Feature Extraction Using Wavelet Power Spectrum for Stellar Spectra Clustering

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

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

Abstract. This paper analyses the capabilities of using wavelet power spectrum for clustering of Be-type stars spectra. We propose a method using discrete wavelet transform for feature extraction and the wavelet power spectrum as a feature vector. We also propose a modification of this method and compare them. We analyse the methods in the clustering of artificial stellar spectra and compare them with a traditional method of wavelet-based feature extraction – keeping k largest coefficients. The results show that the correctness of clustering of our method is significantly better than in the case of a traditional method. We also compare the effect of using different type of wavelet and level of decomposition.

Keywords: be star, stellar spectrum, feature extraction, discrete wavelet transform, wavelet power spectrum, clustering, classification

1 Introduction

Nowadays, astronomy is facing an exponentially growing amount of data due to the evolution of detectors, telescopes and space instruments [1, 2]. Petabytes of data are expected to flow from massive digital sky surveys in the next decade, being stored in the world-wide network of distributed archives. The effective retrieval of knowledge from these massive distributed databases requires new automated approaches of knowledge discovery in databases based on machine learning methods.

The aim of this paper is to analyse the feature extraction method using wavelet power spectrum for automated clustering of simulated spectra of Be stars, which will be further used for classification of real spectra. It seems that wavelets have not been used this way yet in astronomy, although they have been successfully applied in several other domains, mainly on medical data (classification or detection of a disease or an event from EEG/ECG signals [7, 14, 15, 21]). The paper also compares the effect of using different type of wavelet and level of decomposition on the results of clustering. The paper extends the results published in [3].

2 Background

2.1 Classification

In data mining, classification refers to assigning a data item into one of several predefined classes [6]. The piece of input data is represented by a set of characteristics (features), which is usually obtained from the original data by feature extraction.

2.2 Clustering

Clustering refers to assigning a set of objects into groups (clusters) so that the objects in the same cluster are more similar (based on some similarity measure) to each other than to those in other clusters [13].

The accuracy of clustering can be evaluated with the silhouette method [16]. This technique provides an information of how well each object lies within its cluster. This measure ranges from +1, indicating points that are very distant from neighboring clusters, through 0, indicating points that are not distinctly in one cluster or another, to -1, indicating points that are probably assigned to the wrong cluster. The average silhouette value of the entire dataset is a measure of how appropriately the data has been clustered.

2.3 Feature Extraction

Real world data sets are usually not directly suitable for performing data-mining algorithms [9]. They may contain noise, missing values, and usually are too large and high-dimensional. One of the methods of dimensionality reduction is feature extraction. It consists in transforming the input data into a reduced representation set of features known as feature vector. One of popular feature extraction techniques used for signals is wavelet transform.

2.4 Wavelet Transform

The wavelet transform consists in partitioning data (signals) into different frequency components [9]. One major advantage of wavelets is the ability to analyze a local area of a signal [15]. Wavelet analysis is capable of revealing aspects of data that other signal analysis techniques miss, such as trends, breakdown points, or discontinuities. Wavelet transforms have gained popularity in all areas of signal processing and they have also been extensively used in astronomical data analysis during the last fifteen years [17]. A lot of literature can be found about wavelets, e.g. [5, 8, 12, 11, 18].

Discrete Wavelet Transform (DWT) The principle of the DWT consists in passing the original signal through two complementary filters – low-pass and high-pass [15]. 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. This can be repeated so long until we are satisfied with the resolution of components we have created. The wavelet transform of data at a level i of decomposition consists of approximation coefficients at i-th level and all detail coefficients up to i-th level, resulting in *number of levels + 1* coefficient bands. The wavelet coefficients reflect the correlation between the wavelet (at a certain scale) and the data array (at a particular location). A larger absolute value of a coefficient implies a higher correlation.

Wavelet-Based Feature Extraction Common ways of feature extraction from time series using wavelets are [9]:

- keeping the first k coefficients – in this case each time series is represented by a rough sketch, because these coefficients correspond to the low frequencies of the signal
- keeping k largest coefficients – this achieves more accurate representation of the signal

The rest of the signal is approximated with zeros.

