Automated Flowering Time Prediction Using Data Mining and Machine Learning

Runxuan Li[1], Yu Sun[2(✉)], and Qingquan Sun[3]

[1] The Baylor School, Chattanooga, TN 37405, USA
runxuan.lil983548012@gmail.com
[2] Department of Computer Science, California State Polytechnic University, Pomona, Pomona, CA 91768, USA
yusun@cpp.edu
[3] School of Computer Science and Engineering, California State University, San Bernardino, San Bernardino, CA 92407, USA
qsun@csusb.edu

Abstract. This paper presents a solution for the predictions of flowering times concerning specific types of flowers. Since flower blooms are necessarily related to the local environment, the predictions (in months), are yielded by using machine learning to train a model considering the various environmental factors as variables. The environmental factors, which are temperature, precipitation, and the length of day, contribute to the chronological order of flowering periods. The predictions are accurate to a fraction of a month, and it can applied to control the flowering times by changing the values of the variables. The result provides an example of how data mining and machine learning presents itself to be a useful tool in the agricultural or environmental field.

Keywords: Flowering time · Machine learning · Data mining

1 Introduction

Understanding the patterns of nature and making use of it is a topic that is both practical and fascinating in the modern era. Not until the year 1996 did the American scientists applied the technology of genetic engineering and put the products into mass production. By studying and engineering a specific crop or plant, people understand the traits of living plants. Plants, just like humans, respond to a variety of environmental changes; the results of the responds are expressed directly on the exterior part of the plant through the accelerating or tardiness of the growth of leaves, the chronological delay or outstrip of flower blooms, or the time of the last stage of fruit maturity.

Flowers, tend to show more stability than those that are incapable of flowering in terms of the cycle that plants follow. Flowering plants rigidly follows the cycle of seed, germination, growth, reproduction, pollination, and seed spreading. The ultimate goal of any flowering plant is to reproduce fertile offsprings, and by flowering, those plants become capable of receiving or giving pollen to itself or other flowers in the same species [4]. After pollination, the flowers fall down from the plant, and the seed that carries the next generation replace the part where the flowers were. The seeds

X. Gu et al. (Eds.): MLICOM 2017, Part II, LNICST 227, pp. 518–527, 2018.
https://doi.org/10.1007/978-3-319-73447-7_56

eventually grows into the prosperity of the original plant while the original plant goes on through the cycle of life once again.

The context that this paper focuses on is the flowering month of specific flowers. Every plant that goes through the process mentioned above have a flowering period, and in which the plant form the process of reproduction through seed spreading, and this period lasts from 1 to 6 months depending on the plant discussed. Despite so, the flowering month mentioned in this project is concerned with the average month of the year between the month that it starts flowering and the time it falls. For example, if the dahlia flower starts blooming in January and fall in March, the flowering month is set to be February (Fig. 1).

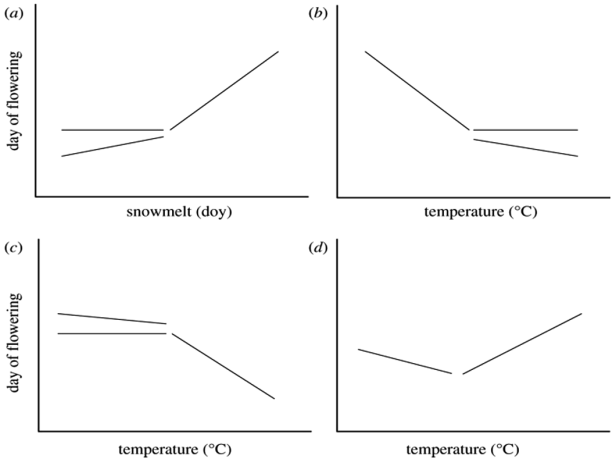

Fig. 1. Flowering time and temperature

Plants in nature take in a lot of different influences, and the time that it flowers can vary directly or implicitly determined by these influential factors. Wind, interference of insects or animals, and even the quality of air can affect the growth of a plant; the determination of which factors are the foremost of all that influence significantly to the flowering month of a plant can be a challenging problem. After making observations on the ten different types of flowers in different climatic environments that this research is based upon, a consensus is reached: the flowering time of the year for plants is extremely dependent on various factors: temperature, precipitation, and sunshine (length of day). These are the major independent variables in terms of the flowering time. The same observation also eliminated the effect of wind along with other factors due to their inconsequential effect that it would not change the flowering process by more than half a month. Finally, a conclusion is drawn that the major independent factors are the temperature, precipitation, and the length of day.

