Selection of Relevant Features for Raman Spectroscopy Using Supervised Classification Techniques

Michael Kemmler and Joachim Denzler

Institute of Computer Science
Friedrich Schiller University of Jena (Germany)

http://www.inf-cv.uni-jena.de michael.kemmler@uni-jena.de

seit 1558

Raman Spectroscopy

Motivation

- Identify microorganisms in medicine, food and pharmaceutical industry
- Using superposition of molecular responses

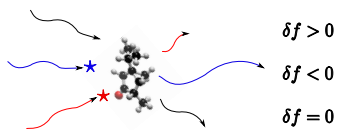
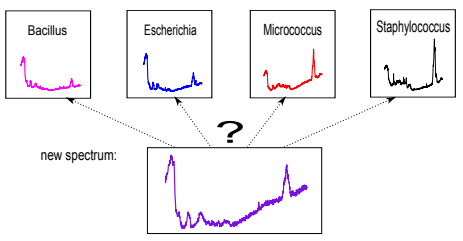


image adapted from [1]

Categorization Problem

- Correctly classify between different genera, species, ...



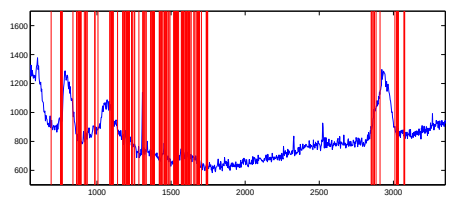
Feature Selection

Advantages

- Reduce computational efforts
- Improve classification accuracy
- Conclusions about important compounds might be possible

Our Aim

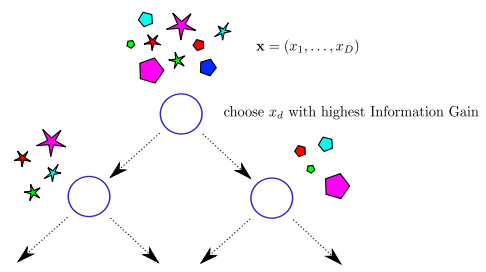
- Finding a subset in input space $\mathcal{X}$ to allow for interpretations



Classifiers I

Random Forest

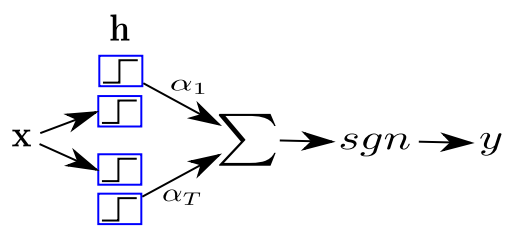
- Splits training examples at each node in order to construct pure nodes



- Random fraction of training samples per tree and features per node

Adaptive Boosting

- Adaptively learns a linear combination of classifiers



Regularized Logistic Regression

- Model class probability via logistic function, i.e.

$$p(y = +1|\mathbf{x}, \boldsymbol{\omega}) = \frac{1}{1 + \exp(-\boldsymbol{\omega}^T \mathbf{x})}$$

- MAP approach: $\boldsymbol{\omega}^* = \text{argmax}_{\boldsymbol{\omega}}\, p(\boldsymbol{\omega}) \cdot p(\mathbf{y}|\mathbf{X}, \boldsymbol{\omega})$
- Using "sparse" prior on $\boldsymbol{\omega}$, i.e. $\omega \sim \mathcal{N}(0, \frac{1}{\lambda})$

Classifiers II

Gaussian Process Classifier

- Probabilistically infer about output $p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)$ [2]:

$$p(f_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = \int p(f_*|\mathbf{X}, \mathbf{f}, \mathbf{x}_*)\, p(\mathbf{f}|\mathbf{X}, \mathbf{y})\, d\mathbf{f}$$

$$p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = \int p(f_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)\, p(y_*|f_*)\, df_*$$

