Self-taught learning for classification of mass spectrometry data: a case study of colorectal cancer

Theodore Alexandrov
Center for Industrial Mathematics (ZeTeM), University of Bremen,
Bibliothekstr. 1, D-28209 Bremen, Germany, theodore@math.uni-bremen.de

Abstract: Mass spectrometry is an important technique for chemical profiling and is a major tool in proteomics, a discipline interested in large-scale studies of proteins expressed by an organism. In this paper we propose using a sparse coding algorithm for classification of mass spectrometry serum protein profiles of colorectal cancer patients and healthy individuals following the so-called self-taught learning approach. Being applied to the dataset of 112 spectra of length 4731 bins, the sparse coding algorithm represents each of them by means of less then ten prototype spectra. The classification of spectra is done as in our previous study on the same dataset [ADM$^+$09], using Support Vector Machines evaluated by means of the double cross-validation. However, the classifiers take as input not discrete wavelet coefficients but the sparse coding coefficients. Comparing the classification results with reference results, we show that providing the same total recognition rate, the sparse coding-based procedure leads to higher generalization performance. Moreover, we propose using the sparse coding coefficients for clustering of mass spectra and demonstrate that this approach allows one to highlight differences between the cancer spectra.

1 Introduction

Mass spectrometry (MS) is an important technique for chemical profiling and is a major tool in proteomics, a discipline interested in large-scale studies of proteins expressed by an organism. In medicine, MS-based proteomics contributes to clinical research by identification of biomarker proteins related to a disease, e.g. produced by a tumor tissue or by the immune system in response to a disease. Since 2002, when it was first proposed to classify cancer patients and healthy individuals based on MS protein profiles, researchers have shown an increased interest in application of mass spectrometry for biomarker detection.

Given a sample of blood, urine or serum, an MS instrument produces a high dimensional histogram-like spectrum. The peaks of the spectrum express chemical compounds with high concentrations. The spectra for different groups of subjects are collected (e.g. cancer patients and control individuals groups) and a quality of classification is studied. If a successful classification is possible, one is interested in interpreting peaks which are used in the classification and in identifying proteins corresponding to those peaks.

In [ADM$^+$09], we investigated the use of Discrete Wavelet Transformation (DWT) together with Support Vector Machines (SVM) for classification of spectra of colorectal cancer patients and healthy individuals. First, we calculated wavelet coefficients for each spectrum. Then statistically different coefficients were classified using SVM. Along with standard DWT we exploited APPDWT ("approximation DWT"), a modified DWT where only approximation coefficients were used. The classification results proved that this type of DWT outperforms the standard DWT. APPDWT can be interpreted as dictionary representation of a spectrum, where the dictionary is constructed by translating and shifting a wavelet scaling function.

Recently, [LBRN06] introduced a sparse coding (SC) algorithm which, given a set of vectors, learns in an unsupervised manner a sparse basis for optimal linear representation of the original vectors. Note that the basis can be overcomplete and its elements are not necessarily orthonormal, i.e., formally speaking, it is not a basis but a dictionary. In [ASKS09] we demonstrated that being applied to MS data, the SC algorithm allows one to pick class-relevant peaks. For this aim, we improved the original SC algorithm replacing l_1-regularization with an elastic-net regularization (combination of l_1- and l_2-regularization terms), for more details see [ASKS09] and [AKL+09].

Later, [RBL+07] proposed using the SC algorithm for classification, calling their approach "self-taught learning" as features used in classification are learned from the data. In this paper we follow this approach, classifying mass spectra of colorectal cancer patients and healthy individuals. The improved version [AKL+09] of the SC algorithm is used. For the classification the same scheme as in [ADM+09] is applied, but instead of DWT (APPDWT) coefficients we exploit the coefficients of the basis learned using the SC algorithm.

