

Raman spectroscopy and chemometrical analysis for cancer research
Bocklitz, T.; Putsche, M.; Rösch, P.; Popp, J.

Abstract:

In this model study we developed a Raman based method to distinguish between breast cancer and normal epithelial cells, which is in principle suitable for online diagnosis. Two cell lines were chosen as model systems for cancer and normal tissue. Both cell lines consist of epithelial cells, but the cells of the MCF 7 series are carcinogenic, where the MCF10A cells are normally growing. An algorithm for distinguishing cells of the MCF7 and MCF10A cell line via its Raman spectra, which has accuracy rates above 99%, is presented. For this purpose two classification steps are utilized. The first one, the so called top level ANN searches for Raman spectra which are measured in the nuclei region. In the second step a wide range of discriminant models are possible, for example ANNs, LDAs or SVMs. These methods are compared. It was found that for the Top Level Classifier as for the Sub Level Classifier a ANN was superior.

Personal Dates:

Thomas Bocklitz
University Jena
Institute for Physical Chemistry
Jena, Germany
thomas.bocklitz uni-jena.de

Summary:

Introduction: Breast cancer is a fatal disease, which is the major cancer affection in women in Germany and Europe. 27,8% of all cancer diagnoses in women are breast cancer. 57000 women are diagnosed annually and 17500 are dying because of the disease. To achieve good treatment an early diagnosis is needed. The gold standard for diagnosis is the inspection by a pathologist. Therefore, a thin cryotome section is taken, stained and the morphology of the tissue is investigated. This method is highly dependent on the experience of the pathologist. An analysis routine which can support the diagnosis is needed.

Breast cancer emerges from mutations of the DNA of epithelial cells which forms the boundary of the lactiferous ducts and glands. This results in a different metabolism of the cells and the cells are abnormal proliferating. In order to do so, they are producing material for cell division, which yields to a different chemical composition of the cells. In principle this change is detectable with Raman spectroscopy. Every substance exhibits its own fingerprint spectrum and the spectrum of a cell is the weighted sum of all component spectra. The differences are small but are detectable, if statistical analysis is used.

In this case study two cell lines are utilized as model system for malignant and benign epithelial cells. MCF7 cells are carcinogenic and MCF10A belong to a healthy cell line.

Data analysis by ANN, SVM, LDA: All computations were done in R [6], a statistical language similar to Matlab. The packages used are called Nnet and e1071 [7, 2]. The supervised methods used for the classification problem are the commonly used methods in chemometric analysis and they are described elsewhere in more detail [1, 5, 8, 3]. Briefly a LDA searches for a linear combination of the wavenumbers to describe the classes. A new spectra is projected on the LD scaling vector and the sign of the result is equivalent to the class assignment

$$class(\vec{x}) = sgn(\vec{x} \cdot \vec{L}) . \quad (1)$$

A SVM belongs to the large margin classifiers and constructs a separating plane in the wavenumber space. In order to deal with non linear separable data the kernel trick is used. The decision function is then expressed with the support vectors $\vec{x}^{(i)}$, a weight vector $\vec{\alpha}$ and the offset b as

$$class(\vec{x}) = sgn\left(\sum_{i=1}^N y_i \alpha_i K(\vec{x}^{(i)}, \vec{x}) + b\right) . \quad (2)$$

A 3-layer Artificial Neuronal Network (ANN) is trained by a backpropagating algorithm, which determines the weights of the neuron connections so, that the calculated output and the known output¹ are minimal. Once the weight matrix is determined new samples

¹belonging to the classes

are classified through

$$class(\vec{x}) = \text{sgn} \left(f \left[\theta'_k + \sum_{j=1}^M w'_{kj} f' \left(\sum_{i=1}^n w_{ji} x_i + \theta_j \right) + \sum_{i=1}^n w''_{ki} x_i \right] \right). \quad (3)$$

Where f, f' and θ, θ' are the activating function and the bias of the hidden layer and the output layer, respectively. $\mathbf{w}, \mathbf{w}', \mathbf{w}''$ are the weight matrices for the input-hidden connection, the hidden-output connection and the shortcut connection.

Data analysis strategy: Because of the complexity of the classification problem, the problem was divided in two easier classification tasks (Figure 1). The idea behind this hierarchical system is *divide and conquer* and is described in [4]. As Top Level Classifier a Artificial Neuronal Net was chosen. This ANN classifies the spectra in 'Nuclei' and 'Rest'. The 'Nuclei' are then presented to the Sub Level Classifier, which decides the belonging to MCF10A and MCF7 cell line.

