POWDERY MILDEW MODELS AND FEASIBILITY TESTS

Nicholas Lee, Gabriel Dima

1 Introduction

Using the Smart-Dart device by Diagenetix[1], we have studied the spread of powdery mildew on grapes and tested the feasibility of using machine learning algorithms to create predictive models for this growth. This document will provide a brief overview of the methods and results of these studies. In Section 2 we detail our initial approach to the problem that combines weather data with Smart-Dart testing to create a predictive model using logistic regression. In Section 3, we create a theoretical 2-D model of powdery mildew that includes a prescription for wind, and then test the feasibility of using Smart-Dart and linear regression to create realistic models.

Section 2 is based on real data from weather stations and the Smart-Dart device, and the results can be used for future studies. Section 3 is based on a very simplistic theoretical model and the results should be only used as a demonstration of future capability.

1.1 Machine Learning Algorithms

Since much of the analysis is based on machine learning algorithms, I will provide a brief background on the basics. There is much more detailed information available from better written sources online[2].

Logistic regression and linear regression are both common forms of “supervised machine learning algorithms”, with the main difference being that logistic regression provides discrete yes/no predictions while linear regression can produce a continuous range of predictions. “Supervised” machine learning refers to the fact that these algorithms are given a set of parameters AND answers. The algorithms then determine the model that best translates from parameters to answers (usually “best” is defined by error minimization).

In our particular case the parameters will be weather & wind data and existing powdery mildew contamination levels, and the answers will be the next day’s expected powdery mildew levels. The machine learning algorithms will then come up with a model that will then be able to predict future powdery mildew levels based on future weather forecasts.

In order to analyze the accuracy of the machine learning models, datasets are usually split into training sets, cross-validation sets, and test sets (we use a 60%/20%/20% split in this analysis). The training set is what is actually used by the machine learning algorithm to find the “best” model, the cross-validation set is used to estimate how well the model has been trained (and can be used to fine-tune certain model parameters), and the test set is used to finally test the accuracy of the model. We will use these splits whenever we discuss testing models.

[1] http://diagenetix.com/

[2] I found this free online book useful: http://neuralnetworksanddeeplearning.com/index.html

2 Weather based Models

2.1 Dataset

In this study we combined weather and Smart-DART data from 3 locations (Frei, Laguna, Two Rock) during the period of March – May, 2014. The weather data included temperature, precipitation, leaf wetness, and relative humidity measurements in 15 minute increments. The Smart-DART measurements were taken weekly and provided a yes/no for the presence of powdery mildew. Weekly fungal treatments were also taken into account for the data, although there was no distinction between different types of treatment. Future work with much larger datasets may be able to distinguish the effect of different fungal treatments.

We designed a model that could take in a location's present powdery mildew status (if already detected or not), future weather forecasts for a week, and determine the likelihood of a positive detection at week's end. We designed it so that all of the entries from each location could be combined in a time-independent and location-independent way so that our dataset would be as large as possible. The parameters that we use are:

- Temperature, recorded as number of hours during the week when the temperature was in the following ranges:
  - < 60°F
  - 60–65°F
  - 65–70°F
  - 70–75°F
  - 75–80°F
  - 80–85°F
  - 85–90°F
  - 90–95°F
  - > 90°F
- Average leaf wetness
- Average humidity
- Previous week's result (1 for detection, 0 for non-detection)
- Treatment usage (1 for treatment, 0 for no treatment)

2.2 Results

From the machine learning algorithms, the parameters that display the strongest positive correlation (indicating they are most likely to lead to a positive powdery mildew detection) are: time in 70–75°F range, time in 85–90°F range, and average humidity. The strongest negative correlations are: time in 75–80°F range, time in < 60°F range, and time in 60–65°F range. The negative correlation with the 75–80°F temperature range is surprising, but may be due to small sample sizes or too fine temperature binning.

In general, the model can predict powdery mildew presence with an error of $\sim$ 20%–25%.

2.3 Tests with new data

In progress

3 Wind models

To test the feasibility of using machine learning algorithms to handle 2-D models that include wind, we created simulations that were then run through the machine learning algorithms. The results from the algorithms were then compared to the simulations to test the accuracy of these models.

3.1 Wind Simulation

The grid we simulate is based on a real area where we were able to obtain matched temperature and wind data for a grid of 51 locations in Napa Valley. For each day in each cell of the grid, we model the present powdery mildew concentration ($C_i(today)$) as a combination of these factors:

- Previous day's concentration ($C_i(yesterday)$)
- Growth of powdery mildew that would be expected in this cell due to temperature ($G(\text{temp})$)
- Inflow of spores due to wind from each of the 4 neighboring cells to north, east, west, and south ($W_{i,in}$). The influx from each neighbor only takes in the component directed into the target cell (meaning we only take the sine or cosine of the wind strength, depending on the angle of the wind)
- (Subtract) Outflow of spores due to wind in target cell ($W_{i,out}$) .

The growth of powdery mildew ($G(\text{temp})$) and the coupling of powdery mildew ($W_{i,out}$ & $W_{i,in}$) are both modeled with simple analytic equations that have easily tuned coefficients. The simulation begins with all cells at an arbitrary concentration of 1 and the model runs for a month, populating each cell in the grid every day.

3.2 Linear Regression

From this simulation of spore concentration, we performed a linear regression model to try to recreate the simulation. The model was fed the following parameters:

- Previous period's concentration in target cell
- Temperature in cell, recorded as number of hours during the week when the temperature was in the same ranges as above
- Wind strength at cell (used for outflow)
- Combined wind strength contributions from all 4 of the neighboring cells

Because we had a simulated concentration grid, we were able to test the model using different time periods between taking concentration measurements.

3.3 Results

The best way to visualize the results of our models is through the interactive GUI provided in the code show_values_nlee.py. There are three different implementations of the model that can be displayed. The first, which assumes daily measurements of the actual spore concentration, has a typical error of $\sim 1\%$. The next, which assumes measurements every 5 days, has errors ranging from 1%–6%. The final implementation is one where the model is fed only the starting conditions and then models the full month with no more spore measurements. The errors in this final implementation climb continuously up to $\sim 40\%$ by the end of the month. Using this model, one can set a maximum allowable error and then determine how often measurements are needed.