Classifying Mass Spectral Data Using SVM and Wavelet-Based Feature Extraction

Wong Liyen[1], Maybin K. Muyeba[1], John A. Keane[2], Zhiguo Gong[3], and Valerie Edwards-Jones[4]

[1] School of Computing, Mathematics and Digital Technology
[2] School of Computer Science, University of Manchester, UK
[3] Faculty of Science and Technology
University of Macau, China
[4] Institute for Biomedical Research into Human Movement and Health
Manchester Metropolitan University, UK
li.y.wong@stu.mmu.ac.uk, {m.muyeba,v.e.jones}@mmu.ac.uk,
jak@cs.man.ac.uk, fstzgg@umac.mo

Abstract. The paper investigates the use of support vector machines (SVM) in classifying Matrix-Assisted Laser Desorption Ionisation (MALDI) Time Of Flight (TOF) mass spectra. MALDI-TOF screening is a simple and useful technique for rapidly identifying microorganisms and classifying them into specific subtypes. MALDI-TOF data presents data analysis challenges due to its complexity and inherent data uncertainties. In addition, there are usually large mass ranges within which to identify the spectra and this may pose problems in classification. To deal with this problem, we use Wavelets to select relevant and localized features. We then search for best optimal parameters to choose an SVM kernel and apply the SVM classifier. We compare classification accuracy and dimensionality reduction between the SVM classifier and the SVM classifier with wavelet-based feature extraction. Results show that wavelet-based feature extraction improved classification accuracy by at least 10%, feature reduction by 76% and runtime by over 80%.

Keywords: SVM, wavelets, MALDI-TOF, parameter search, feature reduction.

1 Introduction

Signal data is a sequence of measurements from instruments that is either continuous or discrete and captured in intervals of time, frequency, distance, wave numbers etc. A signal of particular interest is one that is absorbed or reflected and usually measured in wavelength intervals. In recent years, there have been a number of studies on bacterial diseases and the problem of identifying species of bacteria that cause particular diseases. In particular, when signals are projected on bacterial samples, resulting ions from the compound are allowed to drift (time of flight) towards a detector. The time of flight is measured and is proportional to their mass. This data is called Matrix Assisted Laser Desorption Ionisation (MALDI) Time Of Flight (TOF) [1].

T. Yoshida et al. (Eds.): AMT 2013, LNCS 8210, pp. 413–422, 2013.

The number (or count) of these ions is then plotted against their mass and a spectral graph is produced. This graph shows peaks and troughs of the properties the bacterial species exhibits and is useful for identifying isolates - species, genres, types etc [2][4][6]. The signal data has lots of features, is large and has missing values. These problems motivate the approach used in this paper. Firstly, support vector machines (SVM) [22] have powerful generalisation ability for high dimensional data with missing values. Secondly, feature reduction has been known to improve classification accuracy and wavelets are a favourable choice in signal processing [5]. Wavelets are mathematical tool for decomposing data and complex functions into time and frequency components [26]. Unlike Fourier transform, wavelet transform are better suited for non-stationary signals such as MALDI-TOF mass spectra as they can distinguish different frequency signals at different times (non-stationery).

Signal classification [5] typically has two steps: feature extraction and signal classification on the reduced feature set. Our experiments are based on classification with full features using SVM, compared to classification with reduced features sets (SVM and wavelets) whilst choosing a suitable kernel classifier [25]. Some earlier work regarding initial experimentation and data mining methodology used are given in [28]

The paper is organised as follows: section 2 presents wavelets; section 3 presents mining signal data; section 4 presents experimentation and section 5 a conclusion.

2 Wavelets

Wavelets are a set of mathematical functions used to approximate data and more complex functions by dividing a signal into different frequency and time intervals called wavelets [8]. These intervals are better represented to their scales. Wavelets express a given function in terms of summation of basis functions. The wavelet basis is formed by translation and dilation of the mother wavelet. An example a mother wavelet is a Haar wavelet (fig. 1).

$$\mathrm{H(t)}= \begin{cases} 1, & 0<t<0.5 \\ -1, & 0.5<t<1 \\ 0, & \text{eslewhere} \end{cases}$$

