Front. Optoelectron. 2017, 10(3): 273–279
DOI 10.1007/s12200-017-0726-4

RESEARCH ARTICLE

Recursive feature elimination in Raman spectra with support vector machines

Bernd KAMPE[1], Sandra KLOß[1,2], Thomas BOCKLITZ[1,2], Petra RÖSCH[1,2], Jürgen POPP (✉)[1,2,3]

1 Institute of Physical Chemistry and Abbe Center of Photonics, University of Jena, Helmholtzweg 4, D-07743 Jena, Germany
2 InfectoGnostics Research Campus Jena, Center for Applied Research, Philosophenweg 7, 07743 Jena, Germany
3 Leibniz-Institute of Photonic Technology, Albert-Einstein-Straße 9, D-07745 Jena, Germany

Abstract The presence of irrelevant and correlated data points in a Raman spectrum can lead to a decline in classifier performance. We introduce support vector machine (SVM)-based recursive feature elimination into the field of Raman spectroscopy and demonstrate its performance on a data set of spectra of clinically relevant microorganisms in urine samples, along with patient samples. As the original technique is only suitable for two-class problems, we adapt it to the multi-class setting. It is shown that a large amount of spectral points can be removed without degrading the prediction accuracy of the resulting model notably.

Keywords feature selection, Raman spectroscopy, pattern recognition, chemometrics

1 Introduction

Raman microspectroscopy for the analysis of bacterial microorganisms is seeing increased usage over the last years [1]. One of the main challenges in the construction of a suitable model for those tasks lies in the complexity of the spectra due to the multitude of cell content. To complicate matters Raman spectra of bacteria tend to show a high degree of similarity on the species level [2].

As data sets become more and more complex, there is also a growing need for machine learning approaches that are suitable to process them. Support vector machines (SVMs) have been successfully used with data sets exhibiting hundreds and thousands of different features [3,4].

Traditional methods for the assessment of the importance of certain spectral components, like Student's *t*-test [5], only take the influence of individual components (or absence thereof) into account. This disregards the possible value of a combination of features for the discrimination of classes. By using a machine learning approach, we are able to pick a subset of features and assess the effect of it on the prediction error directly.

The performance of classifiers can degrade in the face of a high number of correlated features as shown by Kohavi and John [6]. Removing the variables not needed by the model to distinguish between sample groups also yields a better focusing of the important spectral differences and might therefore suggest the presence or absence of specific chemical components in the samples. This will also cause the model to be less prone to overfitting [7].

While this technique has already been used on data from DNA micro-arrays [8] and mass spectroscopy [9], to the best of our knowledge it has not been applied to Raman spectra.

Menze et al. [10] reported on the usefulness of random forest classifiers [11] for the removal of unnecessary features. While they feature an inherent robustness to noise and irrelevant patterns when it comes to classification, inducing forests of a few hundreds of decision trees takes several times longer than the procedure outlined here. It has also been reported that highly correlated features can lead to instability in the selection process of random forests [12].

In this paper, we introduce recursive feature elimination based on the weight vector of SVMs (SVM-RFE) into the field and report results on complex Raman spectra of bacteria along with its influence on the classification of patient samples. Since the original formulation of recursive feature elimination only works with two-class problems [13], we also provide an extension to make it work on problems with multiple classes.

