the experiment with feature extraction using DWT in an attempt to identify the best method as well as verification of the results using classification of both extracted feature vectors and original data points.

## 2 Data

The source of data is the archive of the Astronomical Institute of the Academy of Sciences of the Czech Republic (AI ASCR). The data set consists of 2164 spectra of Be stars and also normal stars divided into 4 classes (with 408, 289, 1338, and 129 samples) based on the shape of the  $H_\alpha$  line. The original sample contains approximately  $\sim 2000$  values around  $H_\alpha$  line.

For better understanding of the categories characteristics there is a plot of 25 random samples in Fig. 2 and characteristic spectrum of individual categories created as a sum of all spectra in corresponding category in Fig. 3.

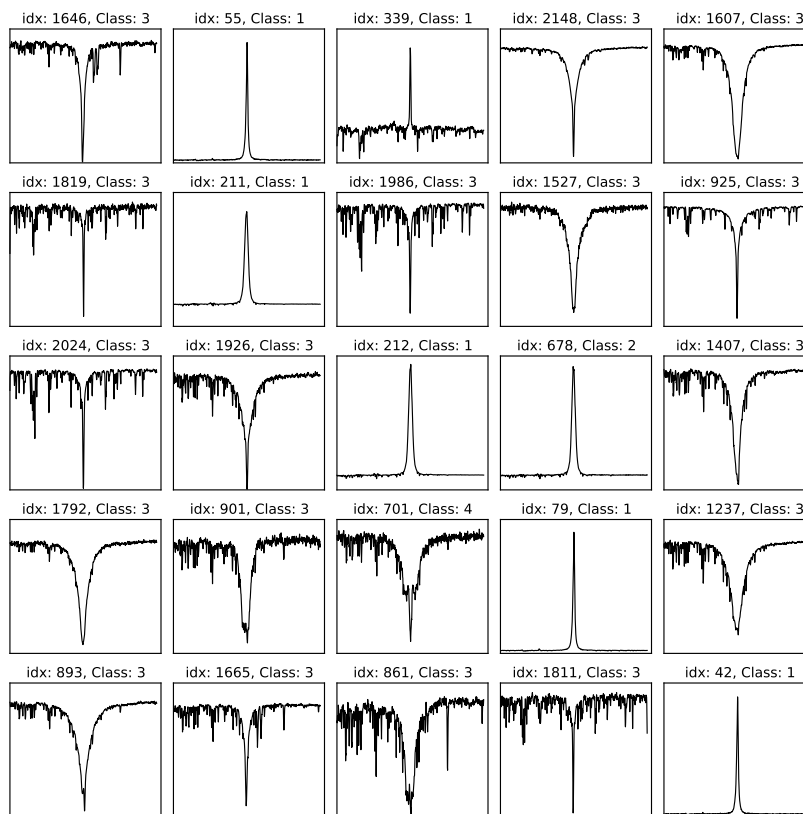


Fig. 2: Random samples of spectra from all categories [2]