Our procedure of classification of mass spectra is as follows. First, given a set of spectra of different classes, we apply the SC algorithm producing a set of few basis vectors and a matrix of coefficients representing each original spectrum in the basis learned. We call each basis vector a prototype spectrum. Use of SC coefficients for MS data processing is promising because peaks of different width can be extracted. In the ideal case, any peak or combination of peaks which take place in sufficiently many spectra and represents a sizable contribution to a large portion of the dataset, will be represented using a SC coefficient. For each original spectrum we build a feature vector consisting of its coefficients. Second, the feature vectors are classified using SVM where the evaluation is done by mean of the double cross-validation, for more details see [ADM+09].

In Section 2 we concisely describe the data investigated, as well as the sparse coding algorithm and the classification scheme used. In Section 3.1 we present the results of SC algorithm. Then, in Section 3.2, we show the classification results and compare them with the reference results of [ADM+09]. Moreover, in Section 3.3 we provide a closer look at the SC results and propose clustering the spectra based on the SC coefficients. Section 4 concludes the paper.

2 Methods

2.1 Mass spectrometry data

The dataset used in this paper consists of matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) serum protein profiles of colorectal cancer patients and healthy individuals, first published in [dNMO$^+$06]. Colorectal cancer is one of the most common malignancies and remains a principal cause of cancer-related morbidity and mortality. Diagnosing colorectal cancer still requires a sensitive test relaying on easily accessible body fluids, like serum. After a preprocessing of spectra and outliers removal, described in [ADM$^+$09], we have 64 cancer and 48 control spectra of length 16331 points covering an mz (mass-over-charge) domain of 960–11163 Da.[1] For this paper we took only a part of the whole mz domain, namely 1100–3000 Da, which contains the most significant peaks for the cancer discrimination according to [dNMO$^+$06] and [ADM$^+$09]. A part of a spectrum in this domain consists of 4731 points. The final data is shown in Fig. 1

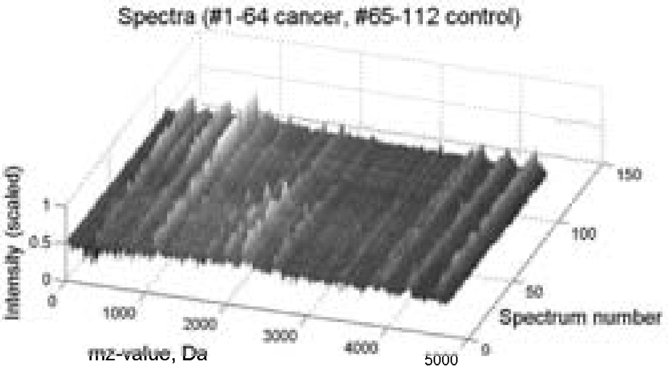

Figure 1: 64 cancer (with numbers 1-64) and 48 control mass spectrometry protein profiles.

2.2 Improved Sparse Coding algorithm with elastic-net regularization

For each original spectrum the SC algorithm calculates its coefficients in a basis expansion, where the basis vectors are learned from the data as follows.

We suppose that the dataset consists of R spectra of length L which belong to D classes ($D \ll R$) each characterized by common peaks at the same positions and with similar heights. Given a matrix $\mathbf{X} \in \mathbb{R}^{L \times R}$ with spectra in columns, the improved SC algorithm with an elastic-net regularization term represents each spectrum (a column of $\mathbf{X}$) in a self-

[1] The data is available at `http://www.math.uni-bremen.de/∼theodore/MALDIDWT`.

taught sparse basis solving the following optimization problem:

$$\min_{\mathbf{B},\mathbf{S}} \quad \frac{1}{2}\left|\left|\mathbf{X}-\mathbf{B}\mathbf{S}\right|\right|_{\mathrm{F}}^2+\alpha\sum_j \left|\left|S_j\right|\right|_1+\frac{\beta}{2}\sum_j \left|\left|S_j\right|\right|^2, \tag{1}$$

$$\text{subject to } \left|\left|B_j\right|\right|^2 \leq \gamma, \tag{2}$$

with respect to a matrix $\mathbf{B} \in \mathbb{R}^{L\times L}$ of basis vectors and a matrix $\mathbf{S} \in \mathbb{R}^{L\times R}$ of the corresponding coefficients, where $||\cdot||_{\mathrm{F}}$ is the matrix Frobenius norm, $||\cdot||_1$ is the vector l_1-norm, and $||\cdot||$ is the standard euclidean norm; S_j and B_j denote the j-th column of $\mathbf{S}$ and $\mathbf{B}$, respectively. The hyperparameters of the optimization problem are the l_1-regularization parameter α, l_2-regularization parameter β and the boundary on the basis vectors norm γ.

