Sigrow

Optimizing Plant Growth

Riemer Dijkstra | 12223255
Dionne Gantzert | 12058866
Otto Márton | 12217735
Mees Meester | 12217255
Johannes Roelink | 11903260

Tweedejaarsproject BSc KI / University of Amsterdam
28.06.2020

Table of Contents

1 | Introduction

One of the goals of horticultural production in greenhouses is to increase the sustainable income of the grower. The investment costs for conventional plant production as well as labour and energy costs are much lower compared to the investment costs for greenhouses. This can only be balanced out with a better utilization of the yielding potential of plants, higher labour productivity and higher energy efficiency (*Tantau, 1991*).
Another goal of horticultural production in greenhouses is related to the global population rapidly increasing, together with the demand for healthy, fresh food. The horticulture industry can play an important role in providing food, but encounters difficulties finding skilled staff to manage crop production (*Hemming et al., 2019*).
In the competitive horticulture industry, small improvements can make or break the competitive advantage of the yield. In order to stay 'ahead of the curve', calculated decisions need to be made about the greenhouse climates (*Sigrow, n.d.*).

Sigrow is a company in the Netherlands which provides state of the art sensor technology tailored to greenhouses. These sensors make it possible for Sigrow's clients to stay ahead of the curve and make those calculated decisions. Since the horticulture industry is very competitive, growers try to find any competitive advantage they can. This means that most businesses that utilize Artificial Intelligence (AI) closely guard their results and techniques and rarely offer them publicly for free. This means that Sigrow has little knowledge about AI and that is why a collaboration of the two fields was needed. Our group of AI students have been given the opportunity to establish this combination with Sigrow.
Over the years, Sigrow has gathered information about the environment in the greenhouses, such as the temperature, air humidity, light intensity and amount of carbon dioxide (CO2) in the air. These four features are the most important factors for plants to grow, apart from water and nutrition (*Hemming et al., 2019*). The task Sigrow had for our group was to optimize plant growth by optimizing the greenhouse climate using AI. This task was dubbed the 'Sigrow Challenge'. For the completion of this challenge, the earlier mentioned greenhouse information was combined in order to create a dataset.

When Sigrow provided this dataset, there were two significant problems:

1. The data lacked information about the plants.
2. The amount of data was scarce.

Sigrow has only collected data about the environment of the plants, but not the plants themselves. The data we received from Sigrow did not contain any information about which kind of plant it was, nor the height, stem diameter or weight (1). In addition to this, the amount of data was so scarce that it was impossible to split the dataset into a training and test set (2). The importance of useful and substantial data is explained in the next section of the text.

Because of the two problems, a new challenge was set up, consisting of two parts.
The first part involved a recommendation. This recommendation had to meet a number of requirements. First, it had to contain a comprehensive analysis about the measurements Sigrow is currently taking. This includes advice about which measurements need to be done and why. This part of the recommendation is meant to inform Sigrow about the importance of useful data and how to acquire that. Second, it had to contain a comprehensive overview of possible feature optimization techniques which Sigrow could use for future work. At last, the recommendation had to contain an understandable explanation about several machine learning techniques.
Besides the recommendation, Sigrow also asked to attempt several discussed implementations. This is the second part of the proposed challenge. This part is similar to the initial Sigrow Challenge. Since the data is not usable, the results would be unreliable. However, Sigrow could utilize the results of attempted implementations in the future. A framework would already be of great help for the company. By providing Sigrow a recommendation on how to optimize the plant growth using machine learning and implementing these techniques for them, we hope Sigrow can optimize the plant growth in the future and stay ahead of the curve.

2 | Proposed Solutions

As stated in the introduction, the provided dataset for the Sigrow Challenge was lackluster. The dataset lacked information about the wellbeing of the plants and the overall data was scarce. AI algorithms are used to unlock the concealed information available in data (*Desik, 2019*). With incomplete data, suboptimal results will be obtained. The solution to the data problem is discussed in this section. Hereafter, the final product for Sigrow is set out, consisting of a recommendation and an implementation.