Received April 4, 2017; accepted May 18, 2017

E-mail: juergen.popp@uni-jena.de

2 Materials and methods

We used a data set already described by Kloß et al. [14] which features common agents of urinary tract infection. It consists of a database of 11 important bacterial species: *Enterococcus faecalis* (DSM 20478), *Enterococcus faecium* (DSM 20477), *Staphylococcus epidermidis* (DSM 20044), *Staphylococcus haemolyticus* (DSM 20263), *Staphylococcus hominis* (DSM 20328), *Staphylococcus saprophyticus* (DSM 20229), *Staphylococcus aureus* (ATCC 43300), two strains of *Escherichia coli* (DSM 10806 and ATCC 35218), *Klebsiella pneumoniae* (ATCC 700603), *Pseudomonas aeruginosa* (ATCC 27853), and *Proteus mirabilis* (DSM 4479). The data set contains also an independent validation data set and spectra of patient samples. In total, there were 2952 spectra to train the model, 514 independent spectra for the validation of it and ten patient urine samples with a combined number of 491 spectra. All strains were provided by the Institute of Medical Microbiology, Jena University Hospital, and were originally purchased from the German Collection of Microorganisms and Cell Cultures (DSMZ) and the American Type Culture Collection (ATCC), except for the isolates from patient samples. All Raman spectroscopic measurements were performed with the Raman microscope BioParticleExplorer (MicrobioID 0.5, RapID, Berlin, Germany) with a 532 nm excitation. The spectral resolution was about 10 cm^{-1} [14]. Although we were tempted to adjust the pre-processing to account for developments in the meantime, to ensure comparability, the data set has been preprocessed exactly as before. This involved background correction with the statistics-sensitive nonlinear iterative peak-clipping (SNIP) algorithm [15], despiking using a robust variant of the upper-bound spectrum algorithm [16] and wavenumber calibration with acetaminophen [17]. After cutting them to the ranges of 450 – 1740 and 2610 – 3100 cm^{-1}, each spectrum consists of 553 spectral points. The exact composition of the training and validation data set is shown in Table 1.

Table 1 Numbers of spectra per data set

bacterial species	in training set	in validation set
E. faecalis	429	53
E. faecium	256	50
S. epidermidis	227	47
S. haemolyticus	225	52
S. hominis	207	49
S. saprophyticus	237	30
S. aureus	285	50
E. coli	360	37
K. pneumoniae	233	40
P. aeruginosa	249	39
P. mirabilis	244	67

SVMs belong to the class of maximum margin classifiers. They are designed to separate two classes of samples with a hyperplane between them that has maximum distance to both classes. This characteristic ensures a good generalization capability of the model [18]. All of the measured intensity values at the different spectral points combined lead to points in a high-dimensional space and spectra belonging to a specific species tend to cluster together due to their similarity. The data points that lie on the margin of these point clouds are called support vectors (SVs). These are usually far fewer than the number of spectra in general.

3 Recursive feature elimination

Conceptionally, SVMs work by calculating the optimal separating (hyper-)plane between pairs of point clouds. To make the separation easier, the data are projected into an even higher high-dimensional space. The projection function in combination with a dot product of two data points is known as the kernel of the SVM. In this work, the kernel used was the linear kernel which amounts to just using the dot products of all points $K(x_i,x_j) = x_i^{\mathrm{T}} x_j$. The breakthrough that makes this technique viable is the fact that the projection can happen after the dot product occurred [18], as long as the kernel function is positive-definite. The projection is only implied and never actually carried out as the target space can be infinite, as is the case for the popular radial basis kernel. This is referred to as the "kernel trick" in Ref. [19].

During the optimization process, most of the data points will never be considered as support vectors and therefore can be left out of the calculation. The resulting decision function is a linear combination of the support vectors. A new spectrum x is classified by

$$f(x) = \sum_{i=1}^{l} y_i \alpha_i K(x,x_j) - b, \tag{1}$$

where x_i is a support vector, y_i is the corresponding class label of it (1 or −1), α_i is a weight and b is the intercept term. The weights are bounded by a constant C, which is the sole user-defined parameter of the model and which will ensure convergence even in the case of non-separability of the data that is used to learn the model [8]. It has been set to a value of 100, which is rather low and in agreement with Ref. [8].

Trying out each and every combination of different spectral points to leave out of the model is only feasible for a really small number of features due to the combinatorial explosion such an approach entails. Two of the favorite approaches for handling large numbers of features are forward selection and backward selection, where one adds or removes one feature at a time, respectively. Backward selection is also known as recursive feature elimination [8].

