Wavelet Based Feature Extraction for Clustering of Be Stars

Pavla Bromová, Petr Škoda and Jaroslav Zendulka

Abstract The goal of our work is to create a feature extraction method for classification of Be stars. Be stars are characterized by prominent emission lines in their spectrum. We focus on the automated classification of Be stars based on typical shapes of their emission lines. We aim to design a reduced, specific set of features characterizing and discriminating the shapes of Be lines. In this paper, we present a feature extraction method based on the wavelet transform and its power spectrum. 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 kmeans algorithm. The results of our method are promising and encouraging to more detailed analysis.

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

Pavla Bromová
Faculty of Information Technology, Brno University of Technology, Božetěchova 1/2, 612 66 Brno, Czech Republic, e-mail: ibromova@fit.vutbr.cz

Petr Škoda
Astronomical Institute of the ASCR, v. v. i., Fričova 298, 251 65 Ondřejov, Czech Republic, e-mail: skoda@sunstel.asu.cas.cz

Jaroslav Zendulka
Faculty of Information Technology, Brno University of Technology, Božetěchova 1/2, 612 66 Brno, Czech Republic, e-mail: zendulka@fit.vutbr.cz

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 lines H_α in their spectrum [11]. 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.

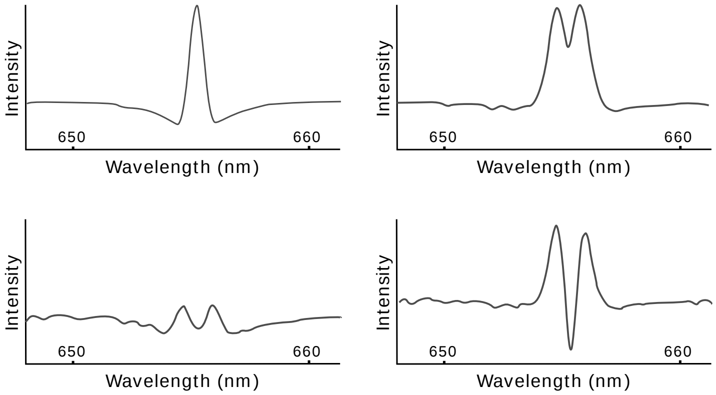

Fig. 1 Typical shapes of emission lines in spectra of Be stars

Our work is focused on the automated classification of Be stars based on typical shapes of their emission lines. There has not been much work on classification of Be stars. The only application found [2] is focused on a broader category of variable stars including pulsating Be stars. However, the method is not suitable for our goals, as it is applied on the whole spectrum where the local differences in the shapes of Be lines are lost. We need to zoom at the small part of a spectrum with the Be line and to design a reduced, specific set of features characterizing and discriminating the shapes of Be lines. Due to a large variety of shapes, it is not easy to construct simple criteria (like e.g. Gaussian fits) to identify the Be lines in an automatic manner.

In this paper, we present the feature extraction method based on the wavelet transform and its power spectrum. 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.

2 Method

The method is based on the wavelet transform and its power spectrum. A resulting feature vector is composed of two parts: 1. wavelet power spectrum, 2. value indicating the orientation of the H_α line (this information is lost in the wavelet power spectrum). The process of creating the feature vector is described here.

2.1 Wavelet Transform

Many different transforms are used in data processing (Fourier transform is perhaps the most widely used) [9]. The goal of these transformations is to obtain a sparse representation of data, and to pack most information into a small number of samples.

The wavelet transform consists in decomposing signals (data) into different frequency components. A wavelet is an elementary oscillatory waveform of a limited duration with an average value of zero. A signal is convolved with a scaled, shifted versions of a wavelet resulting in coefficients, which are a function of scale and position. Their structure leads to a fast computational algorithm. Wavelets are well localized and therefore suitable for revealing local transcient structures in data. Extensive literature exists on wavelets and their applications, e.g. [6, 1, 3, 7, 10].

2.1.1 Continuous Wavelet Transform

The continuous wavelet transform (CWT) is defined as the sum over all time of the signal multiplied by scaled, shifted versions of the wavelet function ψ:

$$C(s,p) = \int_{-\infty}^{\infty} f(t)\psi(s,p,t)dt,$$

where s is a scale of a wavelet and p is a position in the signal. The result of the CWT are wavelet coefficients C, which are a function of scale and position. The wavelet coefficients reflect the correlation between the wavelet and the data array. A larger absolute value of a coefficient implies a higher correlation.