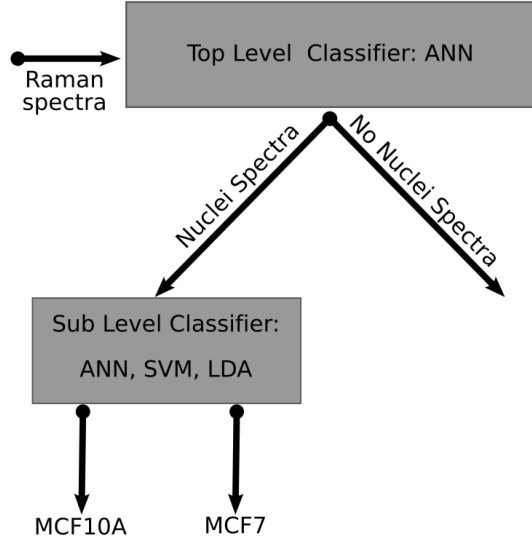


Figure 1: Principle of Classification

Sub level classification: To investigate the prediction ability to separate healthy from malign cells two methods are utilized. First a 25. CV was done the calculate the classification properties of the model. For this propose six supervised methods, namely LDA, linear SVM, radial base SVM, polynomial SVM, and two ANNs, were tested. The results are shown in figure 2. All the methods except the nonlinear SVMs perform quiet well. The accuracies range from 99, 45% (60:15:1 ANN), 99, 59% (60:10:1 ANN), 99, 42% (LDA), 99, 07% (linSVM), 98, 29% (polSVM) to 97, 75% (rbSVM). The sensitivity of all methods were better then 98, 5%, but the selectivity was only for four algorithms better

then 95%.

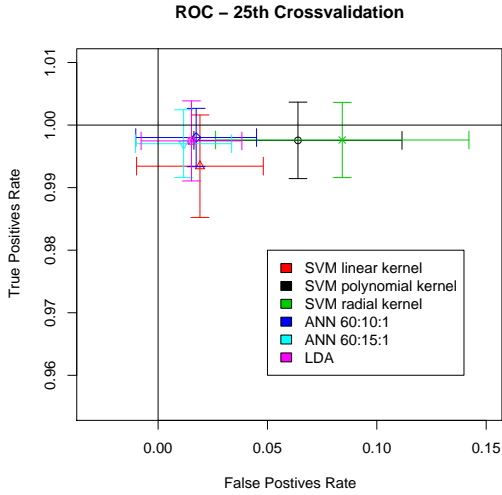


Figure 2: Classifier Performance

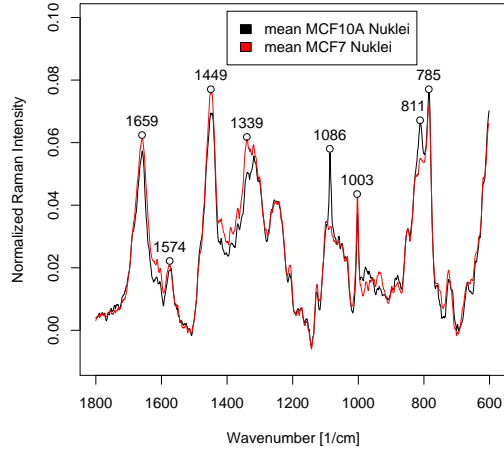


Figure 3: Mean of Classification

Three of these supervised methods were then tested for their identification rates with a holdout method. 34284 spectra were not used to build the model and the prediction of these spectra were compared with their true classes. Then the TP-rate, the FP-rate and the accuracies are calculated and were arranged in table 1. This time the 60:15:1 ANN was superior. The accuracy is 99,11%, the sensitivity 99,71% and the selectivity 96,62%. The mean of both predicted groups is plotted in figure 3.

Table 1: Sub Level Classification

Method	Accuracy	TP Rate	FP Rate
60:15:1 ANN	99,11%	99,71%	3,38%
linear SVM	96,51%	99,65%	16,55%
LDA	97,96%	99,79%	9,66%

Acknowledgment: The funding of the research project Exprimage (FKZ13N9364) within the framework Biophotonik from the Federal Ministry of Education and Research, Germany (BMBF) and the support of the Photonics4life network (Grand Agreement no.: 224014) are gratefully acknowledged.

References

- [1] BURGESS, Christopher J.: A Tutorial on Support Vector Machines. In: *Data Mining and Knowledge Discovery* 2 (1998), S. 121–167
- [2] DIMITRIADOU, Evgenia ; HORNIK, Kurt ; LEISCH, Friedrich ; MEYER, David ; WEINGESSEL, Andreas: *e1071: Misc Functions of the Department of Statistics (e1071)*, TU Wien, 2007. – R package version 1.5-17
- [3] HARZ, M. ; KIEHNTOPF, M. ; STÖCKEL, S. ; RÖSCH, P. ; DEUFEL, T. ; POPP, J.: Analysis of single blood cells for CSF diagnostics via a combination of fluorescence staining and micro-Raman spectroscopy. In: *The Analyst* 133 (2008), S. 1416–1423
- [4] LASCH, Peter ; BEEKES, Michael ; SCHMITT, Jürgen ; NAUMANN, Dieter: Detection of preclinical scrapie from serum by infrared spectroscopy and chemometrics. In: *Anal. Bioanal. Chem.* 387 (2007), S. 1791–1900
- [5] LASCH, Peter ; DIEM, Max ; HÄNSCH, Wolfgang ; NAUMANN, Dieter: Artificial neuronal networks as supervised techniques for FT IR microspectroscopic imaging. In: *Journal of chemometrics* 20 (2006), S. 209–220
- [6] R DEVELOPMENT CORE TEAM: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2007. <http://www.R-project.org>. – ISBN 3-900051-07-0
- [7] VENABLES, W. N. ; RIPLEY, B. D.: *Modern Applied Statistics with S*. Fourth. New York : Springer, 2002 <http://www.stats.ox.ac.uk/pub/MASS4>. – ISBN 0-387-95457-0
- [8] YANG, Husheng: Discriminant Analysis by Neural Networks. In: *Handbook of Vibrational Spectroscopy*. John Wiley & Sons, Ltd, 2002