It has been shown by Couvreur and Bresler that using backward selection leads to an optimal sequence of removed features, when the data has not been affected by noise too much [20]. It can be argued that this poses an inherent problem for Raman spectroscopy due to the low intensity of the effect. Nevertheless, the level of noise as shown in Fig. 1 did not seem to have a mentionable effect. The corresponding rank of the single spectral positions denotes the order in which features were removed over the course of selection with low-ranked points being eliminated early on.

Removing each remaining spectral position at each iteration of the procedure and comparing the resulting classifiers based on their respective accuracy would lead to a considerable and also prohibiting computational overhead.

In the case of a linear kernel, the weight vector w

$$w = \sum_{i=1}^{l} y_i \alpha_i x_i, \tag{2}$$

can be used directly to infer the feature that corresponds to the lowest influence on classification.

At the start of the method, all positions are chosen as input for the SVM. We then build all combinations of models and sum up the absolute values of the weighted support vectors. For the next round the position from all spectra that corresponds to the smallest component was removed, i.e., has the lowest influence on the classification. Afterwards, a new set of models was built on the remaining positions till we have removed all but one or have ended up with a pre-specified number.

Guyon et al. propose to rank the features by the value of $(w_i)^2$ [8]. Taking the square of all of the weights has the effect of turning all values positive. In the context of binary classification, it would not make a difference in the ranking, if we instead take the absolute values of the weights. As we generalize the method to the "all-pairs" scheme, this squaring introduces the effect of favoring few large weights for some classes in contrast to smaller weights for a lot of classes.

$$w = \sum_{j=1}^{k} \left(\sum_{SV \in J} y \alpha x \right)^2. \tag{3}$$

The "all-pairs" scheme is a simple extension of SVMs to the case of multi-class classification [21]. We build a model for each of the $\binom{n}{2}$ pairwise combinations of classes and achieve results by picking the class for each spectrum that receives the most votes. In Eq. (3), k refers to the pairwise combinations and J denotes the current model.

To avoid having to re-compute the models to assess the importance of the removal of a feature, it is assumed that the set of support vectors doesn't change with the removal [8]. Computing the entire feature ranking took about 5 h on an Intel(R) Core(TM) i5 750 at 2.66 GHz using only a single core. The procedure was implemented in R [22] using the kernlab [23] package.

Along with the removal of features from the spectra, we followed the development of the error rate of the 10-fold cross-validation (CV) (Fig. 2) on the training set and the prediction error on the validation set and the spectra of patient samples. Cross-validation works by splitting up the data set into mostly even parts and using all but one of the splits (the folds) to build the model. Prediction is done on the remaining fold. Afterwards, another fold is selected for prediction in turn till all of the folds have been used. As using one and the same data set for feature selection and prediction would lead to biased results, the stopping criterion for the method should be deduced from the error rate of the validation data set as seen in Fig. 3. For comparison, the prediction on the training set itself was flawless even when up to 492 spectral positions had been removed.

4 Results and discussion

As can be seen in Fig. 3, the first 200 selected points that are removed do not lead to a relevant increase of the error

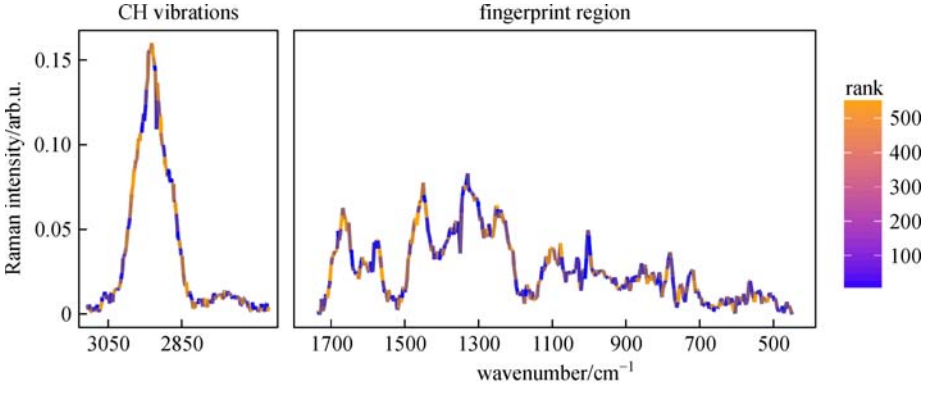

Fig. 1 Sample preprocessed Raman spectrum of *E. faecalis* showing the induced ranking of the method