Fig. 1. Haar wavelet

The wavelet transform is performed on a continuous function, $f(t)$, and defined as

$$f(t) = \int_{-\infty}^{+\infty} h(\omega)\psi_{\omega}(t)dt \qquad (1)$$

where $h(\omega)$ is a weighting function and $\psi_{\omega}(t)$ is an othonorrnal basis function such that $\psi(2^{j}t-k), j,k \in Z$. By dilation and translation of the mother wavelet, we get wavelets compactly supported in their regions [9][10]. Wavelets exhibit other useful properties in addition to dilation and localization, such as smoothness, distinguishing most essential information, feature selection etc and their efficiency makes them candidates for data mining. Wavelets are designed to give excellent time resolution at high frequencies (i.e. for short durations of time at these high frequencies) and poor frequency resolution, and good frequency resolution at low frequencies and poor time resolution. Mass spectra contain noise because of contaminants and matrix material, causing varying baselines [17]. That is, to start preprocessing the data, a baseline correction is needed to remove low-frequency noise. MALDI-TOF MS spectra is recorded in signal form as (mass-to-charge ratio, millivolt signal, see figure 2), the second value shows the strength of the signal. The signal exhibits elongated (or outstanding) features above the baseline noise level and usually unevenly distributed. Feature selection (or removing noise level data) mostly focuses on selecting peaks [16] that are higher than a predetermined signal noise threshold to facilitate biomarker identification [11][12][18]. Biomarkers are measures that indicate normal biological processes, pathogenic (or organism) processes or other pharmacological responses to some therapy. Wavelets are well adapted to removing irrelevant noise level data features (Denoising [14]), sometimes termed smoothing.

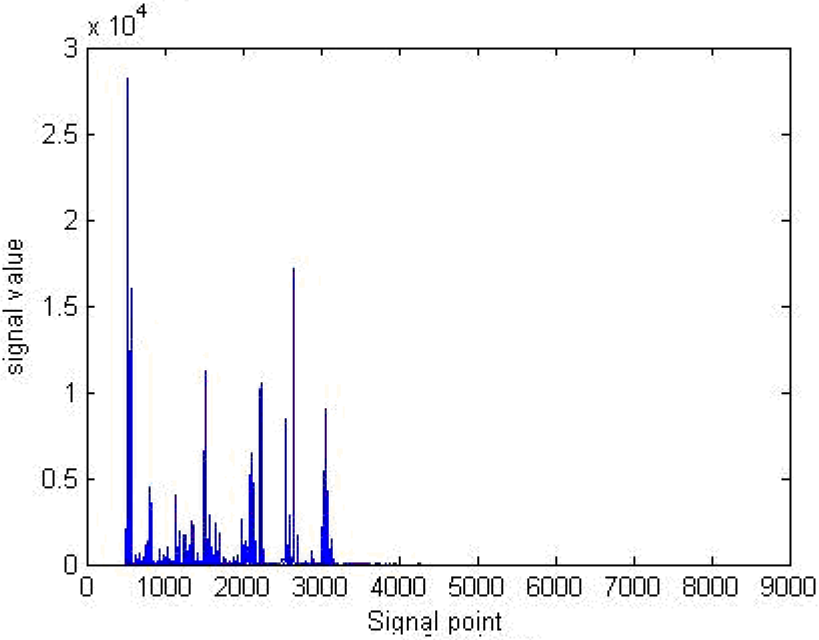

Fig. 2. MALDI-TOF Spectral data

Generally, most MALDI-TOF approaches aim to extract and quantify graph peak features accurately as these are considered to be the most interesting [7]. Other approaches use ant colony optimization to help efficiently select a set of interacting variables (features) by use of heuristic functions [16]. In addition, there are automated techniques for rapidly differentiating similarity between strains of bacteria, and in particular their taxonomic characterisation [2]. Feature selection is mainly concerned with obtaining useful features without loss of information, transforming an m-dimensional domain to a k-dimensional $k << m$.

Two approaches are common:

(1) Keep the largest k-coefficients, approximate rest to zero
(2) Keep the first k-coefficients, approximate rest to zero

Despite these two approaches, there is no guarantee the most important features will be obtained because there might be issues with information granulation (keep some, ignore the rest). In [5] fuzzy wavelet packets are introduced to deal with granulation of signal data into fuzzy relations (or clusters and not fuzzy sets) and reduce feature dimensionality by a fuzzy c-means approach. Information granulation by fuzzy means was presented in [15] and other fuzzy wavelet packet based feature extractions are reported in [19].