2.1 Solution to Data Problem

In order to surpass the hurdle of incomplete data, a new dataset had to be obtained. This dataset had to meet two requirements. First, some features of this external dataset had to be in accordance with features of Sigrow's dataset. This way, the final product is applicable to Sigrow's data and not just to the external data. Secondly, the dataset had to contain a sufficient amount of information so that it could be split into a training and test set.
The external dataset originated from the Autonomous Greenhouse Challenge, a challenge by Wageningen University & Research[1]. The goal of the challenge was to produce a certain crop within six months inside a greenhouse remotely. Teams with people from all around the world tried to tackle this challenge.
The similarities between the Sigrow Challenge and the Autonomous Greenhouse Challenge mainly consist of optimizing plant growth using AI techniques. A limited dataset from this challenge was made available. In collaboration with Sigrow, this set has been chosen as the replacement for Sigrow's dataset.
It contained information about cucumbers and contained features such as temperature, air humidity and the amount of CO2 in the air, all similar to Sigrow's features. Although the new set was not substantial either, the data problem was practically dealt with.

2.2 Proposed Solution to the updated Sigrow Challenge

The final product had to consist of a recommendation of AI techniques for optimizing plant growth, and an implementation of such a technique. So half of the task given by Sigrow was creating a recommendation. This recommendation can be subdivided further into the following parts: recommended methods of measuring and recommended AI techniques. Since Sigrow did not provide a useful dataset initially, it was necessary to advise the company on this matter. This advice could be used in the future to establish a more reliable dataset for AI purposes. Furthermore, Sigrow was interested in learning about AI techniques for growth optimization. This knowledge would also be useful for

[1] Autonomousgreenhouses. (n.d.). Autonomous Greenhouses International Challenge 2019. Retrieved on June 23rd, from http://www.autonomousgreenhouses.com/

future AI projects. It was decided to combine the two parts in one document that serves as a clear, overall recommendation. The following two subsections discuss this subject.

2.2.1 Recommendation of the Measurements

There are multiple features which could be measured to get more data regarding the plants. These features consist for example of the height, stem diameter, number of leaves or weight. The more features there are regarding the plants, the better the results.
To optimize the plant growth, output features are essential. If Sigrow can measure the aforementioned features, it is possible to use AI, which needs data to make accurate predictions. This means that not only the output features are important, but the amount of data is significant as well. The more data there is, the more reliable the predictions made by AI are. The key is to obtain sufficient data about the features of the plants themselves.

If Sigrow wants to optimize the plant growth in order to maximize the production, it is necessary to measure the production. Sigrow can use a similar approach as the Autonomous Greenhouse Challenge, in which the cucumber production is measured by dividing the cucumbers in three classes. The first class is A, which contains the cucumbers with a weight above 375 grams and has no defects. Class B contains the cucumbers with a weight between 300-374 grams or has some defects. Lastly, class C contains all the cucumbers with a weight below 300 grams. Of course, in order to successfully classify the plants, each kind of plant has its own threshold values. Below (figure 1) is an overview of the thresholds for the cucumber distribution. If Sigrow is able to document information about the features of greenhouse plants and information about the production, reliable AI initiatives will be possible.

Class	Weight	Defects
A	> 375 grams	None
B	300-374 grams	At most a few
C	< 300 grams	No limit

Figure 1: Cucumber distribution among three different classes, like with the Autonomous Greenhouse Challenge:

2.2.2 Recommendation of Artificial Intelligence Techniques

The second half of the recommendation contains possible AI methods for solving the Sigrow Challenge. The discussed methods are: Support Vector Regression (SVR), Polynomial Regression, Reinforcement Learning (RL), Deep Deterministic Policy Gradient (DDPG). These techniques were divided into two groups. The first group consisted of methods which can be used to estimate the optimal features values of plants after the process of cultivation. The other group consisted of techniques used for creating the optimal environment for plants during cultivation. SVR and Polynomial Regression have been recommended as methods to use for optimization after cultivation. DDPG was placed in the second group.

Each method from the groups is briefly introduced to give Sigrow a broad overview of the method. After that, the positive and negative aspects are listed. This is to ensure that Sigrow can compare methods easily. Ultimately, the relevance of the techniques, with respect to the new Sigrow Challenge, are discussed. These properties combined, formed the second part of the recommendation. Below is an overview of all the techniques.