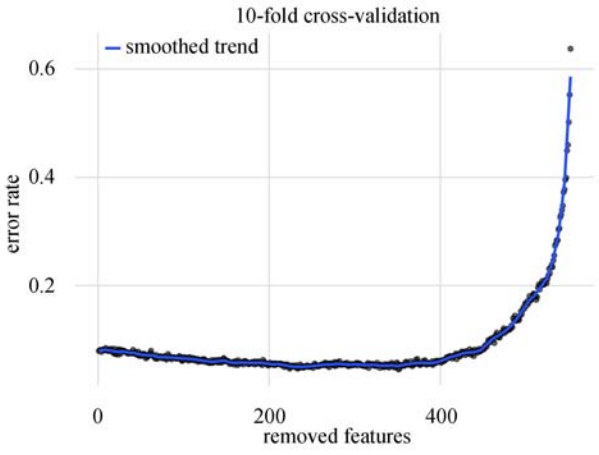

Fig. 2 Development of the cross-validation error rate on the training data set over the course of feature removal

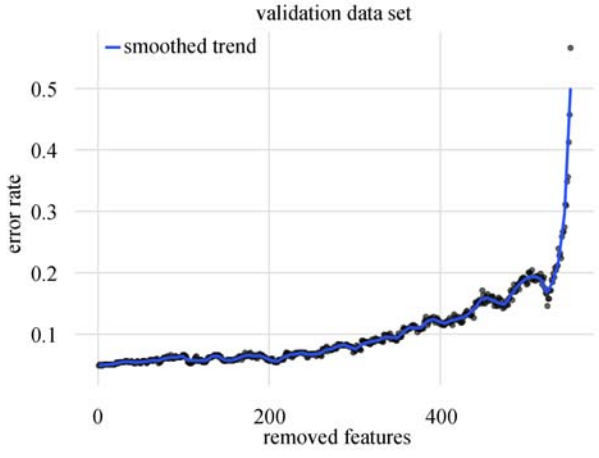

Fig. 3 Development of the classification error rate on the independent validation data set over the course of feature removal

rate of the classifier on the validation data set. Instead, one can even see the error rate of the cross-validation drop slightly over that range. This has been known as the peaking phenomenon [24,25] in the field of statistical pattern recognition for more than four decades, where the recognition rate of a model on the training data reaches a peak for a given number of features and then settles into a plateau or slowly begins to drop again. It should be noted that a lower error rate on the training data leads to an improved generalization ability of SVMs [13].

The behavior of the error rate in Fig. 3 can be roughly divided into a flat region, a region where it is slowly rising and a final part where the prediction capability breaks down entirely when most of the spectral points have been removed. The rise of the error rate coincides with the settling into the plateau phase of the cross-validation setup. Those two marks indicate that all of the measurements that contain duplicated or irrelevant information for the classification task have been removed. One possible stopping condition of the feature elimination would be to keep track of the change of the error rate and stop the procedure, if the decrease was smaller than a certain threshold over a defined last number of steps. Since the partitioning of the folds for the CV happens randomly, care has to be taken not to compare against a false minimum. It is advisable to choose a moving average for the tracking of the progress.

Features of a Raman spectrum are necessarily highly correlated as Raman signals usually occupy several neighboring positions in the digitized spectrum. This also means that the subset of features that was found might not be unique, as spectral positions that are perfectly correlated might be exchanged with each other.

While the cross-validation error fell from 7.9% at the start to 4.8% at the beginning of the plateau, the error on the validation set rose minimally from 4.9 to 5.5%. For comparison, on the patient data set (Fig. 4) we observed slight shifts of the error from 18.1% over a minimum after 107 removals of 17.7% to a minor rise to 18.5%. Parts of the erroneous attribution on the patient data set are due to an overlap of the classes of *E. coli* and *K. pneumoniae*, which is also present in the training data set.