The minimization problem (1) is solved in two steps. First, we learn the coefficients $\mathbf{S}$ keeping the basis fixed using the Feature Sign Search (FSS) algorithm minimizing (1) for a fixed $\mathbf{B}$, then for the learned coefficients we optimize the basis $\mathbf{B}$ using the Lagrange dual. For more details, see [LBRN06]. For motivation of using the elastic-net regularization instead of the original l_1-regularization, see [AKL$^+$09] and [ASKS09].

Finally, for each column X_j of $\mathbf{X}$ we have its sparse representation in the basis $\mathbf{B}$ with only a few basis vectors B_j ($j \in \mathcal{I}$) corresponding to non-zero rows S_j with indices $\mathcal{I}$.

2.3 Classification using Support Vector Machines with double cross-validation

After the SC algorithm produced a matrix $\mathbf{B}$ of basis vectors and a matrix $\mathbf{S}$ of coefficients, we classified the spectra where for each spectrum its coefficients (that is j-th column S_j of $\mathbf{S}$ for j-th spectrum) are used as features. The classification was performed using Support Vector Machine (SVM) of type C-SVM with the gaussian kernel with two-level grid search for the hyperparameters σ (the width of the gaussian kernel) and C (the C-SVM regularization parameter). The tested values are $2^{-4:2:16}$ (a grid with values from 2^{-4} to 2^{16} with a step 2^2) for σ and $2^{-4:2:12}$ for C at the first grid search level and $2^{-1:1:1}$ for both σ and C at the second level of grid search used for refinement. The simultaneous parameters selection and classifiers assessment was done by means of the double cross-validation (double CV) with the leave-one-out cross-validation (i.e. 112-fold) used for the outer loop and 10-fold cross-validation used for the inner loop, again as in [ADM$^+$09]. In this setting, the i-th step of the double CV scheme consists of two stages: (1) the choice of hyperparameters is done using 10-fold CV on all but the i-th spectrum optimizing CV recognition rate (the ratio of spectra correctly classified in CV), (2) a classifier with the chosen hyperparameters is trained using all but the i-th spectrum and applied to the i-th spectrum excluded at the first.

The following characteristics were calculated after the outer loop classification: total recognition rate or TRR (the ratio of correctly classified spectra), specificity, and sensitivity. Moreover, following [BST99] and [ADM$^+$09], we considered the number of support vectors (SV) as a measure of generalization performance of classifiers. The values of these characteristics have been compared with corresponding values reported in [ADM$^+$09], where the same dataset is used (except for the mz-domain as explained in section 2.1).

3 Results

3.1 Sparse coding representation

We applied to the matrix $\mathbf{X} \in \mathbb{R}^{4371 \times 112}$ with spectra in columns the improved SC algorithm with an elastic-net penalty term with different values of parameters α (from 5 to 100 with a step 5) and γ (from 500 to 2500 with a step 500). The used value of the parameter corresponding to the l_2-penalty was $\beta = 10^{-10}$, which was selected as small as possible, as recommended in [AKL$^+$09].

For each pair of parameters (α, γ) we calculated the matrices $\mathbf{B}$ and $\mathbf{S}$ of basis vectors and corresponding coefficients. Recall that only basis vectors with indices $\mathcal{I}$ corresponding to non-zero rows of the coefficients matrix $\mathbf{S}$ are considered. In the following we refer to the computed basis vectors as the prototype spectra because each original spectrum is a linear combination of the basis vectors with weights equal to the corresponding coefficients. Fig. 2 shows the numbers of prototype spectra (sizes of $\mathcal{I}$) for all pairs of α and γ. As can be seen from Fig. 2, for all parameters considered a spectrum can be represented by only a small number of prototype spectra (from 2 to 10). Once, for $\alpha = 45$ and $c = 1000$, only one prototype spectrum is produced. Interestingly, though it is natural to expect that the number of prototype spectra increases as α decreases (because α is a multiplier of the sparsity term), this effect can be hardly observed.