3 Mining Signal Data

Mining signal data is not new and various works exist [3]. Recently, machine learning approaches have been applied to learning MALDI-TOF data, for example use of Support Vector Machines (SVMs) and other techniques [4]. SVMs are popular classifiers that learn by examples and assign labels to objects [23][24]. They can be used for classification and regression as well as other learning tasks. SVM classifies linearly separate data by separating two clusters of data with the optimal hyperplane. The first problem, however, is that real world data is often non-linearly separable. Kernel functions provide a solution by projecting data into a higher dimensional feature space to separate by hyperplane. SVMs have been successfully applied to an increasingly wide variety of biomedical applications, for example microarray gene expression [24].

Secondly, another major problem in classifying features from signal data is handling the dimensionality problem: mapping the reduced data into a space and then classifying the result. In [5] wavelet packets are used to extract features (data/feature issues), classify and rank them (pattern evaluation) using a linear discriminant function (LDF) and others [13]. The approach is then, for a c-class problem with N labeled signal classes $X = \{(x_k, \omega_k), k = 1,2,..,N\}$, $x_k \in \Re^n$ and $\omega_k \in C, C = \{1,2.,c\}$, a feature extraction function i.e. a mapping $f : X \to X'$ where $m << n$ and $X' \subseteq \Re^m$ is the reduced feature space.

To classify the feature space, we find a classifier to map the feature space into the known class labels $g : X' \rightarrow C$. To retain features that are a more accurate representation of the space for classification, the use of fuzzy clusters becomes necessary because of uncertainties in data granulation of the signal data.

To measure pattern extraction, classification metrics such as classification accuracy (rate) is applied to a number of features (or principal components) under varying discriminatory thresholds, r , for example in the fuzzy case where $0 < r < 1$ [5][19].

4 Experiments

Preliminary experiments have been done using high dimensional MALDI-TOF spectral bacteria data provided by the Medical Microbiology Department, Manchester Metropolitan University. The data consisted of 14461 features with 2 columns: the first column being mass/charge ratio and the second column being a millivolt signal also known as strength of signal. There were 14 classes of different strains of S. aureus bacteria with two (2) testing samples each i.e. total of 28 testing data. We used the LIBSVM library [21] for classification.

4.1 Data Pre-processing

The procedure for the experiment was as follows:

(1) Split data into training and testing sets
(2) Perform numerical scaling – prevent dominance in ranges
(3) Perform a Grid Search for best kernel parameters, radial (r) and degree(d)
Parameters with best cross-validation accuracy are chosen for classifier training
(4) Predict the test data with the trained classifier

The basic procedure above is further extended with the case where we perform feature extraction with discrete wavelet transform (DWT) – denoising and decomposition with a thresholding method that only discards the portion of the data that exceeds a certain limit. Further, the signal data is then decomposed into two subsequences – the approximation coefficient and the detail coefficient. The approximation coefficients contain most of the important information (peaks of the data) and are capable of describing the underlying data characteristics [27]. The experiments only used approximation coefficients as they contain most of the peak information of the original signal. The wavelet approximations of a signal at a certain level describe generalised peak lists of the de-noised spectrum [28]. Selecting features this way reduces the dimensionality of the original data while retaining the important features [26].

Figure 3 shows the whole classification process. After data conversion and scaling, a kernel selection is performed based on particular parameter search that best produces the best cross-validation result. We used the following kernels: Linear, Radial Basis Function (RBF), Sigmoid and Polynomial. These are shown in table 1 with parameters for best model selection.