As it can be difficult to estimate the influence of the method on the accuracy of the overall combination of all SVM models, Fig. 5 highlights the results on the model responsible for discriminating between the two species most commonly found in the spectra of our patient samples. Only 12 spectral points are needed for a clean separation. The development of the CV error mirrors that of Fig. 2, albeit with a much sharper rise toward the end of the run. The same could be found for the results on the validation data set. Since the set of points that is important for the discrimination between each of the pairs of classes is different in each case, this will have an effect on the induced ranking in the multi-class setting.

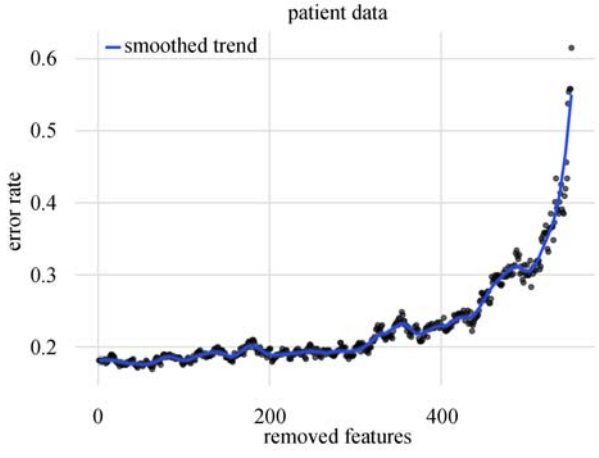

Fig. 4 Development of the classification error rate on the data set of patient samples over the course of feature removal

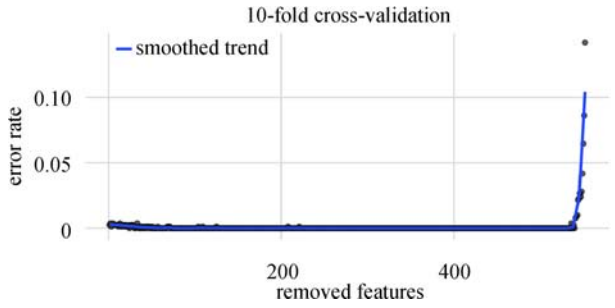

Fig. 5 Trend of the accuracy of the model discriminating *E. faecalis* from *E. coli* based on cross-validation

Another general technique for feature selection is the integration of the least absolute shrinkage and selection operator (lasso). This has e.g. been applied to principal component analysis and linear discriminant analysis (LDA) to obtain sparse and interpretable solutions through regularization [26]. As penalized LDA is readily available as a package (penalizedLDA) in R, we tried to compare the method to SVM-RFE, but found that the cross-validation reached an accuracy of 61.8% at best while retaining close to all of the features. While it is successfully used with DNA microarray data, the structure of our data seems to prevent us from making use of it.

Instead of using an embedded method like SVM-RFE, wrapper methods that combine Genetic Algorithms (GAs) with classifiers – not necessarily SVMs – have been applied to spectroscopic data. Lavine et al. [27] used GAs to separate different types of wood whose spectra were quite similar to another. GAs operate on vector of '0's and '1's and optimization of them over several iterations can lead to an effective filter mask for spectra, where spectral positions with a corresponding 0 in the vector representation will be kept out of the model. In this case, feature subset selection takes place in parallel instead of sequentially by trying out different sets of candidates each iteration and modifying the best ones.

One of the apparent bottlenecks of multi-class SVM-RFE seems to lie in the computation of all pairwise model combinations which scales with $\binom{n}{2}$, n being the number of classes. A possible solution to this could be to switch from the "all-pairs" scheme to the "one-vs-all" (OVA) scheme which was shown by Rifkin and Klautau [21] to be equally well suited for multi-class classification. In OVA, one builds n SVMs where the nth class makes up one class and all the others make up the other class. Then, we choose the model which gives the largest response. This results in fewer, but larger, models. Of course, as the models can be trained independently of the others in both cases, it would also be trivial to parallelize the computation of them.

