Главные редакторы: Виктор Г. Сычёв и Лотар Мюллер

НОВЫЕ МЕТОДЫ И РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ ЛАНДШАФТОВ В ЕВРОПЕ, ЦЕНТРАЛЬНОЙ АЗИИ И СИБИРИ

Том III Мониторинг и моделирование ландшафтов

В содружестве с Академией почвенного плодородия Митчерлиха (МИТАК), Паулиненауэз, Германия

Москва 2018
Main editors: Viktor G. Sychev and Lothar Mueller

NOVEL METHODS AND RESULTS OF LANDSCAPE RESEARCH IN EUROPE, CENTRAL ASIA AND SIBERIA

Monograph in 5 Volumes

Vol. III Landscape Monitoring and Modelling

With friendly support of the Mitscherlich Academy for Soil Fertility (MITAK), Paulinenaue, Germany

Moscow 2018

Team of authors and editors under the guidance of: Vladimir A. Romanenkov (Moscow), Askhad Kh. Sheudzhen (Krasnodar), Lothar Mueller (Muencheberg)

Main editors: Lothar Mueller (Leibniz Centre for Agricultural Landscape Research, Muencheberg, Germany) and Viktor G. Sychev (All-Russian Research Institute of Agrochemistry named after D.N. Pryanishnikov, Moscow, Russia)

This monograph shall inform you about up to date methodologies and recent results in landscape research. It is intended as a guide for researchers, teachers, students, decision makers, stakeholders interested in the topic of landscape science and related disciplines. It provides information basis for decision makers at various levels, from local up to international decision bodies, representing the top level of landscape science in a very short form.

Authors are responsible for the content of their chapters. Neither the authors nor the editors can accept any legal responsibility for any errors or omissions that may be made. The editors make no warranty, express or implied, with respect to the material contained herein.

ISSN 978-5-9238-0246-7
ISBN 978-5-9238-0249-8 (Том 3)
DOI 10.25680/1490.2018.71.71.003
Chapter III/68: THE USE OF HYPER SPECTRAL DATA FOR CROP CLASSIFICATION USING MACHINE LEARNING

Глава III/68: Использование гипер спектральных данных для классификации культур с использованием машинного обучения

Ralf Wieland*; Batunacun

DOI 10.25680/1530.2018.89.32.261
*Email: rwieland@zalf.de
Leibniz Centre for Agricultural Landscape Research (ZALF), Eberswalder Str. 84, 15374 Muencheberg, Germany

ABSTRACT. A set of ground truth hyper spectral data was prepared and stored as open data using a GitHub repository. From this data set a subset was selected according to the satellites bands of Sentinel 2. Five popular machine learning algorithms (adaboost, decision tree, logistic regression, random forest and support vector machine) were selected to train plant classification models. The models were evaluated using the statistical measures: precision, recall and f1score. Additionally, some pros and cons of the used machine learning methods were discussed. The support vector machine and the logistic regression produced the highest statistical measures. The implementation is presented as open source software on the same GitHub repository. The aim of this article is to introduce in machine learning using of hyper spectral data. Additionally, the paper enables the user to experiment independently with the models, to combine them, to compile the data differently, etc.

KEYWORDS: hyper spectral data, satellite based data, Sentinel2, machine learning, Python programming

INTRODUCTION

Remote sensing based on airborne and spaceborne platforms provides valuable data for mapping, environment monitoring, disaster management and civil and military intelligence. In the example the remote sensing is used to determine agricultural crops from satellite images of different spectral bands. However, the satellites have some limitations. Firstly, the spectrum is limited to a set of spectral bands ("Sentinel 2" [1]) has 13 bands), and has specific resolution (10m, 20m, 60m).

The actual Landsat [2] has 12 spectral bands with two resolutions (30m, 100m). Secondly, more important is the satellite data can be only used if the sky is cloud free. The restriction of the spectral bands and the time between two overflights are limiting factors in the use of satellite images.

