COMPARISON OF WAVELET BASED FEATURE EXTRACTION METHODS IN CLASSIFICATION OF EMISSION LINE STELLAR SPECTRA

Pavla Bromová
Doctoral Degree Programme (3), FIT BUT
E-mail: xbromo00@stud.fit.vutbr.cz

Supervised by: Petr Škoda, Jaroslav Zendulka
E-mail: skoda@sunstel.asu.cas.cz, zendulka@fit.vutbr.cz

Abstract: Our goal is the automatic detection of spectra of emission (Be) stars in large archives and classification of 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 computationally expensive. 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, wavelets, and statistical metrics in attempt to find the best method.

Keywords: Be star, stellar spectrum, emission line, feature extraction, wavelet transform, classification, support vector machines, SVM

1 INTRODUCTION

Our goal is to automatically detect spectra of emission stars (Be and B[e] types) 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 computationally expensive. 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 [2] we proposed a feature extraction method which reduces the number of attributes from $\sim$2000 to 10, can reduce the processing time from $\sim$330 minutes to $\sim$1 minute, and increase the accuracy from 96.7 % to 98.1 % at the same time. In this paper, we perform more experiments with feature extraction methods and their parameters in an attempt to find the best method and combination of parameters.

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 spectrograph of Ondřejov Observatory 2 m telescope. The dataset consists of 1565 spectra of Be and normal stars manually divided into 4 classes (178, 172, 1159, and 56 samples) based on the shape of the H_α line. The original spectrum contains approximately 2000 values around H_α line. Examples of spectra typical for individual categories are sketched in Figure 1.

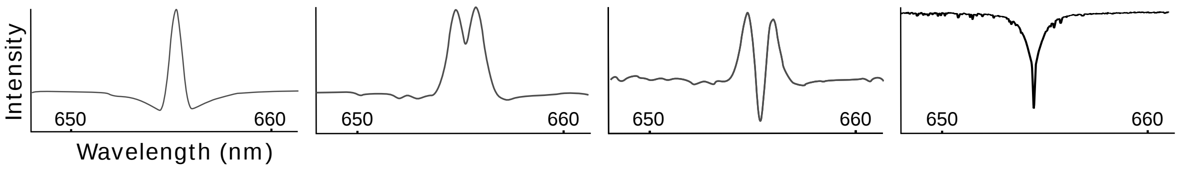

Figure 1: Examples of spectra typical for individual categories

3 FEATURE EXTRACTION

Centering First, the centers of emission lines are aligned to the center of samples, 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 of the sample.

Wavelet Transform The discrete (DWT) and stationary (SWT) wavelet transforms were employed for comparison, using the Cross-platform Discrete Wavelet Transform Library [1]. 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 [4]. 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 compared: wavelet power spectrum (WPS), Euclidean norm, maximum, mean, median, variance, standard deviation, skewness, and kurtosis.

The feature vector

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

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 [6] by division by corresponding scale.

4 CLASSIFICATION

Resulting feature vectors were classified using the support vector machines (SVM) [5] using the LIBSVM library [3]. The radial basis function (RBF) was used as a kernel function. There are two parameters for the 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.

5 RESULTS

We compare the average accuracy of classification using different parameters of feature extraction. There are three parameters: type of wavelet transform, type of wavelet, and aggregate function.

Two types of wavelet transform were used: discrete (DWT) and stationary (SWT); two types of wavelet: CDF 5/3 and CDF 9/7; and nine types of aggregate function: wavelet power spectrum, Euclidean norm, maximum, mean, median, variance, standard deviation, skewness, and kurtosis. More detailed description of parameters is in section 3.

All possible combinations of these parameters were used, resulting in 36 different feature vectors and 36 values of classification accuracy. The average accuracy for each value of each parameter was computed as the average from the accuracies for all feature vectors containing this parameter value (and all combinations of the other parameters). The results are in Table 1.

Parameter	Value	Average accuracy [%]
Wavelet transform	SWT	96.70
	DWT	96.06
Wavelet	CDF 9/7	96.97
	CDF 5/3	95.78
Aggregate function	Euclid. norm	98.40
	Std. deviation	98.31
	Maximum	98.18
	WPS	97.54
	Skewness	97.48
	Variance	95.47
	Kurtosis	95.34
	Mean	94.35
	Median	92.37

Table 1: The average classification accuracy for each value of each parameter of feature extraction.

Table 1 enables direct comparison of values of each parameter.

We can see that the difference between DWT and SWT is not significant, so it doesn't matter which transform we use regarding the accuracy of classification. However, regarding computational demands, SWT is more demanding. Thus, after this experiment we can say that it is more advantageous to use DWT.

The wavelet CDF 9/7 has slightly better result than CDF 5/3. There is no trade-off among different wavelets so we can claim CDF 9/7 to be more preferable.

There is quite big variance among the aggregate functions. We can say that first three of them (Euclid. norm, std. deviation, and maximum) will be among the most preferable, with the accuracy over 98% and very close values.

6 CONCLUSION

Classification of the original data is computationally expensive. In [2] we proposed a method that reduces the number of attributes and the processing time to a small fraction and increases the accuracy in many cases.

In this paper, we described the experiment with classification of spectra of Be stars using different feature extraction methods based on the wavelet transform in an attempt to find the best method. We compared different values of parameters of feature extraction and identified the best combination.

In future work, we will compare more feature extraction methods and different classifiers, and also results of classification and clustering. Based on this, we will try to find the best clustering model, use it for clustering of spectra in large archives, and possibly find some new interesting objects.

ACKNOWLEDGEMENT

This work was supported by the project CEZ MSM0021630528 Security-Oriented Research in Information Technology, the specific research grant FIT-S-11-2, the grant GACR 13-08195S of the Czech Science Foundation, and the project RVO:67985815.

REFERENCES

[1] D. Bařina and P. Zemčík. 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`, 2010–2013.

[2] P. Bromová, D. Bařina, P. Škoda, J. Vážný, and J. Zendulka. Classification of be stars using feature extraction based on discrete wavelet transform. In *Proceedings of conferences Datakon and Znalosti 2013*, pages 95–102. VŠB-Technical University Ostrava, 2013.

[3] 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`.

[4] 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.

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

[6] 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.