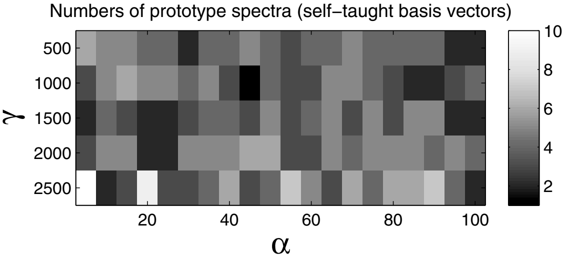

Figure 2: The numbers of prototype spectra for all pairs of α and γ considered.

In the following we consider the results of the SC algorithm for $\alpha = 10$ and $\gamma = 1000$ selected as producing the best classification results (presented later in Section 3.2). Fig. 3 shows the five prototype spectra computed for these parameters. Fig. 4 depicts the non-zero rows of the matrix $\mathbf{S}$ (each row is normalized to have values from zero to one). One can visually observe that the 4-th and 5-th rows highly discriminate cancer (the first 64) and control (the last 48) spectra since their values are visually grouped into two clusters: corresponding to spectra with numbers 1-64 and 65-112. To confirm this observation and to evaluate the separation efficiency of the produced coefficients, we plot a Principal Components Analysis (PCA) score plot, see Fig. 4 which shows clear though not ideal separation between two classes. Here PCA is used only for visualization. In next section we present close to perfect classification results achieved using SVM. A PCA score plots scores of the second principal component against scores of the first principal component and is often used for visualization of high-dimensional data. Fig. 4 demonstrates that the computed coefficients after a linear PCA-transformation allows one to clearly separate the groups of cancer and control individuals. This confirms the potential of using sparse coding coefficients for classifying cancer and control spectra.

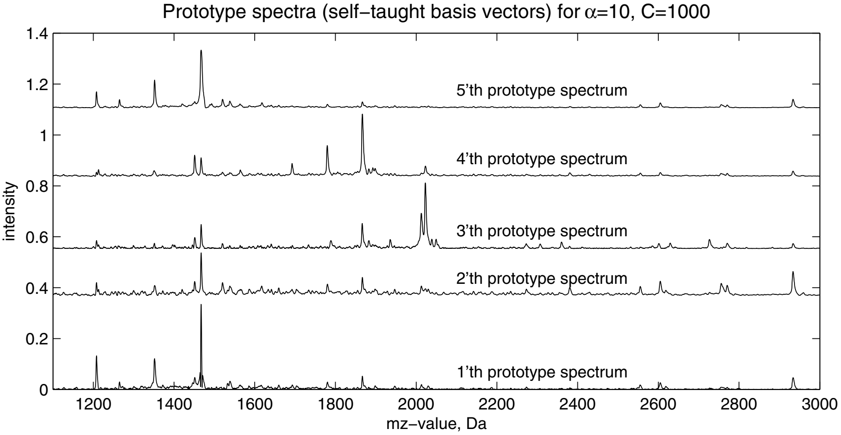

Figure 3: Prototype spectra (self-taught basis vectors) corresponding to non-zero coefficients extracted for $\alpha = 10$, $\gamma = 1000$, shifted in intensity for better visualization.