But of course, the effect of these factors contribute quite differently on distinct types of flowers. The effect exerted by temperature on cherry blossom might have a far more significant effect on the flowering month than it exerts on frangipani. It is indeterminant

that if one factor exert a more significant influence than the other one. Furthermore, the impact of the same variable on two distinct flower types can have astonishingly distinct effects. For example, the lack of watering for sunflower might not be a big thing while iris definitely can not withstand a dry environment.

Due to the fact that the influence of each factor is different as mentioned in the previous paragraph, the amount of influence they can exert is believed to be a constant, and this constant, or fixed amount of influence is somewhat similar to the coefficient of a variable in a function. This suggests that people can look at the month of flowering of a plant as a function of several variables: type of flower, temperature, precipitation, and length of day. Furthermore, a rough model can be build based on these variables; this is viable, and more importantly, practical. Because researchers can use the model to estimate the month of flowering if they input the four variables that are requisite in terms of yielding a result. Nevertheless, since this model contains four separate variables, it is not a viable way to figure out the relations through manual calculations. Also, the correlations between the separate variables can yield a unique effect on the final result.

In this paper, we present a complete approach to model the flowering time using both data mining and machine learning techniques. A hybrid system (i.e., mobile and web) has been built to expose the prediction functionality to end users.

The rest of the paper is organized as follows: In Sect. 2 we explain the 3 major challenges that occurred in predicting the flowering month; Sect. 3 provides the solutions to solve for these problems with technical details; Sect. 4 demonstrates the experiments conducted to improve the accuracy of the prediction; Sect. 5 presents the related work, while Sect. 6 summarizes the project by giving a conclusion and providing future work directions.

2 Challenges

2.1 Challenge 1: The Diversity of Data Factors

As mentioned in Sect. 1, different types of flowers tend to bloom at different times in distinctive locations. For example, the flowering month of Cherry Blossom in Fuzhou Fujian, China in the year 2016 is promptly at April, but the flowering month for the same type of flower in Matsumae Hokkaido, Japan is around May [REF]. The location differences contribute to these distinct factors. In 2016, Fuzhou had a yearly average temperature of 69.25 Fahrenheit, a monthly average precipitation rate of 4.63 in., and an average length of day of 13 h. In the same year, Matsumae had different factors: temperature at 47.92 F, precipitation at 0.5 in., and the length of day at 12.25 h. Thus, the first challenge here is how to package these data factors in a way that leads to an effective prediction model.

2.2 Challenge 2: The Inefficiency of Processing Data Manually

Based on some of the flowering data factors mentioned above, it is a common misconception that the higher the temperature, bigger the precipitation, and the longer the

length of day, the cherry blossom would bloom earlier. However, the problem often involves a large number of data, and it is time-consuming and error-prone to dig into the data and decide the effect of each manually.

Furthermore, the multi-dimensions of the data makes an infeasible task to accurately model manually. In the traditional way of manually doing this problem, researcher have to build 10 models of flowers in terms of the variables x, y, z, and this suggests that each model has to contain 3 separate equations of these three variables. In this case, the traditional method of manually working our way through is excluded from the solution methods. An alternative way of processing data needs to be found.

2.3 Challenge 3: The Unpredictable Variance of Factors Weight

Another similar challenge concerning the factors is that it is hard to understand the weight of each factors in terms of the impact to the flowering time. Apparently some of the factors weight less than others due to the fact that some flowers can blossom even in extreme environment. In general flowering time researches, researchers tend to directly look at the data and decide on the trend that flowers follow, but if the research becomes specific and needs accurate data for the month of flowering, the general analysis would not be very useful.