5 Conclusions

In this contribution, we applied for the first time SVM-based recursive feature elimination to Raman spectra. This involved deriving a suitable generalization for multi-class problems. Even with spectra as complex as bacterial species in urine, this technique manages to reduce the amount of spectral points needed to build a reliable model for the discrimination between them by more than a third. A look at the individual models for pairwise separation enables one to deduce relevant factors for the discrimination between groups. Instead of computing each reduced model from scratch, it is possible to initialize the SVMs with the α values of the previous iteration, thereby speeding up the optimization process [28]. Recursive feature elimination can also be applied to other kernels and we are interested in finding out if the results transfer to them, too.

Acknowledgements Funding of the research project InterSept (13N13852) from the Federal Ministry of Education and Research, Germany (BMBF) is gratefully acknowledged.

References

1. Stöckel S, Kirchhoff J, Neugebauer U, Rösch P, Popp J. The application of Raman spectroscopy for the detection and identification of microorganisms. Journal of Raman Spectroscopy : JRS, 2016, 47(1): 89–109
2. Meisel S, Stöckel S, Rösch P, Popp J. Identification of meat-associated pathogens via Raman microspectroscopy. Food Microbiology, 2014, 38: 36–43
3. Rösch P, Harz M, Schmitt M, Peschke K D, Ronneberger O, Burkhardt H, Motzkus H W, Lankers M, Hofer S, Thiele H, Popp J. Chemotaxonomic identification of single bacteria by micro-Raman spectroscopy: application to clean-room-relevant biological contaminations. Applied and Environmental Microbiology, 2005, 71 (3): 1626–1637
4. Mukherjee S. Classifying Microarray Data Using Support Vector Machines in A Practical Approach to Microarray Data Analysis. Boston: Springer US, 2003, 166–185
5. Bocklitz T, Putsche M, Stüber C, Käs J, Niendorf A, Rösch P, Popp J. A comprehensive study of classification methods for medical diagnosis. Journal of Raman Spectroscopy: JRS, 2009, 40(12): 1759–1765
6. Kohavi R, John G H. Wrappers for feature subset selection. Artificial Intelligence, 1997, 97(1–2): 273–324
7. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics (Oxford, England), 2007, 23(19): 2507–2517
8. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using Support Vector Machines. Machine Learning, 2002, 46(1/3): 389–422