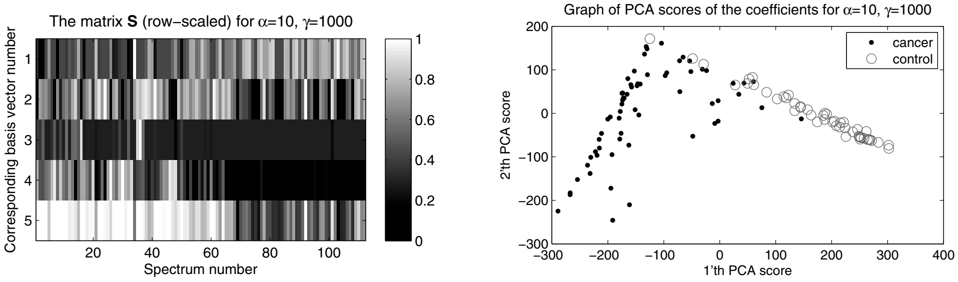

Figure 4: Left: non-zero coefficients (the matrix $\mathbf{S}$) used for representation of original spectra in the basis depicted on Fig. 3; right: a score plot showing the data (as usual, mean-corrected) projection onto the first two principal components.

3.2 Classification results

For each pair of the sparse coding parameters α and γ, we applied the SVM classification where the SVM hyperparameters selection and the classifiers assessment is done using the

		α						
		5	10	15	20	25	30	35
	500	94.64	93.75	93.75	90.18	88.39	90.18	90.18
	1000	92.86	**97.32**	91.07	87.50	89.29	89.29	91.07
γ	1500	90.18	91.07	90.18	89.29	90.18	89.29	90.18
	2000	92.86	96.43	94.64	91.07	87.50	91.07	91.07
	1500	93.75	91.07	87.50	95.54	91.96	92.86	91.07

Table 1: Total recognition rates for different α and γ for SVM classifiers using sparse coding coefficients, calculated through the double cross-validation. The best value (97.32%) is shown in bold.

double cross-validation, as described in Section 2.3 and, more detailed, in [ADM$^+$09]. The computed total recognition rates for all pairs of α and γ as well as the numbers of support vectors used are shown in Fig. 5.

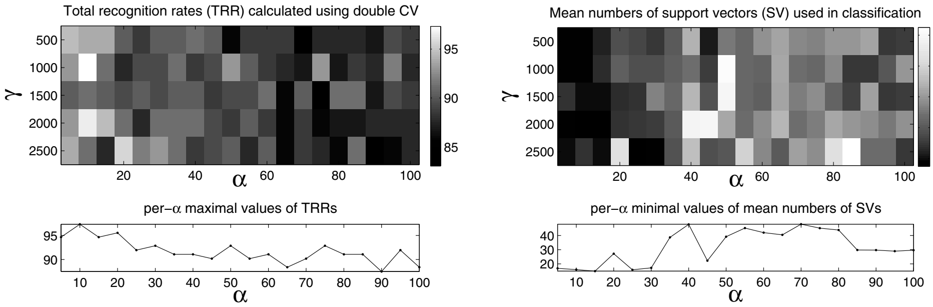

Figure 5: Classification results for different pairs of α and γ for SVM classifiers using sparse coding coefficients, calculated using the double cross-validation. Left: Total recognition rates; right: mean numbers of support vectors.

First, the achieved TRRs are quite high in comparison with the reference results. The best TRR is higher than the results of classification using the reduced-rank Linear Discriminant Analysis also evaluated using the double CV (92.6%) reported by [dNMO$^+$06] and is as hight as the results of the same classification procedure but applied on the DWT coefficients (97.3%) reported by [ADM$^+$09].

Although results presented in Fig. 5 are quite variable, there is a noticeable trend of decreasing TRR and increasing the mean number of SV (as α increases) that is better demonstrated by the plots of per-α maximal TRRs and per-α minimal mean number of SVs. For this reason, we showed in Table 1 and Table 2 the values of TRRs and the mean numbers of SV only for the first considered values of α (from 5 to 35). The best TRR is achieved for $\alpha = 10$ and $\gamma = 1000$ and is equal to 97.3% which is as high as reported by [ADM$^+$09] where DWT coefficients instead of sparse coding coefficients are exploited. The corresponding values of sensitivity and specificity are 96.9% and 97.9%, respectively. The most striking result to emerge from Table 2 is that the same classification efficiency is achieved using only 17 support vectors (corresponding to 15% of a training dataset of