2.4 Challenge 4: The Difficulty of Selecting the Appropriate Training Model

After figuring out how we can solve the previous challenges, the question becomes what is the most accurate model that the predictions can be based upon. In other words, what type of regression should the research data be fitted in. Since the weight of factors represent the coefficients of the variables, the model represent the general equation which the data take form in. In this case, the data can be fit into various models, such as exponential model or linear model. The challenge is to acquire a model which yield the most accurate results.

3 Solution

In order to provide a solid solution that everyone can have access to, we have developed a machine learning approach to predict the flowering time based on a number of factors. The solution has been implemented in both a web-and a mobile-based application.

3.1 Data Model and Collection

As mentioned in the introduction, the task of modeling and dealing with data would be put upon machine learning and data mining. The process of this project is designed to be as the following:

(1) Gather the flowering information needed
(2) Change the information into data points

(3) Choose an accurate algorithm for the prediction
(4) Algorithm and back-end implementation
(5) Front-end development for both mobile and web

The method of the collection of data is similar to researching about the climate of a specific region. But gathering the information is a little more complicated. In order to accomplish step 1 in the procedure, research is one inevitable step. The information section contains the time when one type of flower blossom in one specific area, and by searching the temperature, precipitation and length of day in one year in that area, one data point is obtained. Then, the process is repeated over again, this time finding a new location in which the flower blossoms.

The hidden difficulty appears to be the first step of finding a place that have one type of flower. For some flowers, such as cherry blossom, is easy to extract information due to the fact that a lot of places (Washington DC., Hokkaido, etc.) held cherry blossom festivals. These festivals often have websites which tells the tourists when the cherry blossom tend to bloom and when they would fall. The following step is to get the information of the previous year in which the flower blossom. For example, Fig. 2 shows the cole flower in Fuzhou, Fujian China blossom in March in the year 2016.

Table 1

flower type	month	temp(F)	precipitation(in)	length of day(h)	
primrose 1	3.3, 3, 2, 1, 2.5, 4	65.83, 69.25, 59.67, 65.58, 67.3, 69.1	0.97, 4.63, 2.9, 0, 1.7, 3.2	12.15, 13, 12.15, 13.25, 12.8, 13.2	
tulip 2	4, 3, 6, 2, 1.5	68.58, 69.25, 50.3, 70.1, 70.3	0, 4.63, 2.14, 3.1, 4.5	13.25, 13, 12.7, 13.2, 13.23	
dahlia 3	9, 8.5, 8, 7, 10	55.25, 60.08, 62.3, 63.4, 52.5	3.75, 1.39, 2.1, 3.62, 1.28	12.23, 12.18, 12.12, 13.23, 12.01	
sunflower 4	10, 8, 6.5, 9, 7	50.67, 56.83, 68.17,52.7, 65.8	3.5, 2.93, 5.06, 2.7, 4.87	12.18, 12.23, 12.45,12.19, 12.65	
cole 5	3, 4, 4.5, 6, 2.5,	69.25, 68.58, 59.67, 57.3, 70	4.63, 0, 2.9, 2.48, 3.63	13, 13.25, 12.15, 12.84, 13	
cherry blossom 6	3, 5, 4, 2, 5.5	69.25, 47.92, 61.42, 70.23, 46.23	4.63, 0, 3.48, 4.72, 0,	13, 12.25, 12.17, 13.1, 12.13	
iris 7	3, 6, 4, 1, 3.5	64.58, 59.67, 62.7, 70.43, 63.4	0, 2.9, 2.65, 3.23, 2.56	13.25, 12.15, 13.2, 13.3, 13.18	
magnolia 8	3.5, 2.5, 1, 2, 4,	69.25, 68.58, 59.67, 63.17, 70.16	4.63, 0, 2.9, 1.8, 2.67	13, 13.25, 12.15, 12.07, 13.4	
frangipani 9	8.5, 9, 5, 7, 6	62.25, 61.58, 69.67, 67.2, 68.37,	2.9, 0, 4.63, 3.24, 4.23	13.25, 12.25, 13.4, 13.32, 13.39	
jasmin 10	8, 4, 6, 5, 7	69.25, 68.14, 64.2, 66.38, 68.69	4.63, 3.73, 3.38, 3.47, 2.96	13, 13.25, 12.98, 12.83, 13.3	