Polynomial Regression

Regression uses the relationship between variables to find the best fit line that can be used to make predictions. Since plants do not grow linearly, a nonlinear model has to be fitted to the data. Otherwise the model will be under fitted and thus not plausible. Polynomial Regression is a relatively effortless method of modeling. It is a type of regression analysis, which is concerned with fitting functions on data points (figure 2). For this form, the chosen function is an nth degree polynomial. So by using known data points, a polynomial function can be plotted to reveal the relations between the points. From these relations, certain conclusions can be drawn. For instance, what the optimal value for a variable/parameter is. Polynomial Regression is a great technique to see the correlation between features and helps with the estimation of values between data points. Since the amount of data is scarce, the estimation of data is really valuable. Polynomial Regression is a rather easy way of modeling data. Not much computing power, advanced know-how and time is needed to gather results from this implementation. For Sigrow this could be a reasonable option. Polynomial Regression can acquire results with limited data. These results are unstable, but they can be useful nevertheless. Therefore it would be a solid pick for working with limited datasets, like Sigrow's.

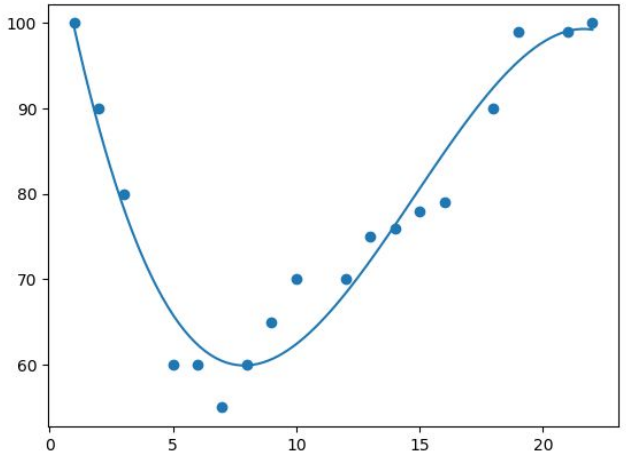

Figure 2: An example of a fitted polynomial, showing the relations between a handful of data points (source: w3schools, n.d.).

Support Vector Regression

Support Vector Regression (SVR) is a form of regression that is concerned with fitting a line through data points by dividing the points into two sections. The line is drawn in between these points with the two groups on opposite sides (figure 3). So essentially SVR decides on data points to group together and then plots a line on the imaginary border. Generally, SVR is used for categorization problems. Which seems logical with regard to the definition. However, the technique is also used for regression problems. This will be the case for Sigrow. So like the Polynomial Regression strategy, certain conclusions can be drawn from the final fit. Estimation of values that are not included in the dataset for instance. This is a great implementation for scarce data.

This technique is relatively easy to implement. It does not require a lot of data and runtime to acquire results. However, the results can be unstable or less informative. A regression approach to the Sigrow Challenge is a trade-off between difficulty and quality.

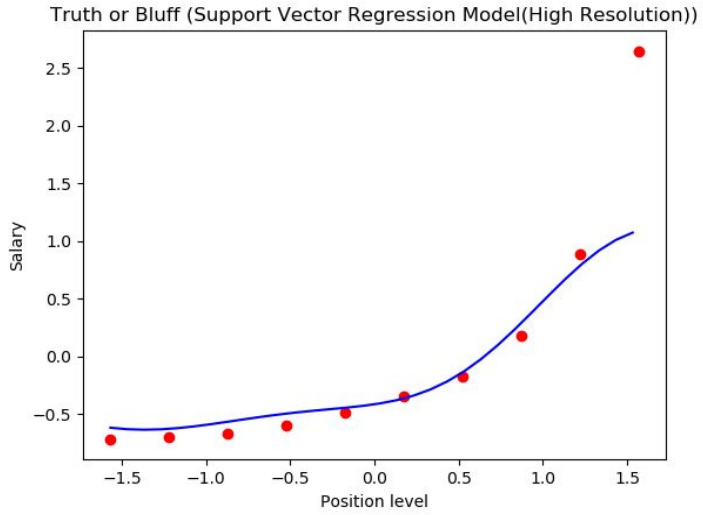

Figure 3: An example of a regression problem tackled with the use of Support Vector Regression (source: Sethi, 2020)

Reinforcement learning

Reinforcement Learning models the way humans learn from experience. Reinforcement Learning problems can be seen as an agent trying to find the best action in a given state. Inherent in this type of machine learning is that an agent is rewarded or penalized based on their actions. If an action or combination of actions leads to the target outcome, the agent is rewarded (reinforced).

