Astronomical Data Analysis Software and Systems XXIII
ASP Conference Series, Vol. 485
N. Manset and P. Forshay, eds.

Classification of Spectra of Emission-line Stars Using Feature Extraction Based on Wavelet Transform

Pavla Bromová,[1] David Bařina,[1] Petr Škoda,[2] Jaroslav Vážný,[2] and Jaroslav Zendulka[1]

[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, Czech Republic*

Abstract. Our goal is to automatically identify spectra of emission (Be) stars in large archives and classify their types based on a typical shape of the H_α emission line. Due to the length of spectra, classification of the original data is very time-consuming. In order to lower computational requirements and enhance the separability of the classes, we have to find a reduced representation of spectral features, however conserving most of the original information content. As the Be stars show a number of different shapes of emission lines, it is not easy to construct simple criteria (like e.g. Gaussian fits) to distinguish the emission lines in an automatic manner. We proposed to perform the wavelet transform of the spectra, calculate statistical metrics from the wavelet coefficients, and use them as feature vectors for classification. In this paper, we compare different wavelet transforms, different wavelets, and different statistical metrics in an attempt to identify the best method.

1. Introduction

Our goal is to automatically identify spectra of H_α emission stars (Be and B[e]) in large archives and classify their types. Due to the length of spectra, classification of the original data is very time-consuming. We can't simply use all points of each spectrum but we have to find a reduced representation of spectral features, however conserving most of the original information content. As the Be stars show a number of different shapes of emission lines, it is not easy to construct simple criteria (like e.g. Gaussian fits) to distinguish the emission lines in an automatic manner.

In Bromová et al. (2013) we proposed a feature extraction method which reduces the number of attributes from ~2000 to 10, can reduce the processing time from ~330 minutes to ~1 minute and increase the accuracy from 96.7 % to 98.1 % at the same time. In this paper, we build on this work by experiments with more feature extraction methods and their parameters in an attempt to find the best method.

2. Data

The source of data is the archive of the Astronomical Institute of the Academy of Sciences of the Czech Republic in Ondřejov. The spectra were obtained with a spectro-

graph of Ondřejov Observatory 2m telescope. Dataset consists of 2164 spectra of Be and normal stars divided into 4 classes (408, 289, 1338, and 129 samples) based on the shape of the H_α line. The original sample contains approximately 2000 values around H_α line. Spectra typical for individual categories are sketched in Figure 1.

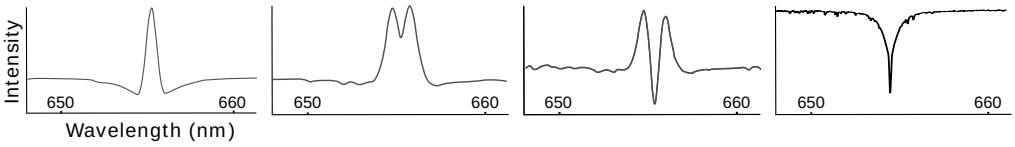

Figure 1. Spectra typical for individual categories.

3. Feature Extraction

Centering. First, the centers of emission 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.

Wavelet Transform. The discrete (DWT) and stationary (SWT) wavelet transforms were employed for comparison, using the Cross-platform Discrete Wavelet Transform Library (Bařina & Zemčík 2013). Here, the selected data samples were decomposed into J scales as

$$W_{j,n} = \langle x, \psi_{j,n} \rangle, \tag{1}$$

where $W_{j,n}$ is a wavelet coefficient at j-th scale and n-th position, x is an input spectrum, and ψ is a wavelet function. Two wavelets were tested: CDF 9/7 and CDF 5/3 (Cohen et al. 1992). These wavelets are employed for lossy or lossless compression in JPEG 2000 and Dirac compression standards.

Aggregate Function. Different functions were used for feature extraction from the wavelet coefficients and then compared: wavelet power spectrum, Euclidean norm, maximum norm, mean, median, variance, and standard deviation.

The feature vector

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

consists of J elements v_j calculated for each obtained subband (scale) j of wavelet coefficients using one of the functions above. All elements in one feature vector were computed using the same function.

Specifically, the wavelet power spectrum for the scale j was calculated as

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

The bias of this power spectrum was further rectified (Liu et al. 2007) by division by corresponding scale.

4. Classification

Classification of resulting feature vectors is performed with the support vector machines (Cortes & Vapnik 1995) using the LIBSVM library (Chang & Lin 2011). The radial basis function (RBF) is used as a kernel function. There are two parameters for a RBF kernel: C and γ. A strategy known as grid-search was used to find the parameters C and γ. Various pairs of C and γ values were tried and each combination was checked using 5-fold cross validation. We have tried exponentially growing sequences of $C = 2^{-5}, 2^{-3}, \ldots, 2^{15}$ and $\gamma = 2^{-15}, 2^{-13}, \ldots, 2^{3}$. The results are given by the combination of parameters with the best cross-validation accuracy.