		α						
		5	10	15	20	25	30	35
γ	500	16.8	18.6	23.5	37.1	33.4	29.6	41.2
	1000	17.9	16.6	27.2	42.2	36.9	37.4	42.6
	1500	28.0	20.3	20.3	27.2	28.4	54.1	44.8
	2000	18.2	15.9	14.9	28.6	30.1	31.4	41.5
	1500	20.2	28.1	20.2	85.8	15.8	17.2	38.7

Table 2: Mean numbers of SVM support vectors for different pairs of α and γ for SVM classifiers using sparse coding coefficients, calculated by means of the double cross-validation (the size of a training dataset is 111).

size 111) vs. 43 reported for the DWT-SVM procedure. It seems possible that the low numbers of the SV are due to the low number of features used in classification (less than 10 according to Fig. 2 vs. 300–600 for DWT and 1500–7000 for APPDWT as reported in [ADM$^+$09]).

As discussed in [ADM$^+$09], the number of support vectors is a proxy-measure of generalization performance of the classifiers. Any significant improvement of the generalization performance is very important in mass spectrometry-based proteomics, since the results should be reproducible when the data is prepared using different protocols, measured in different laboratories and in different conditions. All this leads to additional non-reducible variability in data and imposes high demands on the generalization performance of the exploited classifiers. From this point of view, the achieved advantage in the number of support vectors seems to be relevant and significant.

3.3 Closer look at the prototype spectra and sparse coding coefficients

Let us consider the 4-th and 5-th prototype spectra, see Fig. 6, since as conducted by means of visual inspection, their coefficients are the most discriminative between cancer and control groups. This choice is partially confirmed by the following fact. Considering the values of loadings of the first principal component (a direction of the largest variance) which are 0.1 (for the first prototype spectrum), -0.1 (second), -0.0 (third), -0.4 (fourth), -0.9 (fifth), we see that the 4-th and 5-th prototype spectra have the largest loadings, i.e. the largest contributions into a direction of the highest variance. Fig. 6 shows the scaled cancer and control mean spectra as well. The cancer (control) mean spectrum is manually attributed to the 4-th (5-th) prototype spectrum.

Fig. 6 shows that the prototype spectra are very similar to the per-class means spectra although they are extracted in an unsupervised manner, i.e. not using the labels of spectra.

It is interesting to compare Fig. 6 with Fig. 4a of [ADM$^+$09] showing the biomarker patterns reconstructed by the 1784 most discriminative APPDWT coefficients. In the region of 1100–2400 Da the prototype patterns are very similar to the biomarker patterns which is not surprising since they are similar to the per-class mean spectra. At the same time,

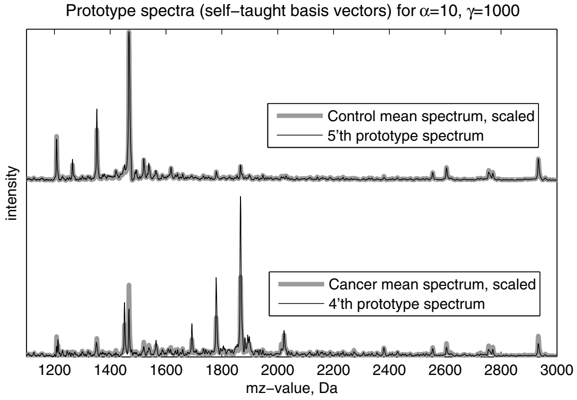

Figure 6: Forth and fifth prototype spectra (self-taught basis vectors) for $\alpha = 10, \gamma = 1000$ (shifted in intensity for better visualization) as well as the scaled cancer and control mean spectra.

note that the DWT biomarker patterns contain only a part of peaks presented in the mean spectra, which is especially noticeable in the region of 2400–3000 Da. This highlights the difference between local properties of wavelets and global (throughout the whole spectrum length) nature of the self-taught basis vectors.