The same idea can be used by Sigrow to optimize the growth of crops. The agent will be rewarded if its actions contribute to a higher plant growth and will be punished when its actions inhibit plant growth. Other factors can be introduced as well, like penalizing the agent when its actions require a lot of resources like energy and water. There are however a few obstacles that need to be overcome.

The first being the fact that any agent cannot directly influence the growth of the plants. It only has control over some properties of the environment of the crops. This creates some delay in the effects of the agents actions and the feedback it receives. This can be fixed by taking this delay into account when calculating the rewards.

The second obstacle is the amount of training data that is needed for such an algorithm. A team in the Autonomous Greenhouse Challenge used artificial data to train such an algorithm because the amount of data needed is inhibitively large.

Another obstacle is the fact that in the Sigrow Challenge the action space, that is the possible actions the agent can take, is continuous. This means that in any given situation there are theoretically infinite possible actions the agent can take. For example, the agent can set the amount of light, water and CO2 received by the plants to any number. In practice the options are limited to the precision of the equipment used. In any

case the amount of possible actions is still immense. Which means that it is computatively infeasible to compute the best possible action., which involves iterating through every possible action. The following technique introduces a solution to this obstacle.

Deep Deterministic Policy Gradient (DDPG)
DDPG is an algorithm that uses a learned policy to decide on actions in any given state. A policy maps the states enountered to policies taken in that state. This algorithm is an example of an off-policy Reinforcement Learning algorithm, which means that the algorithm learns the optimal policy concurrently but independently of the actions it should take in a state.
Normal reinforcement algorithms struggle with continuous action spaces because it is computationally impractical to calculate every possible action in a state. DDPG solves this problem by using a gradient based learning rule instead to approximate the maximum instead of exact calculations.
DDPG would be an excellent technique for learning the optimal environment for plant growth, assuming the changes reported about the measurements are made.

These techniques, along with the information above, are stated in the recommendation to Sigrow. The last item to discuss: the final implementation of our own.

2.2.3 Implementation Introduction
The second half of the Sigrow Challenge was to implement a technique mentioned in the recommendation. Together with the previously discussed first half, this forms the ultimate product.
The implementation of choice was Polynomial Regression. This was concluded to be the most feasible method for the given time period. A structured and interactive file has been made and is ready to be utilized by Sigrow. The file can be used to, interactively, load and read data of choice. This data can then be used for fitting a line with the use of Polynomial Regression. From the use of this technique, certain conclusions can be stated from the data. An example of which is to determine optimal feature values for the cultivation of greenhouse plants. The most important features were found by using random forests. A more elaborate and low-level explanation is found in the next section.

3 | Implementation

All of the implementations are programmed in Python, in a Jupyter Notebook. We tried multiple implementations and methods, and used Github to work on our implementations. All of us programmed and worked on code, a big part of the coding was understanding the data and seeing how it reacted to multiple methods. Due to the scarce amount of data we chose, in the end, to generate the results with a random forest and polynomial fits. The endresults can be found in this github repository: https://github.com/readmees/erudite2020.git. The README.md explains the repository, how to use it and the results of the models.

4 | Conclusion

The final product that was delivered was two fold. The first part consisted of researching and explaining AI techniques that might prove useful to Sigrow in the future. The techniques that were discussed in this report are Polynomial Regression, Support Vector Regression, Reinforcement Learning and Deep Deterministic Policy Gradient. These techniques themselves can be divided into two groups. The first group being the techniques that can be executed after a growth cycle of the crops. They can be used to determine in hindsight what were the best values for some selected features like humidity, light and CO2 levels. These techniques are Support Vector Regression and polynomial fitting. The second group of techniques includes Reinforcement Learning and Deep Deterministic Policy Gradient. These are techniques that can be used while the crops are growing to determine the best possible action one can take based on the current environment of the crops.
The second part of the final product was creating a proof of concept of a model to find the optimal values of the selected features. For this we analyzed the dataset using a Polynomial Regression model, which was made in a way so as to be easily adaptable to different datasets.