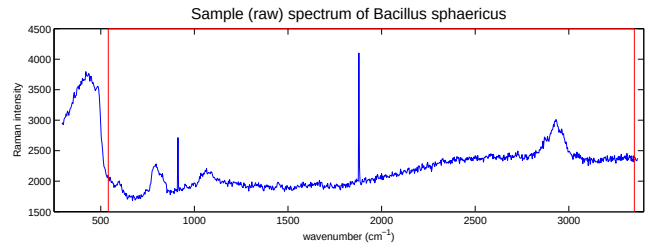
- If $\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma^2\mathbf{I})$, $\mathbf{f}|\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \kappa)$ then $y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_* \sim \mathcal{N}(\mu_*, \sigma_*^2)$
- Automatic Relevance Determination used for feature selection

$$\kappa_{\boldsymbol{\nu}}(\mathbf{x}, \mathbf{x}') = \nu_0^2 \cdot \exp\left(-\sum_{k=1}^{D} \frac{1}{2\nu_k^2}(x_k - x_k')^2\right)$$

Experiments

Dataset

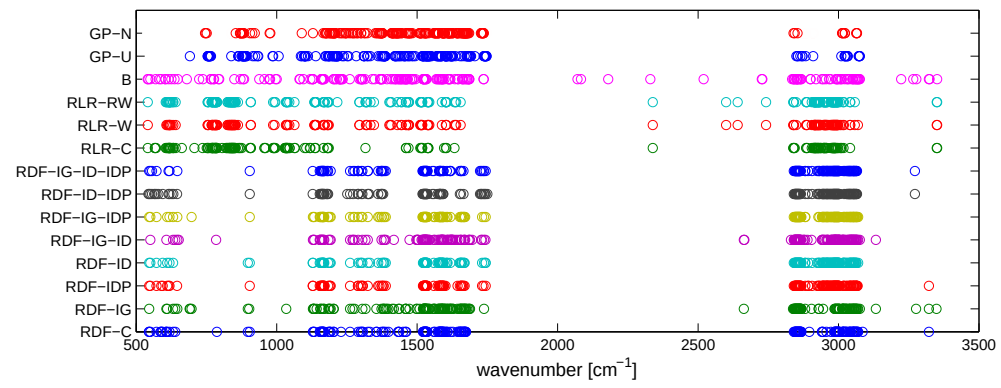
- 6707 Raman spectra from 10 bacterial species



- Quadratic interpolation + Median filtering + Variance normalization

Visual Results

- Visualize most important $d' = 200$ features
- Fingerprint region and high-wavenumber region



Classification Results

- Using GP classifier with Squared Exponential Kernel [3]:

$$\kappa_{\nu_0,\nu_1}(\mathbf{x}, \mathbf{x}') = \nu_0 \cdot \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\nu_1}\right)$$

- Independent dataset (comprising 299 spectra of 7/10 species)
- Measuring selected subsets ($d' = 200$) via prediction accuracy

methods	B.sph.	B.sub.	E.coli	M.lut.	M.lyl.	S.coh.	S.epi.	ARR	ORR
B	4	8	84	51	10	15	26	83.3	90.0
RDF-C	3	8	84	50	10	15	26	82.0	89.1
RDF-IG	0	8	83	51	10	17	26	80.4	88.6
RDF-ID	10	7	83	49	10	21	26	89.9	93.6
RDF-IDP	7	8	83	49	10	23	26	90.0	93.6
RDF-IG-ID	10	8	84	51	10	23	26	93.6	96.4
RDF-IG-IDP	7	7	83	49	10	23	25	87.7	92.7
RDF-ID-IDP	6	8	84	49	10	23	26	89.2	93.6
RDF-IG-ID-IDP	8	7	82	50	10	23	26	89.3	93.6
RLR-C	13	8	84	45	10	16	26	90.7	91.8
RLR-W	13	8	84	49	10	16	27	92.4	94.1
RLR-RW	13	8	84	48	10	18	27	93.3	94.6
GP-U	14	8	83	51	9	23	26	**95.8**	**97.8**
GP-N	11	8	84	51	9	25	27	94.8	97.7
ALL	0	8	82	51	8	24	27	*81.9*	*90.9*
#spectra	15	8	84	51	10	25	27		

Conclusion

- Observed relevant wavenumbers follow spectrochemical reasoning
- Gaussian Process classifier + ARD achieves best results
- TODOs: Find d' automatically

References

[1] 3Dchem.com: Thujone. http://www.3dchem.com/molecules.asp?ID=142 (2007)

[2] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press (2005)

[3] Kemmler, M., Denzler, J., Rösch, P., Popp, J.: Classification of microorganisms via raman spectroscopy using gaussian processes. In: DAGM. (2010)