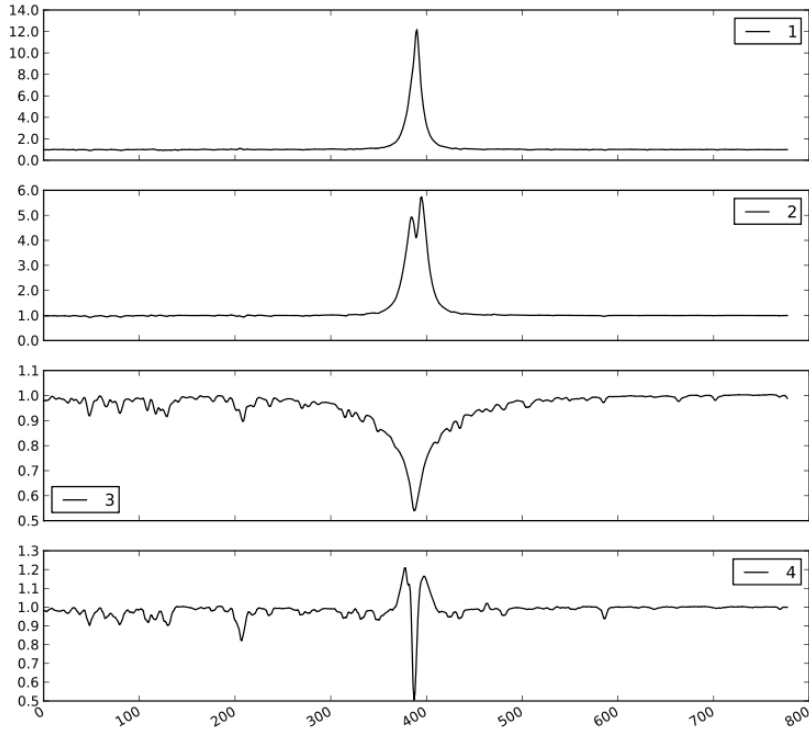


Fig. 3: Characteristic spectrum of individual categories created as a sum of all spectra in corresponding category [2]. Categories 1, 2, and 4 consists of spectra of Be stars, category 3 contains spectra of normal stars. Spectra in cat. 1 are characterized by a pure emission in  $H_\alpha$  spectral line. Cat. 2 contains a small absorption part (less than 1/3 of the height), cat. 3 contains larger absorption part (more than 1/3 of the height). Spectra of normal stars in cat. 3 are characterized by a pure absorption.

### 3 Data Transformation

#### 3.1 Centering

First, the centers of emission (resp. absorption) lines are aligned to the center, so that the influence of the position of the emission in a spectrum on the classification is minimized, as we are interested only in the shape of the emission line.

Centering is done by subtracting the median of a spectrum from the spectrum and alignment of the maximal magnitude of the spectrum to the center.

### 3.2 Wavelet Transform

The wavelet transform was performed using the Cross-platform Discrete Wavelet Transform Library [1]. The selected data samples were decomposed into  $J$  scales using the discrete wavelet transform as

$$W_{j,n} = \langle x, \psi_{j,n} \rangle, \quad (1)$$

where  $W_{j,n}$  is a wavelet coefficient at  $j$ -th scale and  $n$ -th position,  $x$  is a data vector, and  $\psi$  is the CDF 9/7 [4] wavelet. 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.

### 3.3 Feature Extraction

Different methods of computing a feature vector were used and then compared. The feature vector

$$\mathbf{v} = (v_j)_{1 \leq j < J} \quad (2)$$

consists of  $J$  elements  $v_j$  calculated for each obtained subband (scale)  $j$  of wavelet coefficients using one of the methods described below.

All elements in one feature vector were computed by the same method. Some of the individual methods are further explained in details.

**Wavelet 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 [12] by division by corresponding scale. The WPS for the scale  $j$  can be described by

$$v_j = 2^{-j} \sum_n |W_{j,n}|^2. \quad (3)$$

**Euclidean norm** is given by

$$v_j = \left( \sum_n |W_{j,n}|^2 \right)^{1/2}. \quad (4)$$

Similarly, the other descriptors were calculated using the following metrics which are listed without well known definitions.

**Maximum**

**Mean**

**Median**

**Variance**

**Standard deviation**

## 4 Classification

Classification of resulting feature vectors is performed with the support vector machines (SVM) [5] using the LIBSVM [3] library. The radial basis function (RBF) is used as a kernel function. This kernel nonlinearly maps samples into a higher dimensional space so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear. The second reason is the number of hyperparameters which influences the complexity of the model. The RBF kernel has fewer hyperparameters than the other kernels [8].

There are two hyperparameters for a RBF kernel:  $C$  and  $\gamma$ . It is not known beforehand which  $C$  and  $\gamma$  are best for a given problem, therefore some kind of model selection (parameter search) must be done. The goal is to identify optimal  $C$  and  $\gamma$  so that the classifier can accurately predict unknown data. However, it may not be useful to achieve high training accuracy, because of the overfitting problem.

A common strategy to deal with the overfitting problem is known as cross-validation. In  $v$ -fold cross-validation, the training set is divided into  $v$  subsets of equal size. Sequentially one subset is tested using the classifier trained on the remaining  $v-1$  subsets. Thus, each instance of the whole training set is predicted once so the cross-validation accuracy is the percentage of data which are correctly classified. The cross-validation procedure can prevent the overfitting problem.

A strategy known as grid-search was used to find the parameters  $C$  and  $\gamma$  and to give the results of classification. Various pairs of  $C$  and  $\gamma$  values were tried and each combination of parameter choices was checked using 5-fold cross validation. The results are given by the best cross-validation accuracy. We have tried exponentially growing sequences of  $C = 2^{-5}, 2^{-3}, \dots, 2^{15}$  and  $\gamma = 2^{-15}, 2^{-13}, \dots, 2^3$ .

## 5 Results

We present the results of classification using different feature extraction methods and we compare them with the results without using any feature extraction method. Besides formerly described feature vectors, we used also the original data, the data after centering, and all the DWT coefficients. The results are in Table 1. It includes the length of a feature vector, the accuracy and the values of parameters  $C$  and  $\gamma$  of RBF kernel given by the best cross-validation accuracy of grid search, and measured time. The evaluation was performed on desktop PC equipped with AMD Athlon 64 X2 processor at 2.1 GHz.