An advantage of the self-taught sparse coding basis as compared to an APPDWT-induced dictionary is that it is learned in an unsupervised manner. Thus, the coefficients can be used not only for classification but also clustering of the spectra. For demonstration, we performed clustering of the spectra using High Dimensional Discriminant Analysis [BGS07]. The clusters number was set to 10 but the procedure automatically reduced it to 7; the used model is $[a_{ij} b_i Q_i d_i]$; the scree-test threshold is 0.2, for explanations see [BGS07].

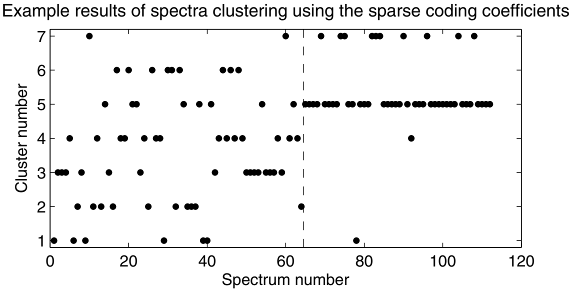

Figure 7: Example results of clustering of spectra using the sparse coding coefficients ($\alpha = 10$, $\gamma = 1000$): number of a cluster assigned to a spectrum against the spectrum number. A dash line is plotted for better visualization and separates cancer (spectra 1-64) from control (65-112) spectra.

Although Fig. 7 is shown mostly to demonstrate the potential of using sparse coding coefficients for spectra clustering, it is surprising to see that the control spectra are attributed only to two clusters. At the same time, the cancer spectra are not so homogeneous and form five clusters that probably indicates the difference in protein profiles of the cancer samples due to several tumor stages used in the measurements or other factors. A special investigation of these results is required which is out of scope of this paper.

4 Conclusions

In this paper we proved the potential of the sparse coding classification scheme proposed by [RBL+07] using the improved sparse coding algorithm of [AKL+09] for applications in mass spectrometry. The combination of SC and SVM demonstrated the same accuracy as DWT-SVM procedure [ADM+09] but with a significantly higher generalization performance measured by the number of support vectors. We demonstrate that the SC coefficients can be used not only for classification but also for clustering of the spectra.

Acknowledgements. The author thanks Stefan Schiffler for his implementation of the Feature Sign Search algorithm.

References

[ADM+09] T. Alexandrov, J. Decker, B. Mertens, A. M. Deelder, R. A. E. M. Tollenaar, P. Maass, and H. Thiele. Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation. *Bioinformatics*, 25(5):643–649, 2009.

[AKL+09] T. Alexandrov, O. Keszoecze, D. A. Lorenz, S. Schiffler, and K. Steinhorst. An active set approach to the elastic-net and its applications in mass spectrometry. In *Proc. Int. Workshop on Sparsity in Signal Processing (SPARS)*, 2009. Available at `http://hal.archives-ouvertes.fr/docs/00/36/93/97/PDF/19.pdf`.

[ASKS09] T. Alexandrov, K. Steinhorst, O. Keszöcze, and S. Schiffler. SparseCodePicking: feature extraction in mass spectrometry using sparse coding algorithms. In *Proc. IFCS'09, submitted*, 2009. Available at `http://arxiv.org/abs/0907.3426`.

[BGS07] C. Bouveyron, S. Girard, and C. Schmid. High-dimensional data clustering. *Comp. Stat. Data Anal.*, 52:502–519, 2007.

[BST99] P. Bartlett and J. Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in kernel methods: SV learning*, pages 43–54. MIT Press, 1999.

[dNMO+06] M. de Noo, B. Mertens, A. Ozalp, M. Bladergroen, M. van der Werff, C. van de Velde, A. Deelder, and R. Tollenaar. Detection of colorectal cancer using MALDI-TOF serum protein profiling. *Eur. J. Cancer*, 42(8):1068–1076, 2006.

[LBRN06] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *Proc. NIPS'06*, pages 801–808, 2006.

[RBL+07] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *Proc. 24th Int. Conf. on Machine learning (ICML)*, pages 759–766. ACM, 2007.