Fig. 2. The data table for all flowers and factors

In this case, we went to the weather website containing the previous monthly average temperature, sum of precipitation, and average length of day; we extracted these data from April 2015 to March 2016 and averaged them in order to get exactly three numbers representing the three factors. After doing that, we put the data into a table ready for edit in order to export to the python program.

3.2 Machine Learning

Generally speaking, machine learning is a method of analyzing data and making predictions with auxiliary programs like Java, Matlab, C, etc. The program that we used

in the research is Python. Unlike manually operating with a pen and paper, machine learning already have written programs for various types of algorithms, and the only job we are doing is to figure out which algorithm yield a result that is the closest to the precise flowering time.

The Python library has numerous algorithms that is provided by programmers, so it is really convenient to use. In the current edition of Python, to use the algorithms that the Python Community provided, programmers can type in the order "import", which represent the importation of a long written program and running it on the user's Python program. This is an easy and solid way to model the data, and furthermore, it save us time to try more than one or two algorithms in order to determine which one can make the prediction that is closest to the precise precise result.

In addition, we believe the issue around the question "At what conditions would a flower blossom in a specific month?" is worth researching upon. So we started a new group of codes based on the predictions of the old one. Given the the flowering month and the type of flower, the program would provide the user with the information concerning the three factors.

3.3 Web Service

After the completion of the program concerning the algorithm of the prediction, the file is working only locally. So the next task is to upload the program onto the web and let everyone with internet have excess to it. we bought a domain recently, so we uploaded the converted front end file to one of the sub-domains: http://flower.runxuanli.com. The client was developed using one side language: HTML, since it can work on any browser.

The webpage contains four boxes:

(1) The flower type (labeled in numbers)
(2) The monthly average temperature for one year
(3) One year of precipitation
(4) Average length of day in one year

After clicking the submit button when the data are typed in, another web page will pop up stating your predicted month. The month is expressed accurate to the hundredth place, which can be converted to days by multiplying 30 or 31 depending on the month. The flowering time is accurate to one hundredth of a month, and the result of the conversion is generally in the form of month plus days (Fig. 3).

3.4 Mobile App

The last step of the project is the development of a mobile app for the users to have a more easy access to the application. Some changes on the program were made in order to adapt the requirements of the mobile app. The basic preview of it is quite similar with front end web page; the user is asked to fill in the four boxes and hit submit (Fig. 4).

The challenges in the previous section is more or less addressed and solved for in this section. The research of gathering the data was made into a table and later applied in the Python Program for machine learning. Also, the deficiency of manual operation