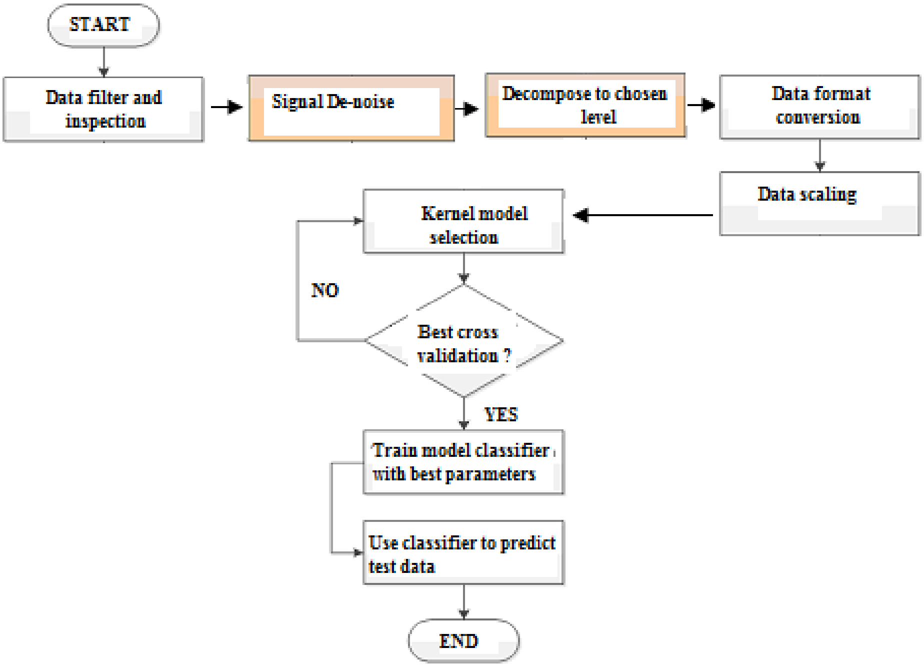
START
Data filter and inspection
Signal De-noise
Decompose to chosen level
Data format conversion
Data scaling
Kernel model selection
NO
Best cross validation ?
YES
Train model classifier with best parameters
Use classifier to predict test data
END

Fig. 3. SVM classification with feature selection– search for best cross validation result

Table 1. Kernelselection model parameters (Y=yes, N=no)

		Parameters			
		c Cost)	γ Gamma	γ (Radial)	d (degree)
Kernels	Linear	Y	N	N	N
	RBF	Y	Y	N	N
	Sigmoid	Y	Y	Y	N
	Polynomial	Y	Y	Y	Y

The experiments make use of grid search method provided in LIBSVM package. However the grid search method are time consuming and only optimize the pairs (C, γ) , therefore the experiment used, for a start, coarse grid values for the pair and later optimising the remaining parameters with finer grids.

All experiments used 5-fold cross validation. We note that Occam razor's theory mentioned in [29] suggests that smaller parameter values (in Table 1) are preferable to larger ones if they both have the same level of accuracy since building the SVM model with larger parameters takes longer. Thus our parameter search is simply based on (1) best level accuracy, (2) smallest parameter values.

4.2 Results for Experiment 1

For model selection using parameter settings in Table 1 and comparing both coarse and fine grid search, we obtained results shown in tables 2 and 3. Experiment 1 results are for signal data that does not use wavelet feature selection but only use model search for the best kernel and then performing a classification.

Table 2. Coarse Grid search cross-validation (CV%) (LIBSVM)

Kernel	c Cost)	γ Gamma	CV%
Linear	0.3125	2	76.2
RBF	1024	0.0000305	76.2
Sigmoid	1024	0.0000305	76.2
Polynomial	0.0000305	1	52.4

Table 3. Finer Grid Search cross-validation (CV%)(after MATLAB optimisation codes)

Kernel	c Cost)	γ Gamma	γ (Radial)	d (degree)	CV%
Linear	0.0078125	2	N/A	N/A	76.2
RBF	362.039	0.0000108	N/A	N/A	76.2
Sigmoid	256	0.0000305	0.0001	N/A	76.2
Polynomial	0.00195313	4	0.1	1	76.2

We found the same cross-validation accuracy of 76.2 (tables 2 and 3) across the kernels in both coarse and finer grid searches except for the polynomial kernel coarse grid which was 52.4%. Using the SVM classification on the trained data set, and computing accuracy (Equation 2), we obtained results as shown in table 4 with an equal number of support vectors (nSV).

$$Accuracy = \frac{\#Correctly\ classified\ signal\ data}{Total\ \#\ of\ \ signal\ data} \tag{2}$$

Table 4. Prediction accuracy –experiment 1

Kernel	Prediction accuracy	nSV	CV%	Elapsed time (s)
Linear	64.3	42	76.2	0.1412
RBF	64.3	41	76.2	0.1452
Sigmoid	64.3	42	76.2	0.1800
Polynomial	64.3	42	76.2	0.1851

Clearly as shown in table 4, linear kernel executes faster on average, given the same classification accuracy of 64.3% for all kernels. This meant that out of 28 classes (14x2), 64.3% of 28~18 classes correctly classified. It is also clear that as kernel complexity increases (e.g. linear kernel has one parameter compared to polynomial kernel with four), elapsed time also increases. Of note is the fact that cross-validation and prediction accuracy are the same. For experiment 2, we were compelled to use the linear and the RBF kernels for the reasons given above.