After Grow Cycle Techniques
The acquired datasets that we have now do not contain much information about the plants, which limits the amount of techniques that we could use. The two techniques we discussed for time series analysis were Polynomial Regression and Support Vector Regression. Polynomial Regression is a rather easy way of modeling data. Not much computing power, advanced know-how and time is needed to gather results from this implementation. Polynomial Regression can acquire results with limited data, which makes it a suitable solution for Sigrow. However, these results can be unstable because of the major influence of outliers, so Sigrow should take this into account.

Like Polynomial Regression, a Support Vector Regression solution to Sigrow's problem is relatively easy to implement. It does not require a lot of expertise or runtime to acquire results. However, the results can be unstable depending on the number and severity of the outliers and the results are less informative than the results of Polynomial Regression. A regression approach to the Sigrow Challenge is a trade-off of difficulty and quality. SVR methods can perform better with a lot of data.

Live Grow Cycle Techniques
The techniques discussed for a live model, which update themselves as the plants grow, were Deep Deterministic Policy Gradient and Reinforcement Learning. Reinforcement Learning is a very computationally expensive technique since significant computing power is needed. However, the main problem for Sigrow is the amount of training data that is needed. Deep Reinforcement Learning would be an excellent technique for learning the optimal environment for plant growth, assuming the changes reported about the measurements are made.

When evaluating everything together the most important conclusion is that data is one of the most important factors in machine learning and AI by extension. The main hurdle that was encountered was the lack of usable data. Future research should focus on obtaining reliable data.

5 | Acknowledgements

This project would not be possible without the unfaltering support of some key individuals. First and foremost we want to thank our teaching assistant Bob Leijnse for aiding and supporting us throughout this project even though we had some trouble in the start, he kept believing the project with Sigrow might work.
We want to give thanks to Sigrow for allowing us to work on this project. Even though we only had contact with Loek one time we would like to thank her for getting us started on the project and teaching us about Sigrow and their work. In particular we want to thank Javier. On a personal note we want to thank Wouter Zwerink for teaching us an extremely useful trick for importing packages in our Jupyter Notebooks.

6 | References

[1] Alhnaity, B., Pearson, S., Leontidis, G., & Kollias, S. (2019). Using Deep Learning to Predict Plant Growth and Yield in Greenhouse Environments. *arXiv preprint arXiv:1907.00624*.

[2] Desik, A. (2019). Making the Foundation Strong: The Importance of Data Processing in Machine Learning/Artificial Intelligence. Retrieved from: https://www.tcs.com/blogs/making-the-foundation-strong-importance-of-data-processing-in-machine-learning.

[3] Hemming, S., de Zwart, F., Elings, A., Righini, I., Petropoulou, A. (2019). Remote control of greenhouse vegetable production with Artificial Intelligence—greenhouse climate, irrigation, and crop production. *Sensors*, *19*(8), 1807.

[4] Hemming, S., de Zwart, F., Elings, A., Righini, I., Petropoulou, A. (2019). Autonomous Greenhouse Challenge, First Edition (2018). 4TU.Centre for Research Data. Dataset. https://doi.org/10.4121/uuid:e4987a7b-04dd-4c89-9b18-883aad30ba9a

[5] Sethi, A. (2020). Support Vector Regression Tutorial for Machine Learning. Retrieved on June 23rd, from: https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/

[6] Tantau, H. J. (1991). Optimal control for plant production in greenhouses. *IFAC Proceedings Volumes*, *24*(11), 1-6.

[7] Autonomousgreenhouses. (n.d.). Autonomous Greenhouses International Challenge 2019. Retrieved on June 23rd, from http://www.autonomousgreenhouses.com/

[8] w3schools. (n.d.). Polynomial Regression. Retrieved on June 23rd, from: https://www.w3schools.com/python/python_ml_polynomial_regression.asp