3 Data

Be stars are hot, rapidly rotating B-type stars with equatorial gaseous disk producing prominent emission lines in their photospheric spectrum [19, 20]. Be stars show a number of different shapes of emission lines, like double-peaked profiles with or without narrow absorption, or single peak profiles with various deformations, as we can see in Fig.1.

The analysis of the method is performed on simulated spectra generated by computer. A collection of 1000 spectra has been created trying to cover as many emission lines shapes as possible. Each spectrum is 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 is then convolved with a gaussian function, which simulates an appropriate resolution of the spectrograph.

The source of real data is the archive of the Astronomical Institute of the Academy of Sciences of the Czech Republic.

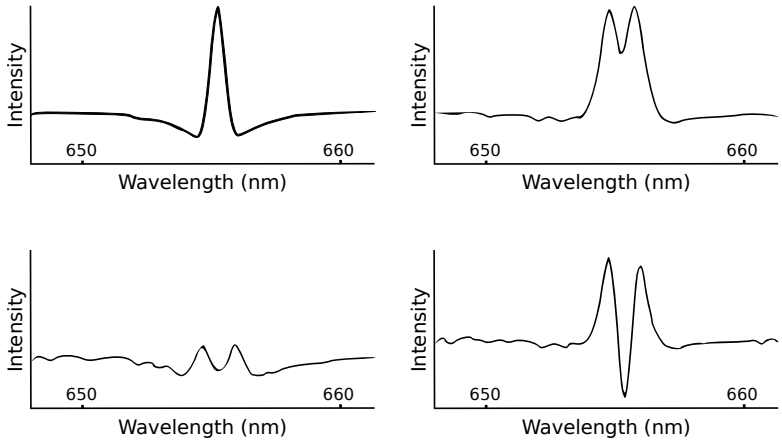

Fig. 1: Different shapes of emission lines [4]

4 Method

We propose a feature extraction method using the wavelet power spectrum (WPS) for stellar spectra clustering. The WPS is a useful way how to determine the distribution of energy within the signal [10]. 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 [15]. 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:

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

A disadvantage of WPS is that the information about the positive/negative direction of the peak in the spectrum is lost, as results from its definition, so it doesn't distinguish spectra with the same shape of the peak but the opposite direction. Therefore we propose a modified version of WPS – WPSD (WPS keeping Direction) which retains this information. WPSD is defined as

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

where variables have the same meaning as for WPS. An example of WPS and WPSD of a simulated spectrum is in Fig.2.

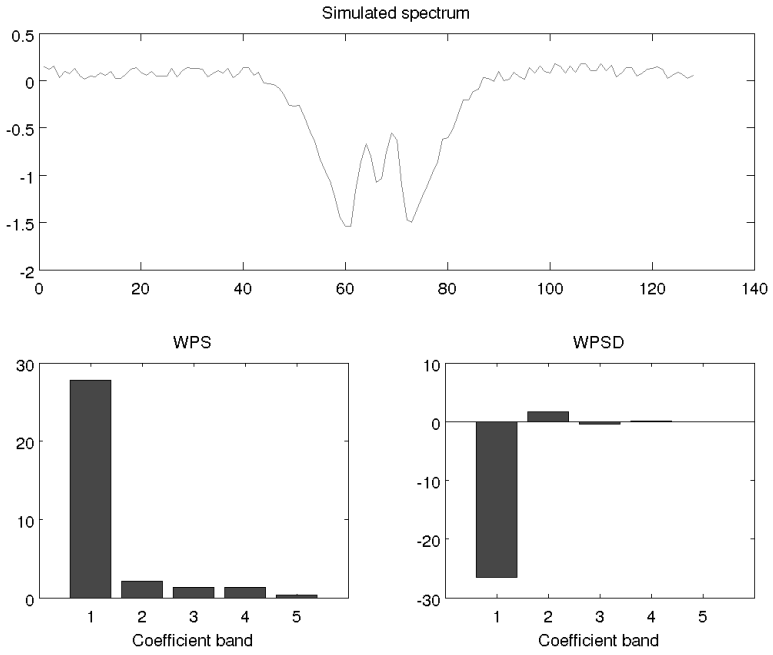

Fig. 2: An example of WPS and WPSD of a simulated spectrum. Number of coefficient bands = level of decomposition + 1 (see DWT in 2.4).