Hyper spectral data provide larger spectral range and can be used without the limitations of real satellite images. So is the spectra not divided into small bands and the sampling do not depend on the overflight of the satellite.

The presented hyper spectral data were sampled using a field spectrometer: "FieldSpec Pro JR (A 110080) of the American company "Analytical Spectral Devices". The hyper spectral data comprises a spectral range from 350nm to 2500nm with a spectral resolution of 1nm. The modeler can use the complete spectra or select some bands for modeling. This opens new options: the modeler can investigate how available satellites images can be used by selecting the bands of a satellite, he can combine data from
different satellites, the modeler can estimate the most informative spectral range for crop classification, etc.

The third part of the use of spectral data is machine learning. For a crop classification based on satellite images the modeler has to build a model which estimates the crop. The hyper spectral data can be used for training of model using machine learning algorithms. The open source software "scikit" [3] provide a set of well tested machine learning algorithms implemented in Python. Additionally, Scikit includes methods for preprocessing the data and of the statistical analysis of the results which are really helpful for implementation of machine learning programs (Scikit acts as Python module inside a Python program which has to be programmed by the modeler).

DATA AND METHODS

The data were sampled next to Muencheberg(Fig.1), a small town located 50km east of Berlin using the FieldSpec Pro JR (A 110080) in the season of 2002. The data are stored in Excel for a set of selected agricultural crops: (Alfalfa, Cocks Grass, Lupin, Pea, Potato, Silo Maize, Triticale, Winter Barley, Winter Rape, Winter Rye, Winter Wheat). Each file contains the hyper spectral data for different plots and different sensing times. The Python script "gen_sample_all.py" reads the data and transform it so that it can be used for training.

As machine learning methods were used a set of popular algorithms: Adaboost [4], Decision Tree (DTree) [5], Logistic Regression (LogReg) [6], Random Forest [7] and the Support Vector Machine (SVM) [8]. This set can be easily expanded using scikit.

The machine learning algorithms implemented as Python program (Adaboost.py, DTree.py, LogReg.py, Random_Forest.py and SVM.py) uses the "gen_sample_all.py" to read the data. Each model can run autonomous and depends on Python, the numeric package numpy and scikit which should be installed. The programs run in an terminal and write the output back into the terminal. The user can modify the inputs, the hyper parameters (to control the training) and the output statistics according his wishes.

The data set was randomly split into a part used for training and a part used for testing (training:testing=67:33). The testing part was used to produce the confusion matrix and the classification statistics based on precision and recall. The precision is defined as:

\[
\text{precision} = \frac{TP}{TP + FP}
\]

(1)

The recall is defined as:

\[
\text{recall} = \frac{TP}{TP + FN}
\]

(2)
and f1score used precision and recall:

\[
f1score = 2 * \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]  

(3)

Tp means true positive (the true value was positive and the predicted values was positive too), Fp means false positive (true value is negative, predicted values was positive), Fn means false negative (the true value is positive the predicted value was negative).

**RESULTS AND COMPARISON OF THE ALGORITHMS**

In the example the crops: (Silo Maize, Potato, Winter Rape, Winter Wheat, Winter Rye) was used to build the model. The user can select the time of the sample by an index \( \in 0..4 \), the medium sample (index=2) is used by default.

The table 1 summarizes the training quality using a confusion matrix for the DTree.

The table 1 shows some differences from the main diagonal, that means training errors. It should be remarked that the result based on one selected training run and should repeated for statistical valid results.