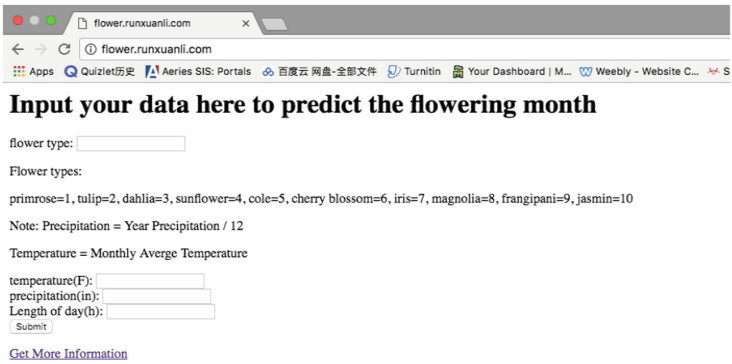

Fig. 3. Overview of the webpage

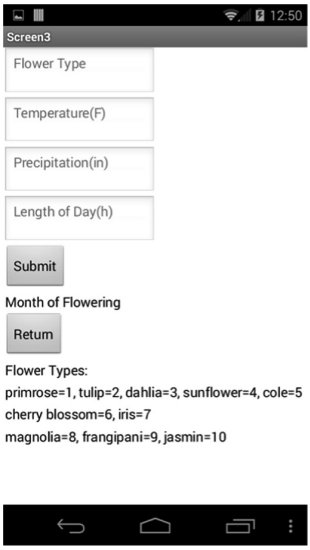

Fig. 4. Android App

and the problem concerning the weight of the factors is resolved by the powerful machine learning. Lastly, by various tentative approaches, the model which the data are fitted into is determined. Therefore, all of the major problems in the research are solved.

4 Experiments

In this section, we verify the accuracy of the algorithm that is currently in use. In addition, the verification also includes the experiments conducted using the variables other than the temperature, precipitation and length of day. Based on the result, we either verify that the algorithm in use is the most precise or change and improve the algorithm.

4.1 Experiment 1: The Accuracy of the Overall Prediction

The following experiments 2 and 3 will test the accuracy of the model in use in terms of algorithms and the factors chosen. However, the foremost auxiliary experiment that experiment 2 and 3 are persistently using but not mentioning the way to conduct it is the experiment that proves the model is accurate to what extent. When doing any kind of verifying, it is important to match the predicted result to the actual result. This method of experimenting is to take a number of data point that is one less than the number of data points that you obtained. After training the data points and fitting them into a model, try predicting by using the variables in the last data point. In this way, the result of the prediction can be compared with the result of the actual situation.

After doing the experiments 2 and 3, experiment 1 is always applied, and the result will be described in the following experiments.

4.2 Experiment 2: The Accuracy When Using Different Machine Learning Algorithms

Often the hardest part of solving a machine learning problem is finding the right estimator for the job. Different estimators are better suited for different types of data and different problems. And by far, there are too many models for the classical guess and check method. Therefore, we have to manually exclude some of the algorithms from our list.

There are two major groups of algorithms designated for machine learning: regression and classification. After providing data, the regression algorithm yields specific numbers as the result. This works similar as a function: when you input the "x value", the function gives you the "y value". In the classification algorithm, however, the computer groups the datas into clusters and designate which general cluster belongs to which group. For example, 3 apple trees have the following tree trunk lengths: 3 m, 3.5 m, and 3.6 m; Cherry trees have the following tree trunk length: 1 m, 1.2 m, and 1.3 m. If given a tree with the tree trunk length of 1.4 m, which tree is it? The classification algorithm would first group the datas together and tell you that the tree is a cherry tree. However, if you represent the apple tree with the number 1 and cherry tree with the number 2 and use one kind of regression to solve for the type of tree, the computer would not tell you what kind of tree it is, but would give an ambiguous answer like 2.3. And since we are operating the data to get a month, accurate to a hundredth of a month, regression will be our first choice.

However, by far the experiment had only solved half of the problem by eliminating one group of algorithms. We still have numerous algorithms in the regression group to choose from. After researching about the relationship between temperature and flowering, we found out that the flowering month corresponds to the temperature in the form of a parabola [2]. So we experimented on the even polynomials, and it turns out that the secondary polynomial gives the answer that is the most precise (Fig. 5).

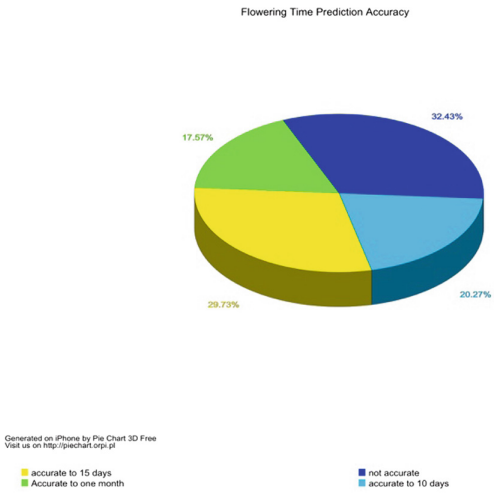

Fig. 5. Accuracy of prediction

4.3 Experiment 3: The Accuracy When Using Different Sets of Factors

Experiment 3 conducts an experiment that decides which factors should be added or subtracted that are used in coming up with the result. Like mentioned in the introductory section, temperature, precipitation, and length of day are not the only factors determining flower blooming; therefore, we decided to take two other factors: wind speed and humidity. We consider these two factors to be the most influential on the prediction and put them into the data set for the prediction.