5 Experiments

At first, the discrete wavelet transform of simulated spectra is performed. In DWT, the type of wavelet and the level of decomposition must be determined. We perform experiments comparing the effect of different values of these parameters on the results of clustering and choose the parameters with the best results for the final comparison of feature extraction methods. So, there are 3 experiments – comparison of clustering results depending on different values of 3 parameters:

- level of decomposition
- type of wavelet
- feature extraction method

Clustering is performed using k-means algorithm into 3-10 clusters and the silhouette method is used for the evaluation. Different number of iterations of the clustering process is used in experiments and the average silhouette values are presented as the results.

5.1 Level of decomposition

Tested levels of decomposition were from 2 to 5. The simplest Haar (db1) wavelet was used, the number of iterations of clustering was 50.

On Fig. 3 we can see that the results of different levels differ only in hundredths of unit, showing that more values in WPS do not necessarily contribute to better results and 2 levels are sufficient.

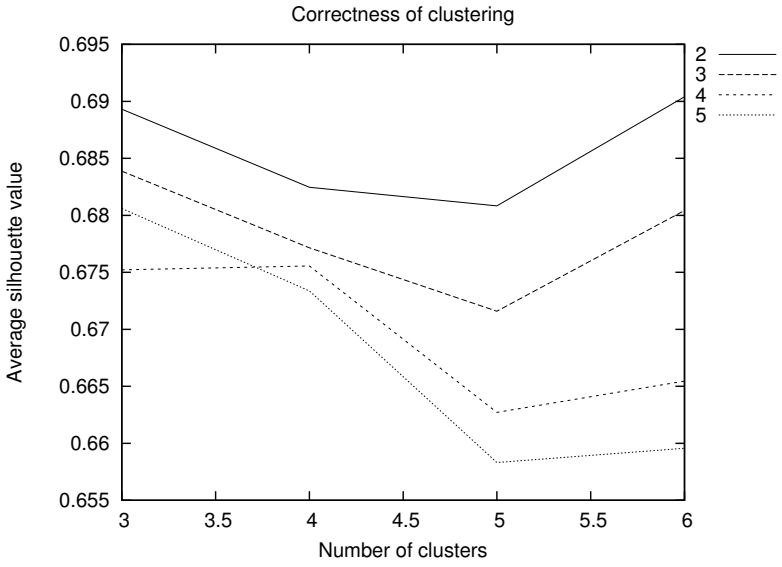

Fig. 3: The effect of different level of decomposition on the results of clustering

5.2 Type of wavelet

We tried to find the wavelet best describing the character of our data, based on its similarity with the shape of emission lines. We have chosen from the set of wavelets available for DWT in Matlab, i.e. daubechies, symlets, coiflets, biorthogonal, and reverse biorthogonal wavelets family. Two representative wavelets from each family were chosen and db1 (or Haar) wavelet was used for comparison as the simplest wavelet. The list of tested wavelets:

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

Based on the previous results, the level of decomposition was set to 2. The number of iterations of clustering was 50.

On Fig. 4 we can see that there are minimal differences between wavelets (hundredths of unit), which suggests that the type of wavelet has not big effect on the clustering results.