9. Granitto P M, Furlanello C, Biasioli F, Gasperi F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometrics and Intelligent Laboratory Systems, 2006, 83(2): 83–90
10. Menze B H, Kelm B M, Masuch R, Himmelreich U, Bachert P, Petrich W, Hamprecht F A. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics, 2009, 10(1): 213
11. Breiman L. Random forests. Machine Learning, 2001, 45(1): 5–32
12. Toloşi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics (Oxford, England), 2011, 27(14): 1986–1994
13. Cortes C, Vapnik V. Support-vector networks. Machine Learning, 1995, 20(3): 273–297
14. Kloß S, Kampe B, Sachse S, Rösch P, Straube E, Pfister W, Kiehntopf M, Popp J. Culture independent Raman spectroscopic identification of urinary tract infection pathogens: a proof of principle study. Analytical Chemistry, 2013, 85(20): 9610–9616
15. Morháč M, Kliman J, Matoušek V, Veselský M, Turzo I. Background elimination methods for multidimensional coincidence γ-ray spectra. Nuclear Instruments & Methods in Physics Research Section A, Accelerators, Spectrometers, Detectors and Associated Equipment, 1997, 401(1): 113–132
16. Zhang D, Jallad K N, Ben-Amotz D. Stripping of cosmic spike spectral artifacts using a new upper-bound spectrum algorithm. Applied Spectroscopy, 2001, 55(11): 1523–1531
17. Dörfer T, Bocklitz T, Tarcea N, Schmitt M, Popp J. Checking and improving calibration of Raman spectra using chemometric approaches. Zeitschrift für Physikalische Chemie, 2011, 225(6–7): 753–764
18. Boser B E, Guyon I M, Vapnik V N. A training algorithm for optimal margin classifiers. In: Proceedings of the 5th Annual Workshop on Computational Learning Theory. New York: ACM, 1992, 144–152
19. Vapnik V. The Nature of Statistical Learning Theory. 2nd ed. New York: Springer Science & Business Media, 2013
20. Couvreur C, Bresler Y. On the optimality of the backward greedy algorithm for the subset selection problem. SIAM Journal on Matrix Analysis and Applications, 2000, 21(3): 797–808
21. Rifkin R, Klautau A. In defense of one-vs-all classification. Journal of Machine Learning Research, 2004, 5: 101–141
22. R Core Team. R: A language and environment for statistical computing, R Foundation for Statistical Computing, 2016
23. Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab – An S4 package for kernel methods in R. Journal of Statistical Software, 2004, 11(9): 1–20
24. Van Campenhout J M. Topics in measurement selection. In: Handbook of Statistics. Elsevier, 1982, 793–803
25. Sima C, Dougherty E R. The peaking phenomenon in the presence of feature-selection. Pattern Recognition Letters, 2008, 29(11): 1667–1674
26. Witten D M, Tibshirani R. Penalized classification using Fisher's linear discriminant. Journal of the Royal Statistical Society Series B, Statistical Methodology, 2011, 73(5): 753–772
27. Lavine B K, Davidson C E, Moores A J, Griffiths P R. Raman spectroscopy and genetic algorithms for the classification of wood types. Applied Spectroscopy, 2001, 55(8): 960–966
28. Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research, 2003, 3: 1157–1182

Bernd Kampe studied bioinformatics at the Friedrich Schiller University Jena, Germany. He subsequently joined the work group of Jürgen Popp in July 2010 to start his Ph.D. studies focused on the identification of microorganisms with micro-Raman spectroscopy and chemometric methods, especially support vector machines. He is currently with the Jena University Language & Information Engineering Lab (JULIE Lab), where his research is aimed at the extraction of information about protein-protein interactions from text.

Sandra Kloß studied chemistry in Jena. In 2015 she received her Ph.D. at the Friedrich-Schiller-University Jena. Currently she is working as a post-doctoral researcher in the work group of Jürgen Popp. Her main research interests are the isolation of microorganisms from complex matrices and their subsequent Raman spectroscopic and molecular biological investigation.

Thomas Bocklitz studied physics at the Friedrich-Schiller-University. He received his diploma in theoretical physics in 2007 and the Ph.D. in chemometrics in 2011. Dr. Bocklitz is a junior research group leader for statistical data analysis and image analysis mostly for biophotonic applications. Dr. Bocklitz research agenda is closely connected with the translation of physical information, measured by AFM, TERS, Raman-spectroscopy, CARS, SHG, TPEF, into medical or biological relevant information. This research led to over 60 publications in peer-reviewed journals and his habilitation, which he completed in 2016.

Petra Rösch studied chemistry at the University of Würzburg. Actually she is research associate at the chair of Jürgen Popp at the University of Jena. Her research interests are focused on the investigation of all kind of biological, medical, and pharmaceutical relevant problems with various vibrational spectroscopic methods. Her main focus lays on the characterization and identification of microorganisms with Raman spectroscopy.

Jürgen Popp studied chemistry at the universities of Erlangen and Würzburg. After his Ph.D. in Chemistry he joined Yale University for postdoctoral work. He subsequently returned to Würzburg University where he finished his habilitation in 2002. Since 2002 he holds a chair for Physical Chemistry at the Friedrich-Schiller University Jena. Since 2006 he is also the scientific director of the Leibniz Institute of Photonic Technology, Jena. His research interests are mainly concerned with biophotonics. In particular his expertise in the development and application of innovative Raman techniques for biomedical diagnosis should be emphasized.