There is a correspondence between wavelet scales and frequency as revealed by wavelet analysis: low scale → high frequency, high scale → low frequency.

In CWT (unlike DWT), it is possible to use every scale, and the analyzing wavelet is shifted smoothly over the full domain of the analyzed function.

2.1.2 Discrete Wavelet Transform

In the discrete wavelet transform (DWT), scales and positions are based on powers of two (dyadic scales and positions). An efficient way to implement this scheme using filters was developed yielding a fast wavelet transform.

The principle can be described as passing the original signal through two complementary filters – low-pass and high-pass [8]. This results in two signals, referred to as approximation and detail. The approximation is a high-scale, low-frequency component of the signal, the detail is a low-scale, high-frequency component. After each pass through filters, downsampling (removing every alternative coefficient) is performed in order to avoid doubling the amount of data.

The decomposition process can be iterated by splitting the approximation part of a signal as it still contains some details. So a signal is broken down into many lower-resolution components. The decomposition can proceed only until the individual details consist of a single sample. The wavelet transform then consists of approximation coefficients at last level and detail coefficients at all levels. For more details see e.g. [6].

2.2 Wavelet Power Spectrum

The wavelet power spectrum (WPS) is a useful way how to determine the distribution of energy within the signal [12]. By looking for regions of large power within WPS, we can determine which features of the signal are important.

The WPS at a particular decomposition level is calculated by summing up the squares of wavelet coefficients at that level [8]. For a set of wavelet coefficients $c_{j,k}$, where j is the level of decomposition and k is the order of the coefficient, WPS is given by:

$$wps(j) = \sum_{k=0}^{2^j-1} c_{j,k}^2$$

2.2.1 Rectification

It has been shown in [5] that wavelet power spectra are biased in favor of low frequencies. For example, for a signal comprising two sine waves of the same amplitude but distinct frequencies, a wavelet analysis will result in two spectral peaks of a different magnitude, the one on the low frequency being larger. This counters our expectation and is also in contrast to the result of any classical global analysis (such as Fourier transform), making comparison of the peaks across the scales impossible.

In [5], they established theoretically that the bias actually results from the traditional definition of energy for the wavelet power spectra which is not physically consistent. They present a physically consistent definition of energy: the transform coefficients squared divided by the scale they associate. The traditional biased power spectra are therefore easily rectified.

2.3 Normalization

WPS is normalized so that its total energy equals to 1, so it consists of percentages of energy corresponding to individual levels.

2.4 Orientation of Spectral Line

The information about the orientation of a spectral line is lost in the wavelet power spectrum, so we need to add it somehow 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. So far we have tried three values: 1, 0.1, and the amplitude of a spectral line, measured from the continuum of value 1.

3 Experiments

The experimental verification of the feature extraction method is performed using clustering. So far, the whole process has been implemented in Matlab, using its embedded algorithms. The stages are described in following sections.

3.1 Data Selection

We use spectra of Be stars from the archive of the Astronomical Institute of the Academy of Sciences of the Czech Republic. The spectra intended for spectral data mining are divided into 11 categories based on the shape of the H_α line. From them, 3 categories contains spectra of unstable stars, 1 category is composed of uncategorized (unknown) samples, 3 categories contains not enough samples for data mining, so only 4 categories are suitable for our experiments. The number of samples in these 4 categories are: 66, 150, 164, 276. We select at most 200 samples from each category, resulting in 656 samples in total.

From each spectrum, we select only a small part containing the emission line, so that the sample has 256 values and the emission line is approximately in the center. The number 256 was chosen according to the average width of the H_α line and according to DWT requirement on the length of the input data being a power of 2.

The spectra are normalized – lying on a continuum of a value 1.

3.2 Feature Extraction

The feature extraction method is described in the previous chapter. We use both continuous and discrete wavelet transform. As we do not have any reference method for Be stars for comparison, we compare our results with a common way of feature extraction from time series using wavelets – keeping N largest coefficients of wavelet transform [4].

From resulting coefficients of the wavelet transform we create different kinds of feature vectors which are used for comparison in experiments:

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 a discrete wavelet power spectrum, 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**: continuous wavelet power spectrum (normalized), CWT performed with 16 scales. Same orientation as in the previous cases with DWPS.
10. **CWPS 8 + orientation 1**: continuous wavelet power spectrum (normalized), CWT performed with 8 scales. Same orientation as in the previous case.

