STELLAR SPECTRA CLASSIFICATION USING WAVELET POWER SPECTRUM

Pavla Bromová
Doctoral Degree Programme (1), 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: This paper analyses the capabilities of using wavelet power spectrum for classification of spectra of Be stars. We propose a method of using the wavelet power spectrum as a feature vector and apply it to clustering of artificial spectra. The method is compared with a traditional approach of feature extraction and the results of our method are significantly better.

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

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 capabilities of using wavelet power spectrum for automated classification of spectra of Be stars. 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 [6, 10, 11]).

2 BACKGROUND

2.1 CLASSIFICATION

In data mining, classification refers to assigning a data item into one of several predefined classes [5]. 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 [9].

The accuracy of clustering can be evaluated with the silhouette method [12]. 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 [7]. 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 [7]. One major advantage of wavelets is the ability to analyze a local area of a signal [11]. 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 [13]. A lot of literature can be found about wavelets, e.g. [4].

Discrete Wavelet Transform (DWT) The principle of the DWT consists in passing the original signal through two complementary filters – low-pass and high-pass [11]. 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 (1) 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) and (2) keeping the k largest coefficients (this achieves more accurate representation of the signal) [7]. The rest of the signal is approximated with zeros.

3 DATA

The source of the data will be the archive of the Astronomical Institute of the Academy of Sciences of the Czech Republic, which has a long-term experience with the research of Be stars.

Be stars are hot, rapidly rotating B-type stars with equatorial gaseous disk producing prominent emission lines in their photospheric spectrum [14]. 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. Each line in the graph corresponds to one spectrum.

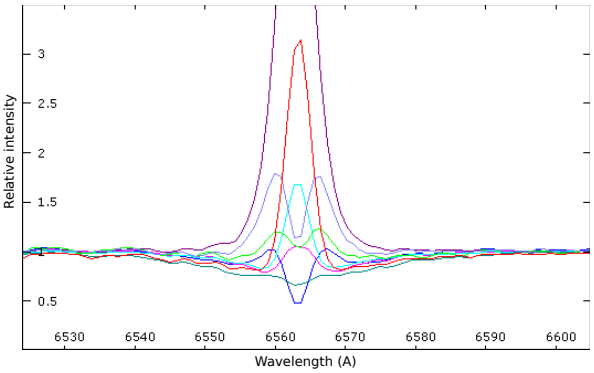

Figure 1: Different shapes of emission lines [3]

4 METHOD

We propose a method of using the wavelet power spectrum (WPS) of a stellar spectrum as a feature vector for classification. The WPS is a useful way how to determine the distribution of energy within the signal [8]. 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 [11]. 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.

For comparison we use the traditional approach of keeping k largest coefficients with $k = 20$. The analysis of WPS potential is performed on clustering of simulated spectra generated by computer. The complete process of the analysis is described in following steps:

1. **Simulated spectra generation**: 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.

2. **Feature extraction**: At first, the discrete wavelet transform of simulated spectra is performed using the Haar wavelet and 4 levels of decomposition. Then the WPS and WPSD are calculated and 3 types of feature vector are created: (1) WPS, (2) WPSD, and (3) keeping k largest coefficients with $k = 20$.

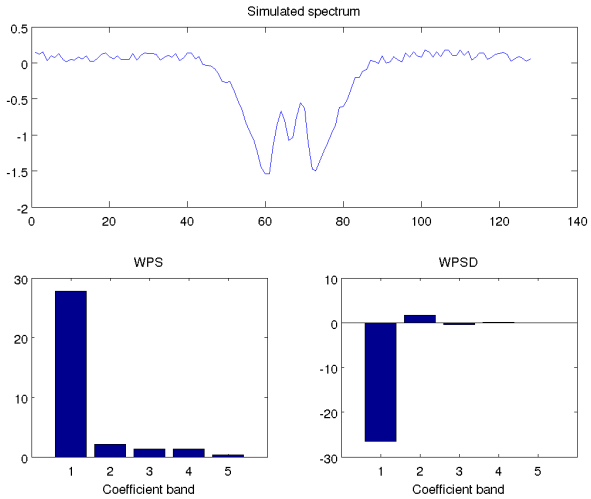

Figure 2: An example of WPS and WPSD of a simulated spectrum. Number of coefficient bands = number of decomposition levels + 1 (see DWT).

3. **Clustering**: Clustering is performed using k-means algorithm into 2-10 clusters and the silhouette method is used for the evaluation. This process is repeated 50-times and the average silhouette values are presented as the results.

4. **Classification**: The feature extraction method with the best results is optimized for using on large-scale data and applied to the classification of the real spectra. Currently, neural network based classification is assumed.

5 RESULTS

The results of the clustering step are presented here. Figure 3 shows a dependency of the correctness of clustering (average silhouette value) on the number of clusters for different types of feature vector. We can see that using WPS and WPSD has very similar results for most numbers of clusters and in both cases the results are significantly better than using k largest coefficients.

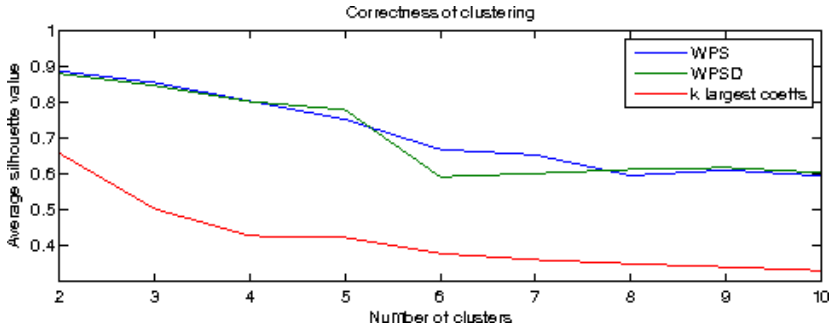

Figure 3: Correctness of clustering for different types of feature vector

6 CONCLUSION

In this paper, we have analysed the capabilities of using wavelet power spectrum for classification of spectra of Be stars. We have proposed a method of using the WPS as a feature vector and also proposed and examined a modified version of WPS. The method has been applied to clustering of

artificial spectra and compared with a traditional approach of keeping k largest coefficients. The results show that using WPS (both the original and modified version) is significantly better.

The feature extraction method may be further enhanced by a selection of an optimal wavelet type and a number of wavelet decomposition levels, which determines the length of WPS. In the final step, the wavelet power spectrum will be applied in the classification of the real spectra.

ACKNOWLEDGEMENT

This work was partially supported by the research plan MSM0021630528, the specific research grant FIT-S-11-2 and the IT4Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070.

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] Christian Buil. The spectroscopic be stars atlas. `http://www.astrosurf.com/buil/us/bestar.htm`.

[4] I. Daubechies. *Ten lectures on wavelets*. CBMS-NSF regional conference series in applied mathematics. Society for Industrial and Applied Mathematics, 1994.

[5] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996.

[6] 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.

[7] 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.

[8] 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.

[9] B.G. Mirkin. *Mathematical classification and clustering*. Nonconvex optimization and its applications. Kluwer Academic Pubishers, 1996.

[10] 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.

[11] S. Prabakaran, R. Sahu, and S. Verma. Feature selection using haar wavelet power spectrum. *BMC Bioinformatics*, 7:432, 2006.

[12] 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.

[13] J.L. Starck and F. Murtagh. *Astronomical image and data analysis*. Astronomy and astrophysics library. Springer, 2006.

[14] O. Thizy. Classical Be Stars High Resolution Spectroscopy. *Society for Astronomical Sciences Annual Symposium*, 27:49, 2008.