**Table 1:** Confusion matrix Dtree, X=true input, Y=simulated

<table>
<thead>
<tr>
<th>X</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

For one training run a table 2 containing the precision, the recall and the f1score can be made from the confusion matrix:

**Table 2:** RTree, precision, recall, f1score for on training run

<table>
<thead>
<tr>
<th>Name</th>
<th>Precision</th>
<th>Recall</th>
<th>f1score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winter Rape</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>Winter Rye</td>
<td>1.00</td>
<td>0.75</td>
<td>0.86</td>
</tr>
<tr>
<td>Winter Wheat</td>
<td>0.89</td>
<td>1.00</td>
<td>0.94</td>
</tr>
<tr>
<td>Silo Maize</td>
<td>0.86</td>
<td>1.00</td>
<td>0.92</td>
</tr>
<tr>
<td>Potato</td>
<td>1.00</td>
<td>0.67</td>
<td>0.80</td>
</tr>
<tr>
<td>total</td>
<td>0.91</td>
<td>0.90</td>
<td>0.90</td>
</tr>
</tbody>
</table>

To overcome the variance of different runs a cross validation (with cv=10) of the training was applied. The table 3 summarizes the results of all used machine learning algorithms:

**Table 3:** Result of cross validation c=10

<table>
<thead>
<tr>
<th>Name</th>
<th>Precision mean</th>
<th>Precision std</th>
<th>Recall mean</th>
<th>Recall std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaboos</td>
<td>0.8844</td>
<td>0.078</td>
<td>0.887</td>
<td>0.105</td>
</tr>
<tr>
<td>DTree</td>
<td>0.733</td>
<td>0.182</td>
<td>0.742</td>
<td>0.168</td>
</tr>
<tr>
<td>LogReg</td>
<td>0.921</td>
<td>0.071</td>
<td>0.911</td>
<td>0.085</td>
</tr>
<tr>
<td>Random_Forest</td>
<td>0.908</td>
<td>0.092</td>
<td>0.905</td>
<td>0.072</td>
</tr>
<tr>
<td>SVM</td>
<td>9.926</td>
<td>0.067</td>
<td>0.921</td>
<td>0.092</td>
</tr>
</tbody>
</table>

The SVM was the best training algorithm in this example but all the other did a good job too. The training time was so short that no significant different could be seen. For a larger data set or combination of different training algorithm the training time could be an important value.
All algorithms need so called 'hyper parameter' which controls the training. SVM used: kernel='rbf', C=1000, gamma=0.1 as hyper parameters. These parameters must be updated (optimized using grid search) during the model construction procedure. This can be a time consuming process.

REMARKS ABOUT MACHINE LEARNING
Before a machine learning procedure can be applied to the data set, the data set itself has to be prepared. In the example the hyper spectral measurement has to calibrated, the data has to be corrected, for example by removing the water bands and interpolation the gaps using a polynomial of fourth order. In general the modeler has to take care about the data, remove outliers, fill gaps etc. the scikit can support the modeler but the modeler is responsible for the data integrity.

The provided hyper spectral data were sampled at different dates. This can be used to train models for each sample date and combine the models. At different dates the crops are in different grow states. The knowledge about the growth of the plants can help to get better classification results for selected crops. Such a sequential model refinement (using models with different f1scores for the crops) can be helpful in many ways. For example an satellite image made early in the season can be used for a first estimation of the crop distribution which can be refined using later satellite images. Another idea is to include the a priory knowledge about well known crop distribution from the last years can help to improve the models accuracy [9].

The machine learning algorithms used in this paper are selected due to it popularity. For a real project each algorithm has its own specific which has to take into account. For example the DTee can be easily interpreted by the modeler which could be a reason to prefer it despite of its poor accuracy. A grid search to find appropriate hyper parameters can be a time consuming process. A random forest algorithm has a lot of hyper parameters (n_estimators, criterion, max_features, ...) but the default parameters produce often a satisfying result so the random forest can be used without adapting the hyper parameters in many cases. The SVM on the other hand has less hyper parameters but without an fine tuning it should not be used.

The presented data set and the open source software can be used as a basis for training in machine learning. You can expand the software according your interests. You could combine different models, could select different sample data, could combine models at different sample dates, etc. Even when the spectral data is not portable to your region you can check if a field spectrometer could be useful for you. The data and the software can be found at: https://github.com/Ralf3/Remote-Sensing.