In experiments up to now, we have used ”symlet 4” wavelet. In DWT, the maximum possible level of decomposition = 5 has been used.

3.3 Clustering

So far, the k-means algorithm in Matlab has been used for clustering. Squared Euclidean distance is used as a distance measure. Clustering is repeated 30 times, each with a new set of initial cluster centroid positions. Kmeans 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.

4 Results

Fig. 2 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.

These results are promising and indicate the continuous wavelet transform being more suitable for our purpose. This encourages to focus on the CWT and perform more experiments, as we have focused more on the DWT up to now.

5 Conclusion

The goal of our work is to create a feature extraction method for classification of Be stars based on typical shapes of their emission lines. We aim to design a reduced, specific set of features characterizing and discriminating the shapes of the H_α emission lines.

In this paper, we have analysed the capabilities of using wavelet power spectrum for classification of spectra of Be stars. We have presented a feature extraction method based on the wavelet transform and its power spectrum. Different feature

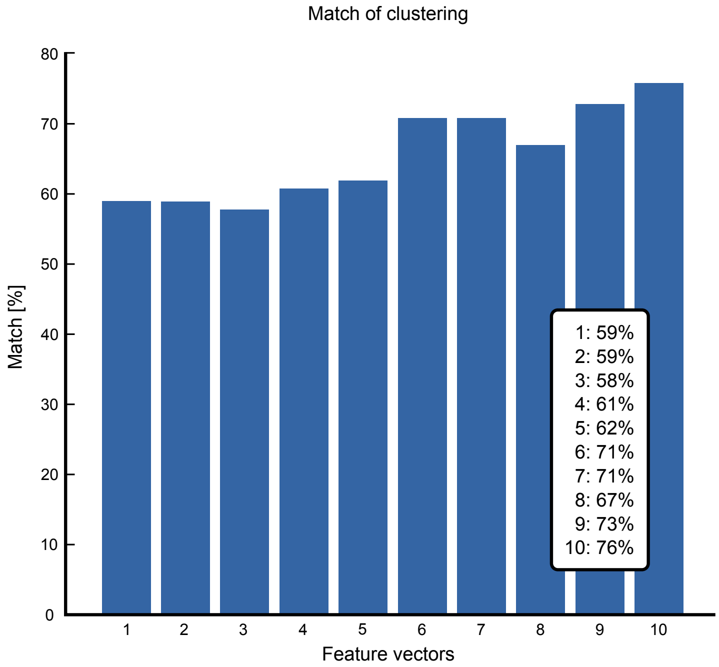

Fig. 2 The match of the clustering using different feature vectors

vectors have been created and compared on clustering of Be stars spectra. The evaluation was performed utilizing our knowledge of ideal classification of spectra based on a manual categorizing.

The best results are given by the feature extraction method based on the CWT. Results are promising and indicate the CWT is more suitable for our purpose than DWT, which encourages to focus on the CWT and perform more detailed analysis.

Acknowledgement

This work was supported by the specific research grant FIT-S-11-2.

References

1. I. Daubechies. *Ten lectures on wavelets*. CBMS-NSF regional conference series in applied mathematics. Society for Industrial and Applied Mathematics, 1994.
2. J. Debosscher. *Automated Classification of variable stars: Application to the OGLE and CoRoT databases*. PhD thesis, Institute of Astronomy, Faculty of Sciences, Catholic University of Leuven, 2009.

3. G. Kaiser. *A friendly guide to wavelets*. Birkhäuser, 1994.
4. 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.
5. 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.
6. S. Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, 3rd edition, 2008.
7. Y. Meyer and D.H. Salinger. *Wavelets and Operators*. Number sv. 1 in Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1995.
8. S. Prabakaran, R. Sahu, and S. Verma. Feature selection using haar wavelet power spectrum. *BMC Bioinformatics*, 7:432, 2006.
9. J.L. Starck and F. Murtagh. *Astronomical image and data analysis*. Astronomy and astrophysics library. Springer, 2006.
10. G. Strang and T. Nguyen. *Wavelets and filter banks*. Wellesley-Cambridge Press, 1996.
11. O. Thizy. Classical Be Stars High Resolution Spectroscopy. *Society for Astronomical Sciences Annual Symposium*, 27:49, 2008.
12. C. Torrence and G. P. Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79:61–78, 1998.