Support Vector Regression for estimation of biophysical parameters and the identification of relevant feature subsets

Benjamin Jakimow
Sebastian van der Linden
Andreas Rabe
Patrick Hostert
Introduction: Support Vector Machines (SVM)

- avoid assumptions on statistical properties of input data
- able to handle hyperspectral collinearity
- often used for data classification


Only few studies used support vector regression
Introduction: Support Vector Machines (SVM)

SV Classification

\( x_1 \)
\( x_2 \)

divides training labels into two classes

SV Regression (SVR)

\( y \)
\( x \)

function estimates labels
Introduction: Support Vector Regression (SVR)

Chlorophyll estimations with an increasing number of training pixels

- Input data: simulated EnMAP scene with 220 features
- 50 repetitions of SVR per number of training pixel, 45 validation pixel
- each training & independent validation set was random sampled
Objectives & Questions

- How many features do we really need?
- Reduction of feature space, e.g. to decrease computational costs
- Find spectral features that are most relevant for SVR

- Use of Forward Features Selection
- Wrapper approach (Kohavi & John 1997), imageSVM (HU Berlin)
Methods: Forward Feature Selection using wrapper approach

- Ranking by explanatory power a single feature gives to the set of features that was ranked before
- detection of interrelationships vs. computational costs
Datasets

EnMAP Scene, 220 features used, simulated data

- Chlorophyll A+B [µg/cm²]
- 962231 labeled pixels
- 30 x 30 m ground resolution

R: 828 G: 1654 B: 661 nm
blue to red = low to high values

HyMap, 126 features used, “real world data”

- Gravimetric soil moisture
- 1806 labeled pixels
- 4 x 4m ground resolution

R: 830 G: 1652 B: 663 nm
blue to red = low to high values
Results: Error of estimations during forward selection

Chlorophyll A+B from simulated EnMAP data

- 30 repetitions of forward selection
- Each repetition on new sampled training and validation sets
- 360 training pixels, 90 validation pixels
Results: Error of estimations during forward selection (II)

Gravimetric Soil Moisture from HyMap data

- 100 repetitions of forward selection
- each repetition on new sampled training and validation sets
- 100 training pixels, 60 validation pixels
Results: Feature ranking

Chlorophyll data: 2D-Histogram of ranks for each feature

30x forward selections on Chlorophyll data

Feature index (in spectral order)
Results: A simple feature score (I)

Basic Idea:

• combine results of all forward selections \(\rightarrow\) generalize them
• a single feature is as more relevant the more it was allocated to a low/good rank

For each single feature:

1. weight the relative frequencies of each rank
2. add up the weighted frequencies = score value
Results: A simple feature score (II)

EnMAP: Chlorophyll AB

HyMap: Gravimetric Soil Moisture

linear weighting
$r_m = 20$
1/rank weighting
Conclusions

- SVR showed good results, $r^2 > 0.9$ using more than 30 training pixels
- Variations of errors & rankings due to sampling of training data
- Spectral feature space can be reduced by around half of all features / for both datasets ~30-50 features were sufficient
- Feature score
  - Allows to generalize rankings from repeated forward selections
  - Highlights most relevant spectral regions
  - Regions with low scores not necessarily irrelevant!
Acknowledgements

Thank you very much for your attention!

- Sören Haubrock & Sebastian Schmidtlein for data and advice
- VISTA and GFZ for simulated data EnMAP data
- Funding was provided by the Space Agency of the German Aerospace Center (Deutsches Zentrum für Luft- und Raumfahrt e.V.) with funds of the Federal Ministry of Economics and Technology

Contact:
benjamin.jakimow@geo.hu-berlin.de
Datasets: I. synthetic EnMAP Scene „Alpine Foreland“
Datasets: II. HyMap Scene „Welzow“

CIR

Clay
Dune
Tertiary
Quartery

GSM [%]

Quartery with Fe crust
Quartery with salt crust
Quartery with reed
Accuracy Assessment – CAB [µg/cm² | µg/l]

- 45 VPx
- 9 to 405 TPx
- 50x SVR per #TPx
Accuracy Assessment – GSM [%]

- 6 validation pixel
- 6 to 408 training pixel
- 100x SVR per per number of training pixel
Results: Error of estimations during forward selection

Chlorophyll A+B

- 30 repeated forward selection on one set of 45 training & 90 validation pixels
- variance of errors / accuracies due to model internal cross-validation
Results: Error of estimations during forward selection

Chlorophyll A+B

- 30 repeated forward selections, each on new sampled sets of 45 training & 90 validation pixels
- variance of errors / accuracies due to model internal cross-validation
Feature score values
Feature score values
Statistics dataset gravimetric soil moisture