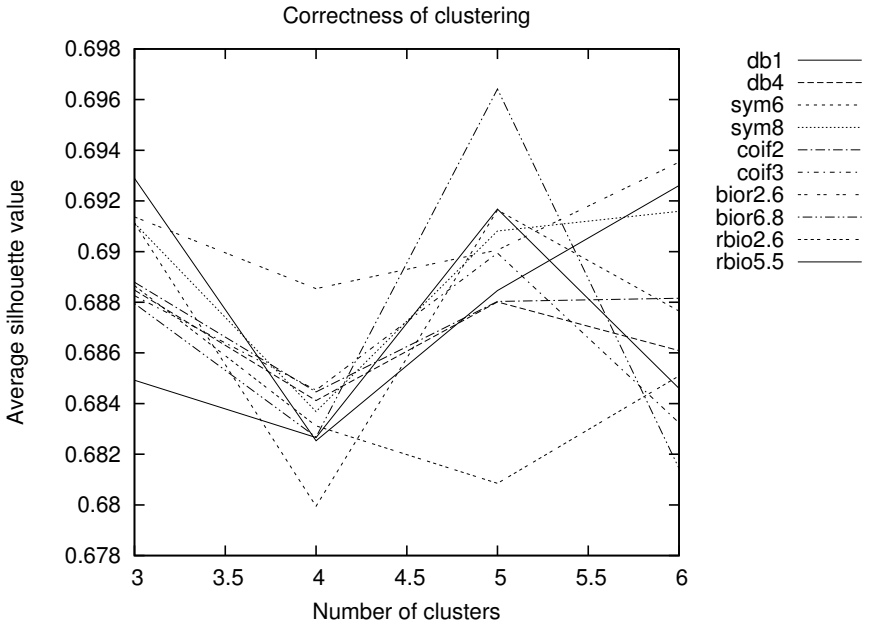

Fig. 4: The effect of different type of wavelet on the results of clustering

5.3 Feature extraction method

Finally, different types of feature vector were created. For comparison of our method we use the traditional method of wavelet-based feature extraction – keeping k largest coefficients. We use various values of k for better comparison. The list of tested feature vectors:

- WPS
- WPSD
- keeping 5 largest coefficients
- keeping 20 largest coefficients
- keeping 50 largest coefficients

Based on the previous results, the simplest Haar wavelet and 2 levels of decomposition were used in DWT. Clustering was performed in 30 iterations.

From Fig. 5 we can see that there are almost no differences between WPS and WPSD, which suggests that the information about the direction of the peak is not significant in this method as it could seem to. We can also notice differences between the results of keeping k largest coefficients depending on different k. But mainly the graph shows that the results of our method are in all cases significantly better than in the case of keeping k largest coefficients.

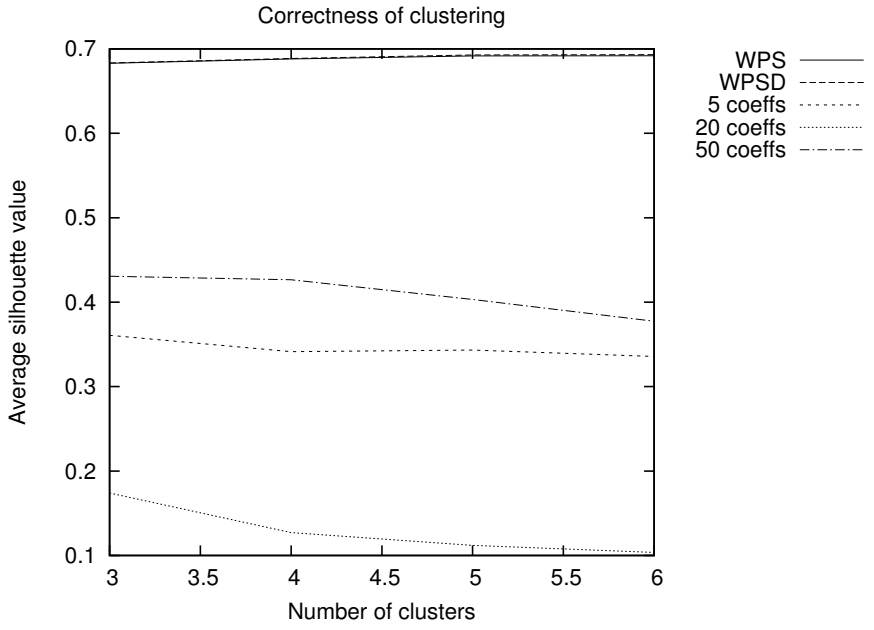

Fig. 5: The effect of different feature extraction method on the results of clustering