4.3 Results for Experiment 2

Table 4 shows experiment 1 results for cross validation, prediction accuracy, nSV and elapsed time for linear and RBK kernels using a pre-processing stage shown in figure 3.

To compare with a wavelet-based approach (experiment 2), several filters and levels were tested for the wavelet de-nosing part with two threshold levels, 50 and 70. Wavelet filters included the "bior1.1", "bior2.6", "bior3.7", "sym15" and "dmey". Results obtained are shown in table 5. The wavelet filter "bior2.6" achieved overall best performance with 75% prediction accuracy and a short elapsed time of 0.023389s. Comparing experiments 1 and 2, this represents an improvement in elapsed time of (0.14124-0.023389)/0.14124*100~83% if we compared with the best kernel "Linear kernel" from experiment 1. Prediction accuracy improved by (75-64.2857)/75*100~14.3%.

Similarly, the RBF kernel improved elapsed time by (0.1452-0.025458)/ 0.1452*100~82%, and accuracy by (71.4286-64.2857)/71.4286*100~10%.

Table 5. Comparison of results between experiments 1 and 2

Linear Kernel	**CV%**	**Elapsed Time**	**Prediction Accuracy**
Experiment 1	76.2	0.14124	64.3
Experiment 2.	76.2	0.023389	75.0
RBF kernel			
Experiment 1	76.2	0.1452	64.3
Experiment 2	76.2	0.025458	71.4

Results for experiment 2 and in table 5 confirm that classification of signal data with wavelet-based feature [26] extraction improves classification accuracy by at least 80% and runtime by at least 10%.

Table 6. File size comparison

Data Files	**Exp. 1 (KB)**	**Exp. 2 (KB)**	**% Reduction**
Training data	4329	987	77.2
Testing data	2917	685	76.5

Comparing data file sizes in kilobytes (KB) after data decomposition by the wavelet filter "bior2.6", table 6 shows that experiment 2 (with wavelet feature selection) had more than 77% and 76% feature reduction in the training and the testing data respectively. Both the literature in [26] and experiments shown here confirm that using SVMs with wavelet-based feature reduction improves the prediction accuracy, runtime and reduces features to be classified for MALDI-TOF signal data. These improvements would be significantly high with larger data.

5 Conclusion

The paper has presented SVM classification of MALDI-TOF signal data from bacteria colony using feature extraction by discrete wavelet transforms. Experiments tested the data with four kernels using various parameters so that a more suitable kernel was used for classification. Results showed that classification accuracy with feature selection improved accuracy by at least10%, and feature reduction by 76% and runtime by over 80%.

Further work is planned to explore other peak feature selection methods and identification of bacteria types in those peaks. Particular known denoising wavelet methods will also be used to check the cross-validation results and overall classification accuracy. Further, wavelet lifting schemes [20] may be tested with our approach. Wavelet lifting schemes are more efficient implementations of first generation wavelets and are not necessarily translates and dilates of one function, and thus do not rely on polynomial factorizations, as do Fourier transforms.