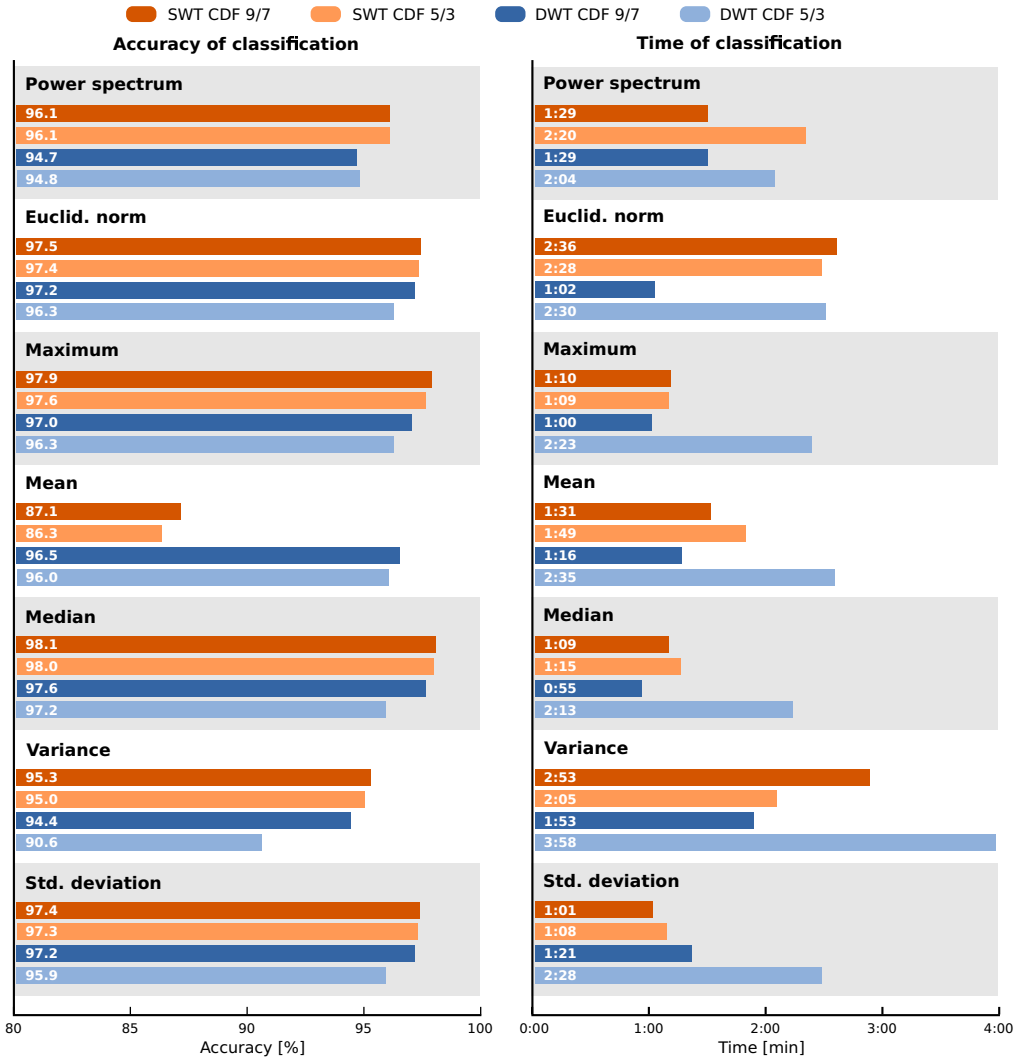

Figure 2. The accuracy and measured time of classification using different combinations of feature extraction methods.

5. Results

We present the results of classification using different combinations of feature extraction methods. The results are in Figure 2 which shows the accuracy and measured time for each tested combination of methods. The evaluation was performed on desktop PC equipped with AMD Athlon 64 X2 processor at 2.1 GHz.

From Figure 2 we can see that median has the best accuracy in all cases and the best processing time in case of using SWT with wavelet CDF 9/7. In overall accuracy, SWT outperforms DWT, and wavelet CDF 9/7 outperforms wavelet CDF 5/3. In overall processing time, DWT with wavelet CDF 9/7 has the best results.

6. Conclusion

Classification of the original data is very time-consuming. In Bromová et al. (2013) we showed that classification of data without feature extraction takes ~330 minutes with the accuracy 96.7 %. The proposed method reduces the number of attributes and the processing time to a small fraction and increases the accuracy in many cases.

In this paper, we describe the experiment with classification of spectra of Be stars using different feature extraction methods based on the wavelet transform in an attempt to identify the best method. From the results we can conclude that the best out of tested methods is SWT with wavelet CDF 9/7 followed by median as the aggregation function.

In future work, we will compare different classifiers and compare classification and clustering results. Based on this, we will try to find the best clustering model, and use it for clustering of any spectra and possibly to find new interesting candidates.

Acknowledgments. This work was 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), the national Technology Agency of the Czech Republic (TACR) project RODOS (no. TE01020155), and the project RVO:67985815.

References

Bařina, D., & Zemčík, P. 2010–2013, 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`

Bromová, P., et al. 2013, in Proceedings of conferences Datakon and Znalosti 2013 (VŠB-Technical University Ostrava), 95

Chang, C.-C., & Lin, C.-J. 2011, ACM Transactions on Intelligent Systems and Technology, 2, 27:1. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

Cohen, A., Daubechies, I., & Feauveau, J.-C. 1992, Communications on Pure and Applied Mathematics, 45, 485

Cortes, C., & Vapnik, V. 1995, Machine Learning, 20, 273

Liu, Y., San Liang, X., & Weisberg, R. H. 2007, Journal of Atmospheric and Oceanic Technology, 24, 2093