6 Conclusion

In this paper, we have analysed the capabilities of using wavelet power spectrum for clustering of spectra of Be stars. We have proposed a feature extraction method using WPS and also proposed and tested its modified version. The methods have been applied to clustering of artificial spectra and compared with a traditional method of keeping k largest coefficients. The results show that differences between the two variants of our method are neglecting and in both cases the results are significantly better than in the case of the method of keeping k largest coefficients.

We have also performed experiments comparing the effect of different type of wavelet and level of decomposition on the results of clustering. Experiments

show that the differences in the results are neglecting, thus suggesting that the choice of these parameters has not significant effect on the results of clustering.

In the next step, the feature extraction method will be optimized for application on large-scale data and applied to the classification of the real spectra. Currently, neural network based classification is assumed.

Acknowledgement

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

References

1. N. M. Ball and R. J. Brunner. Data Mining and Machine Learning in Astronomy. *International Journal of Modern Physics*, D19:1049–1106, 2010.
2. M. Brescia, G. Longo, and F. Pasian. Mining Knowledge in Astrophysical Massive Data Sets. *Nuclear Instruments and Methods in Physics Research*, A623:845–849, 2010.
3. Pavla Bromová. Stellar spectra classification using wavelet power spectrum. In *Proceedings of the 18th Conference STUDENT EEICT 2012 Volume 3*, pages 366–370. Brno University of Technology, 2012.
4. Christian Buil. The spectroscopic be stars atlas. http://www.astrosurf.com/buil/us/bestar.htm.
5. I. Daubechies. *Ten lectures on wavelets*. CBMS-NSF regional conference series in applied mathematics. Society for Industrial and Applied Mathematics, 1994.
6. U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996.
7. P. Jahankhani, K. Revett, and V. Kodogiannis. Data mining an eeg dataset with an emphasis on dimensionality reduction. In *CIDM*, pages 405–412. IEEE, 2007.
8. G. Kaiser. *A friendly guide to wavelets*. Birkhäuser, 1994.
9. 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.
10. 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.
11. S.G. Mallat. *A wavelet tour of signal processing*. Wavelet Analysis and Its Applications Series. Academic Press, 1999.
12. Y. Meyer and D.H. Salinger. *Wavelets and Operators*. Number sv. 1 in Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1995.
13. B.G. Mirkin. *Mathematical classification and clustering*. Nonconvex optimization and its applications. Kluwer Academic Pubishers, 1996.
14. M. Murugappan, M. Rizon, R. Nagarajan, and S. Yaacob. Fcm clustering of human emotions using wavelet based features from eeg. *Biomedical Soft Computing and Human Sciences*, 14(2):35–40, 2009.
15. S. Prabakaran, R. Sahu, and S. Verma. Feature selection using haar wavelet power spectrum. *BMC Bioinformatics*, 7:432, 2006.
16. 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.

17. J.L. Starck and F. Murtagh. *Astronomical image and data analysis*. Astronomy and astrophysics library. Springer, 2006.
18. G. Strang and T. Nguyen. *Wavelets and filter banks*. Wellesley-Cambridge Press, 1996.
19. O. Thizy. Classical Be Stars High Resolution Spectroscopy. *Society for Astronomical Sciences Annual Symposium*, 27:49, 2008.
20. J. Vážný. Virtual observatory and data mining. diploma thesis, Department of Theoretical Physics and Astrophysics, Masaryk University, Brno, Czech Republic, 2011.
21. W. Zhao and C. E. Davis. Swarm intelligence based wavelet coefficient feature selection for mass spectral classification: An application to proteomics data. *Analytica Chimica Acta*, 651(1):15 – 23, 2009.