The Table 1 shows that the accuracy of all tested feature vector descriptors is very similar and moreover in most of the cases even better than the accuracy of the original data. We can also see that the classification of the original data is very time-consuming. Using proposed method of data transformation (centering, DWT, and one of the descriptors) we can reduce the number of attributes and the processing time to a small fraction, and moreover increase the accuracy.

Feature vector	Length	Accuracy [%]	Time [min.]	$\log_2 C$	$\log_2 \gamma$
Original data	1997	96.716	~330		
Centering	1997	98.2886	~330		
DWT	1997	98.1961	~330		
Median	10	97.6411	1.72	7	-3
Std. deviation	10	97.2248	2.02	7	1
Euclid. norm	10	97.1785	1.93	9	-5
Maximum	10	97.0398	1.85	7	-3
Mean	10	96.531	2.08	13	-1
WPS	10	94.6809	2.22	15	1
Variance	10	94.4496	2.48	15	-5

Table 1: Results of classification of different feature vectors. The table includes the length of a feature vector, the accuracy and the values of parameters  $C$  and  $\gamma$  of RBF kernel given by the best cross-validation accuracy of grid search, and measured time.

## 6 Conclusion

In this paper, we describe the experiment with classification of spectra of Be stars using different feature extraction methods based on the discrete wavelet transform in an attempt to identify the best method. We present the results of classification of different extracted feature vectors as well as the original data.

From the results we can see that the classification of the original data is very time-consuming. Using proposed method of data transformation (centering, DWT, and one of the descriptors) we can reduce the number of attributes and the processing time to a small fraction, and moreover increase the accuracy.

In 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 use for clustering of all spectra in the archive of AI ASCR, and possibly to find new interesting candidates.

## Acknowledgement

This work has been supported by the grant GACR 13-08195S of the Czech Science Foundation, the project CEZ MSM0021630528 Security-Oriented Research in Information Technology, the specific research grant FIT-S-11-2, the EU FP7-ARTEMIS project IMPART (grant no. 316564) and the national Technology Agency of the Czech Republic project RODOS (no. TE01020155).

## References

1. D. Bařina and P. Zemćik. A cross-platform discrete wavelet transform library. Authorised software, Brno University of Technology. Software available at [http://www.fit.vutbr.cz/research/view\\_product.php?id=211](http://www.fit.vutbr.cz/research/view_product.php?id=211), 2010-2013.
2. P. Bromova, P. řkoda, and J. Vazny. Classification of spectra of emission line stars using machine learning techniques. *International Journal of Automation and Computing*, 2013. Submitted.
3. C.-C. Chang and C.-J. 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>.
4. A. Cohen, I. Daubechies, and J.-C. Feauveau. Biorthogonal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 45(5):485–560, 1992.
5. C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
6. I. Daubechies. *Ten lectures on wavelets*. CBMS-NSF regional conference series in applied mathematics. Society for Industrial and Applied Mathematics, 1994.
7. U. o. T. Dept. Physics & Astronomy. Stars, galaxies, and cosmology. <http://csep10.phys.utk.edu/astr162/lect/index.html>. Accessed: 16/01/2013.
8. C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2010.
9. P. Jahankhani, K. Revett, and V. Kodogiannis. Data mining an EEG dataset with an emphasis on dimensionality reduction. In *IEEE Symposium on Computational Intelligence and Data Mining*, volume 1,2, pages 405–412, 2007.
10. G. Kaiser. *A friendly guide to wavelets*. Birkhuser, 1994.
11. T. Li, S. Ma, and M. Ogihara. Wavelet methods in data mining. In O. Maimon and L. Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 553–571. Springer, 2010.
12. 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.
13. S. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, 3rd edition, 2009.
14. M. Manteiga, D. Ordonez, C. Dafonte, and B. Arcay. ANNs and wavelets: A strategy for gaia RVS low S/N stellar spectra parameterization. In *Publications of the Astronomical Society of the Pacific*, volume 122, pages 608–617, 2010.
15. Y. Meyer and D. Salinger. *Wavelets and Operators*. Number sv. 1 in Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1995.
16. G. Strang and T. Nguyen. *Wavelets and filter banks*. Wellesley-Cambridge Press, 1996.
17. O. Thizy. Classical Be stars high resolution spectroscopy. *Society for Astronomical Sciences Annual Symposium*, 27:49, 2008.
18. P. řkoda and J. Vazny. Searching of new emission-line stars using the astroinformatics approach. In *Astronomical Data Analysis Software and Systems XXI, Astronomical Society of the Pacific Conference Series*, volume 461, page 573, 2012.