References

1. Lay, J.O.: MALDI-TOF Mass Spectrometry of Bacteria. John Wiley (2002)
2. Bundy, J., Fenselau, C.: Lectin-based Affinity Capture for MALDI-MS Analysis of Bacteria. Analy. Chem. 71(7), 1460–1463 (1999)
3. Li, T., Li, Q., Zhu, S., Ogihara, M.: A Survey on Wavelet Applications in Data Mining. SIGKDD Explorations 4(2), 49–68 (2003)
4. Bruyne, K.D., et al.: Bacterial Species Identification from MALDI-TOF Mass Spectra through Data Analysis and Machine Learning. Syst. and Appl. Microb. 34, 20–29 (2011)
5. Li, D., Pedrycz, W., Pizzi, N.J.: Fuzzy Wavelet Packet Based Feature Extraction Method and its Application to Biomedical Signal Classification. IEEE Trans. Biom. Eng. 526, 1132–1139 (2005)
6. Biotyper 2.0, `http://www.bdal.com/products/software/maldi-biotyper/overview.html`
7. Morris, J.S., Coombes, K.R., Koomen, J., Baggerly, K.A., Kobayashi, R.: Feature Extraction and Quantification for Mass Spectrometry in Biomedical Applications using the Mean Spectrum. Bioinformatics 21, 1764–1775 (2005)
8. Chui, C.K.: An Introduction to Wavelets. Academic Press, Boston (1992)
9. Daubechies, I.: Orthonormal Bases of Compactly Support Wavelets. Comm. Pure Appl. Math. 41, 909–996 (1988)
10. Daubechies, I.: Ten Lectures on Wavelets. Capital City Press, Montpelier (1992)
11. McDonough, R.N., Whale, A.D.: Detection of Signals in Noise, 2nd edn. Academic Press, San Diego (1995)
12. Conrad, T.O.F., Leichtle, A., Hagehülsmann, A., Diederichs, E., Baumann, S., Thiery, J., Schütte, C.: Beating the Noise: New Statistical Methods for Detecting Signals in MALDI-TOF Spectra Below Noise Level. In: Berthold, M., Glen, R.C., Fischer, I. (eds.) CompLife 2006. LNCS (LNBI), vol. 4216, pp. 119–128. Springer, Heidelberg (2006)
13. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. (2001)
14. Shin, H., Sampat, M.P., Koomen, J.M., Markey, M.K.: Wavelet-based Adaptive Denoising and Baseline Correction for MALDI-TOF MS. J. of Integr. Biol. 14(3), 283–295 (2010)

15. Pedrycz, W., Vukovich, G.: Feature Analysis through Information Granulation and Fuzzy Sets. Pattern Recog. 35, 825–834 (2002)
16. Resson, H.W., et al.: Peak Selection from MALDI-TOF Mass Spectra using Ant Colony Optimisation. Bioinformatics 23(5), 619–626 (2007)
17. Malyarenko, D.I., et al.: Enhancement of Sensitivity and Resolution of Surface-enhanced Laser Desorption Ionisation Time-of-flight Mass Spectrometric Records for Serum Peptides using Time-series Analysis Techniques. Clin. Chem. 51, 65–74 (2005)
18. Alexandrov, T., et al.: Biomarker Discovery in MALDI-TOF Serum Protein using Discrete Wavelet Transformation. Bioinformatics 25(5), 643–649 (2009)
19. Khushaba, R.N., Al-Jumaily, A.: Fuzzy Wavelet Packet Based Feature Extraction Method for Multifunction Myoelectric Control. J. of Biol. and Life Sci. 2(3), 186–194 (2007)
20. Sweldens, W.: Lifting Scheme: A New Philosophy in Biorthogonal Wavelet Constructions. In: SPIE Wavelet Applications in Signal and Image Processing III, vol. 2569, pp. 68–79 (1995)
21. Chih-Chung, C., Chih-Jen, L.: LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology 2(27), 1–27 (2011)
22. Hsu, C.W., Chang, C.C., Lin, C.J.: A Practical Guide to Support Vector Classification. Bioinformatics 1(1), 1–16 (2010)
23. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A Training Algorithm for Optimal Margin Classifiers. In: 5th Annual ACM Workshop on COLT, pp. 144–152 (1992)
24. Ramaswamy, R., et al.: Multiclass Cancer Diagnosis using Tumor Gene Expression Signatures. Proceedings of the National Academy of Sciences of the United States 98(26), 15149–15154 (2001)
25. Savchuk, O.Y., Hart, J.D., Sheather, S.J.: Indirect Cross-validation for Density Estimation. Amer. Stat. Ass. 105(489), 415–423 (2010)
26. Shutao, L., Chen, L., James, K.: Wavelet-based Feature Selection for Microarray Data Classification. In: Proc. Int. Joint Conference on Neur. Net. (IJCNN), pp. 5028–5033 (2006)
27. Frank-Michael, S., et al.: Support Vector Classification of Proteomic Profile Spectra Based on Feature Extraction with the Bi-orthogonal Discrete Wavelet Transform. Comp. and Visual. in Sci. 12(4), 189–199 (2009)
28. Wong, L., Muyeba, M., Keane, J.: Towards Adaptive Mining of Spectral Features. In: Proceedings of UK Workshop on Computational Intelligence, pp. 213–216 (2011)
29. Smith, M., Martinez, T.: Improving Classification Accuracy by Identifying and Removing Instances that Should Be Misclassified. In: Proc. Int. Joint Conference on Neur. Net. (IJCNN), San Jose, pp. 2690–2697 (2011)