It turns out that due to the erratic behavior of the wind speed, a discrepancy occurs in the prediction with the two factors and without the factors. And using the method described in experiment 1 of leaving out a data point to test the two methods, the one without adding the two factors appears to be the closest to the actual flowering time.

5 Related Work

There had been many studies concerning the reasons of flower bloom [1, 3, 5]. Most of them are focusing on the genetic part of the flowering process. Furthermore, these researches often focus only on one flower and how variety of influences tend to change the normal flowering time. This approach is tremendously useful for mass agricultural studies related to the biological field. However, these studies does not provide a solid reference for individual gardening practice since flowers as a genre of decorative plants, can be controlled in a more practically of watering and controlling the temperature.

6 Conclusion and Future Work

In conclusion, this project has derived information from more than 50 data points, using one machine learning based algorithm model to make predictions of flowering time. The practicability of machine learning applied on gardening practice is the foremost objective of this project. With a website and a mobile app, the practical achievement of the research is made accessible and tangible to the audience using the product.

In the long term, we hope to develop or get the permission to use a flower recognition app on the cellphone to correlate with the present project. If that app is developed, we will write a code using the Python Program to extract the information in terms of temperature, precipitation, and length of day in the specific region that the flower is spotted and taken a picture of. Then the data points will be added onto the original data sets and make the predictions more transparent and precise.

References

1. An, H., et al.: CONSTANS acts in the phloem to regulate a systemic signal that induces photoperiodic flowering of Arabidopsis. Development **131**, 3615–3626 (2004)
2. Shull, C.A.: Temperature and flowering. Bot. Gaz. **78**(2), 244–245 (1924)
3. Fornara, F., de Montaigu, A., Coupland, G.: SnapShot: control of flowering in Arabidopsis. Cell **141**, 550.e1 (2010)
4. Zeevaart, J.A.D.: Physiology of flower formation. Annu. Rev. Plant Physiol. **27**, 321–348 (1976)
5. Wellmer, F., Riechmann, J.L.: Gene networks controlling the initiation of flower development. Trends Genet. **26**, 519–527 (2010)
6. Iler, A.M., Høye, T.T., Inouye, D.W., Schmidt, N.M.: Nonlinear flowering responses to climate: are species approaching their limits of phenological change?
7. Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., Struhl, K.: Current Protocols in Molecular Biology. Wiley, New York (1995)
8. Blazquez, M.A., Ahn, J.H., Weigel, D.: A thermosensory pathway controlling flowering time in Arabidopsis thaliana. Nat. Genet. **33**, 168–171 (2003)
9. Hutvagner, G., McLachlan, J., Pasquinelli, A.E., Balint, E., Tuschl, T., Zamore, P.D.: A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. Science **293**, 834–838 (2001)
10. Hutvagner, G., Zamore, P.D.: A microRNA in a multiple-turnover RNAi enzyme complex. Science **297**, 2056–2060 (2002)
11. Kawasaki, H., Taira, K.: Hes1 is a target of microRNA-23 during retinoic-acid-induced differentiation of NT2 cells. Nature **423**, 838–842 (2003)
12. Olsen, P.H., Ambros, V.: The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. Dev. Biol. **216**, 671–680 (1999)
13. Tang, G., Reinhart, B.J., Bartel, D.P., Zamore, P.D.: A biochemical framework for RNA silencing in plants. Genes Dev. **17**, 49–63 (2003)
14. Xu, P., Vernooy, S.Y., Guo, M., Hay, B.A.: The Drosophila microRNA Mir-14 suppresses cell death and is required for normal fat metabolism. Curr. Biol. **13**, 790–795 (2003)