CONCLUSIONS
1. Hyper spectral data is high precision ground truth data.
2. Data needs to be preprocessed before they can be used for machine learning.
3. Machine learning methods differ in accuracy, the number of hyper parameters and many others and should be selected with care.
4. Only the open science idea: to combine data and methods as open source give the user the freedom that he needs to make science.

ACKNOWLEDGEMENT
We want to thank our former Colleague Dr. Bernd Zbell for the data collection and we want to thank the Python community providing the the wonderful scikit-learn. This paper was funded by the Federal Ministry of Food and Agriculture and the Ministry for Science, Research and Cultures of the State of Brandenburg (MWFK), Germany.

REFERENCES
[1]http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Sentinel-2
Глава III/69: МЕТОДЫ ПРОГНОЗА УРОЖАЙНОСТИ ОЗИМОЙ ПШЕНИЦЫ В БАССЕЙНЕ Р. ОКИ С ИСПОЛЬЗОВАНИЕМ РЕЛЬЕФА, КЛИМАТА И ПОЧВ

Chapter III/69: Methods for Forecast of Winter Wheat Yield in the Oka Basin Using Relief, Climate and Soil Data

Петр А. Шарый*1, Лариса С. Шаря2, Ольга В. Рухович3

DOI 10.25680/2042.2018.37.32.262
*Эл. Почта: p_shary@mail.ru
1. Институт физико-химических и биологических проблем почвоведения РАН, ул. Институтская д.2, 142290 г. Пущино Московской области, Россия
2. Институт экологии Волжского бассейна РАН, ул. Комзина д.10, 445003 г. Тольятти Самарской области, Россия
3. Всероссийский научно-исследовательский институт агрохимии РАН, ул. Прянишникова д.31а, 127550 г. Москва, Россия

Резюме. Хотя в глобальном изменении климата с 1998 г. наблюдается пауза, ее нет в России; поэтому для России актуален прогноз урожайности сельскохозяйственных культур. Анализ одновременного влияния природных факторов (климата, почв и рельефа) на урожайность озимой пшеницы в Нечерноземной области для современного периода выявил, что ведущим является микроклимат, описываемый освещенностью мезосклонов, и который не учитывался ранее в исследованиях. Построена прогнозная модель урожайности озимой пшеницы на 2050 г. с использованием прогнозной климатической модели Е GISS NASA. Для учета важных для урожаев хронологически последовательных климатических факторов, запрещенных к введению в регрессионные модели из-за тесной связи между ними, введен новый фактор – климатический инвариант. Пространственная гетерогенность прогнозируемой урожайности существенна и определена не только климатическими изменениями, но и освещенностью мезосклонов: на более освещенных склонах прогнозируется ее возрастание в 1.6–1.9 раз по сравнению с менее освещенными. Эти различия важны при планировании посевных площадей к 2050 г.

Abstract. Although some hiatus is observed in global climate change since 1998, it is absent in Russia; hence a forecast of crop production is actual for Russia. Our analysis of influence of natural factors (climate, soils, and topography) to winter wheat in Non-Chernozem region of Russia showed that the impact of microclimate is a leading factor and was not taken into account in previous studies. We have developed a forecast model for winter wheat for 2050 using climate projection using the E GISS model of the NASA. To take into account important for crop yields chronologically sequential climatic factors into regression models, we suggest a new factor – climatic invariant. The spatial heterogeneity of crop yield is essential and is defined not only by the climate change, but also by slope insolation: crop yields on more insolated slopes are 1.6–1.9 times greater than on other slopes. These distinctions are important for planning crop areas in 2050.

Ключевые слова: озимая пшеница, прогноз, рельеф, регрессия
Keywords: winter wheat, forecast, topography, regression

Введение
Необходимость адаптации сельского хозяйства России связана с глобальным изменением климата: при глобальном росте температуры 0.17 °С/10 лет за 1976–2015 гг., за тот же период рост температуры в России составил 0.45 °С/10 лет [1]. Хотя с 1998 года наблюдается пауза в глобальном