Haensel AMS
Advanced Mathematical Solutions

Case Study

yield pop

Data-Mining on Agriculture Trial Data

Summary

Task

- develop a clustering algorithm to identify very similar performing crop varieties
- estimate the impact of different treatments and general conditions on the varieties' performance per type group

Solution

- develop tailored classification and prediction algorithms
- built R and Python programs able perform the clustering and impact prediction within fast computation times
- visualization of results with different perspectives

Result

- Yield Pop is able to identify similar crop types, even when they are never tested in the same trial
- the clustering results provides a schema for different grouping cutoff points
- attribute impacts on crop performance are visualized for further testing and development focus

Organization:

Yield Pop is an internet startup in the agriculture space and its mission is to make farm decisions easier. They help farmers and industry specialists to improve practices through greater transparency about product performance, e.g. chemicals and seeds, and more relevant knowledge exchange.

Challenge:

Yield Pop assembled a large database of US crop trial data, containing lots of information about the varieties/ types tested. Attributes about the trials included location, time, soil characteristics, chemicals used, and resulting yield performance, etc. The first and main question was: Given this data can we find varieties which perform similarly under similar conditions? The dataset contains thousands of products, but as any individual trial only tests 30-50 products, products rarely share more than a handful of trials. The second question was to derive an approach to estimate the attribute/treatment impact for different individual types and or groups. E.g. if group XYZ is planted on soil with pH-value greater 7, then certain treatments should be given to increase yields.

"We had assembled a giant dataset and we knew it held a treasure trove of information. The challenge was to figure out how to unpack it and generate actionable insights for our customers. Hansel AMS intuitively understood the challenge and took on the work with rigorous professionalism."
— Matthew Perkins, Co-founder Yield Pop

Solution:

The challenge in this assignment was of course the large number of treatment and environmental variables, which are correlated and have indirect effects. Also the dataset is sparse in terms observed common trial of type pairs. And finally the data is very noisy because not all attributes which effect the growing of a plant are recorded in the dataset. The first task was to bring the attribute values into a common structure and scale for comparison. Then we started to test some clustering algorithms on the data, with reasonable results for initial approaches but the cluster distances where still considerably high. We further improved the algorithms by adding statistical similarity measures to the clustering algorithms. Weather information had not yet been considered in the clustering and therefore we downloaded the historic daily weather information of the stations closest to the trial locations from the US weather and climate agency.

Haensel AMS
Advanced Mathematical Solutions

Case Study

Data-Mining on Agriculture Trial Data
developed for

Yield Pop

November 2013
till January 2014

Haensel AMS
www.haensel-ams.com

The additional information reduced the noise in the data significantly and provided very good clustering results. For the impact estimation of attributes we choose to work with decision tree classification methods to identify which attributes have significant impact and also combinations of them. Further developments will focus on the quantification of these impacts and for that we are planning to use advanced regression models.

"The main challenge was to combine all the different given data types into one model, which captures also the indirect attribute relations and influences. We choose to model this project mainly with R and Python programs to enable the use of most current developments in machine learning and statistical computing."
— Alwin Haensel, PhD, Founder Haensel AMS

Result:
Yield Pop obtained a clear picture of the clustering possibilities and results. With the customized algorithms it is straight forward to identify similar performing varieties. In our result reports we highlighted all development steps, which made it easy for Yield Pop to understand how the algorithms worked. We also started the estimation of attribute impacts by identifying group specific major impact attributes, i.e. treatments or environmental conditions. Further developments will focus on the quantification of these impacts.

"The insights which Alwin and his team generated for us go a long way towards helping us unlock the value of the dataset, both for us and for our customers. Alwin was great at simplifying the results of his data-mining such that we could understand the implications and start to think about how to integrate the findings into our platform."
— Alex Wimbush, Co-founder Yield Pop

"This data-mining project was so far a very interesting and challenging assignment. It was fascinating to see the results developing while our algorithms where able to explain more and more of the underlying complexity."
— Alwin Haensel, PhD, Founder Haensel AMS