[Bar chart showing CAB (µg/cm²) percentages for different categories: total dataset, coniferous forest, meadow, cereals, maize, rape seed, deciduous forest, natural vegetation, and other crops. Each category is represented by a bar with a percentage label.]
Statistics dataset gravimetric soil moisture
SVR Specifications

- imageSVM Humboldt-Universität zu Berlin (EnMAP Box)
- uses ENVI file format
- input values scaled to [0, 1]
- \( \varepsilon \)-SVR, RBF-Kernel
- \( \varepsilon = 0 \)
- internal grid-search with 3-fold cross-validation
- Chlorophyll: \( \gamma = [10^{-2}, 10^2], C=[10^1, 10^5] \)
- Gravimetric Soil Moisture: \( \gamma = [10^{-4}, 10^2], C=[10^0, 10^6] \)
The feature score

1. Relative frequency of rank

$$\varphi_{rel}(f_i, r_j) = \frac{1}{d} \sum_{i=1}^{d} f_i = r_j$$

- $f_i = \text{feature } i$
- $r_j = \text{rank } j$
- $d = \text{dimensions / total number of features}$

2. Weighing function

$$w(r_j) = \begin{cases} \frac{1}{r_j} & \text{1/rank} \\ 1 - \frac{1}{r_m} & \text{linear, } r_m = \text{max. rank of interest} \\ 0 & \text{if } j > m \end{cases}$$

$$w(r_j) = 0 \text{ if } j > m$$

- weights for all ranks $> r_m$ are zero

3. Score for feature $i$

$$\text{score}(f_i) = \sum_{j=1}^{d} \varphi_{rel}(f_i, r_j) w(r_j)$$
Table 4.4.: Comparison between SVR and PLSR based on 20 random sampled pairs of training- and validation sets. * denotes significant differences between averages (EV: $l_t = 20$, $l_v = 10$, GSM: $l_t = 240$, $l_v = 120$, avg = averaged values, p50 = median, sdev = standard deviation)

<table>
<thead>
<tr>
<th></th>
<th>MAE</th>
<th></th>
<th>RMSE</th>
<th></th>
<th>r2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SVR</td>
<td>PLSR</td>
<td>SVR</td>
<td>PLSR</td>
<td></td>
</tr>
<tr>
<td>$log(GSM)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$r_u^2 = 0.89$</td>
</tr>
<tr>
<td>avg</td>
<td>0.1097*</td>
<td>0.2564</td>
<td>0.1635*</td>
<td>0.3598</td>
<td>0.90</td>
</tr>
<tr>
<td>p50</td>
<td>0.1080</td>
<td>0.2532</td>
<td>0.1603</td>
<td>0.3571</td>
<td>0.91</td>
</tr>
<tr>
<td>sdev</td>
<td>0.0139</td>
<td>0.0251</td>
<td>0.0271</td>
<td>0.0364</td>
<td>0.03</td>
</tr>
<tr>
<td>$log(EV F)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>avg</td>
<td>0.0147*</td>
<td>0.0172</td>
<td>0.0191*</td>
<td>0.0209</td>
<td>0.39*</td>
</tr>
<tr>
<td>p50</td>
<td>0.0146</td>
<td>0.0165</td>
<td>0.0207</td>
<td>0.0211</td>
<td>0.30</td>
</tr>
<tr>
<td>sdev</td>
<td>0.0033</td>
<td>0.0042</td>
<td>0.0044</td>
<td>0.0045</td>
<td>0.21</td>
</tr>
<tr>
<td>$sqrt(EV N)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>avg</td>
<td>0.1349</td>
<td>0.1193</td>
<td>0.1629</td>
<td>0.1441</td>
<td>0.82</td>
</tr>
<tr>
<td>p50</td>
<td>0.1255</td>
<td>0.1138</td>
<td>0.1497</td>
<td>0.1414</td>
<td>0.85</td>
</tr>
<tr>
<td>sdev</td>
<td>0.0389</td>
<td>0.0261</td>
<td>0.0474</td>
<td>0.0304</td>
<td>0.12</td>
</tr>
<tr>
<td>$EV R$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>avg</td>
<td>0.8996*</td>
<td>0.6833</td>
<td>1.1507*</td>
<td>0.8466</td>
<td>0.78*</td>
</tr>
<tr>
<td>p50</td>
<td>0.9181</td>
<td>0.6957</td>
<td>1.1317</td>
<td>0.8365</td>
<td>0.79</td>
</tr>
<tr>
<td>sdev</td>
<td>0.1751</td>
<td>0.1914</td>
<td>0.2679</td>
<td>0.2208</td>
<td>0.11</td>
</tr>
</tbody>
</table>