Initial Machine Learning Framework Development of Agriculture Cyber Physical Systems

Gregorius Airlangga¹, Alan Liu²

Information System Department, Atma Jaya Catholic University of Indonesia, Indonesia¹
Electrical Engineering Department, National Chung Cheng University, Taiwan²

gregorius.airlangga@atmajaya.ac.id, aliu@ee.ccu.edu.tw

Abstract. A cyber-physical system (CPS) shows its potential in integrating the computer systems with the physical environment. The progress in such systems has brought large potential to improve human life quality, such as improved health care services, energy consumption and food supply chain system (FSC). In particular, FSC plays an important role in human beings. It comes from the complex agricultural system and lead up to human dining table. The fusion of CPS and agricultural system could improve the quality of food and environment. Therefore, many studies have been conducted to tackle the challenges in this domain, such as lack of information systems and infrastructures, poor collaboration toward larger Internet of Things solutions and dynamic changes of intrinsic and extrinsic conditions of enabler technology in precision agriculture. In this work, we focus on developing an initial framework that could improve prediction rate and handling imprecise data due to the dynamic problem in precision agriculture. As an evaluation, first we predict a rainfall event using weather sensors data and after that the prediction result will be continue to set up as an additional attribute for predicting the action of water sprinkle monitoring system. The results confirm the good accuracy for farmers and could be applicable in real-time.

1. Introduction
The advancement of Information and Communication Technology (ICT) has been leading towards the realization of Industry 4.0 that commonly referred to as fourth industrial revolution [1] [2]. One of the current issues in this area is Agricultural Cyber-Physical Systems (ACPS) which are seen as an orchestration of computer systems and agricultural environment [3]. The progress in the systems has brought large potential to improve food supply chain system (FSC) quality [4]. It plays an important role in human beings and has a wide cycle that comes from farm and includes farm production, food processing packaging, sales & marketing, logistic & distribution, retail, consumer interactions and waste management [5].

Furthermore, those cycles have different challenges to face especially by utilizing CPS technology. One of the well-known challenges in this area is precisions agriculture which is a management strategy that engages circumstances and specific information to precisely manage production input such as soil and crop characteristic unique to each part of the field, precision agriculture aims to optimize the production input within small portions of the field, such as reduced use of water, fertilizers, herbicides and pesticides besides farm equipment in real-time [6] [7]. Most recently, one of the enabling precision agricultural technologies is Internet of Underground Things (IOUT) [8]. IOUT represents autonomous devices that gather any relevant information about the Earth and are interconnected with communication...
and networking systems that facilitate information transactions from the fields to the farmers and decision making system. The advantage of IOUT is real-time sensing. Sensing has led to adoption of technology in the precision agriculture and it also improved efficiency of agricultural production and practices [9]. Practically due to the dynamic changes in the intrinsic and extrinsic conditions of underground things such as communication medium in soil, seasonal changes, crop growth cycle. The technology should be able to autonomously adapt the conditions in order to give a precise decision making system to farmers in time fashion [10].

In this work, we propose an initial machine learning framework that resolve adaptability problem. The machine learning framework consists of five dimensions including normalizing data, handling incorrect data, time-series prediction, information fusion and classifier for decision making system. As an accuracy evaluation, we did an experiment using dataset comprised of fifteen sensors. The sensors information contains several attributes, such as soil moisture, air temperature, soil temperature, winds velocity and air pressure. This work is organized in five sections. Section I presents an introduction, section II Framework Theory, Section III the experiment, Section IV the result and analysis, and the last the conclusion.

2. Proposed Framework
The machine learning framework is comprised of three main steps containing data preprocessing, data forecasting and data classifier. There are three steps of data preprocessing to support machine learning classifier, such as normalizing data, handling incorrect data and information fusion. After that, the data will be computed in forecasting algorithm in order to prevent the lack of data or imprecise data collection for the next sensing cycle. As a data classifier, we present Naïve Bayes classifier to predict a real-time event based on the discretized fusion of forecasted data and prior knowledge based. The event prediction is calculated in each time cycle.

2.1. Normalizing Data
Practically it is often found that each feature is located in different values interval. One of methods to resolve this problem is called normalization technique that has a purpose to produce data into same interval. On this work, we utilized the normalizing method to scale each feature element to be an interval values that have a range between [0-1]. The implementation is shown in first equations. Where \( i \) is represented the each source index that have same feature and \( j \) is represented as an element index in sensor \( i \) that would be normalized.

\[
\hat{x}_{ij} = \frac{x_{ij} - x_{im\text{ax}}}{x_{im\text{ax}} - x_{im\text{in}}} \quad (1)
\]

2.2. Handling Incorrect Data
Sometimes data obtained from various sources such as sensors can be false or null due to the various factors. In order to resolve this problem, we propose a novel approach to predict the incorrect data instead of throwing all the data or changing the wrong value into random. First, we normalize the data in vector of sources, after that we calculate the average of proximity between all existing normalized values in all data at the time \( (t) \) and source \( (i) \) where incorrect values event occur. The second, third and fourth equations depict the process to forecast incorrect data.

\[
\phi(i) = \sum_{j=1}^{N-1}(\hat{x}_j - \hat{x}_{(j+1)}) \text{ for } x_{(j+1)} \neq \text{null} \quad (2)
\]

\[
k_i = \phi_i x_{im\text{ax}} \quad (3)
\]

\[
xf(i)_j = \frac{\sum_{k=1}^{N}k_{(j+1)} + \sum_{d=1}^{N}k_{(j-1)}}{2} \quad (4)
\]
2.3. Information Fusion
In the information fusion approach, we use a simple weighting average that implemented each data source from same feature. We calculate the information fusion value with an interval weight value between [0-1] in order to form information in each feature collected from several sensors. The implementation of information fusion is based on the 5th equation.

\[ \hat{x}_t = \frac{\sum_{i=1}^{N} x_i w_i}{N} \]  

2.4. Prediction Method
In order to predict the subsequent data at the time, we utilize single exponential smoothing forecasting method. The method is selected since the actual value does not have a significant trends and seasonal behavior at each subsequent time. The exponential smoothing method is defined as the 6th equation.

\[ F_t = F_{(t-1)} + (A_{(t-1)} - F_{(t-1)}) \]  

2.5. Naïve Bayes Classifier
In order to make a decision based on the data sources at time \((t)\). We use Bayesian theorem and the probabilities or conditional opportunities are expressed as the 7th equations

\[ P(H \mid X) = \frac{P(X \mid H) P(H)}{P(X)} \]  

Where \(X\) is a proof, \(H\) is a hypothesis and \(P(H \mid X)\) is probability that have the correct value of \(H\) hypothesis for true \(X\) evidence for \(H\) hypothesis or posterior probability of \(X\) with the condition of \(H\). \(P(H)\) is the prior hypothesis probability of \(H\), \(P(X)\) is the prior probability of proof \(X\). Usually \(X\) is represented by a tuple \(X\) or data object, \(H\) is a hypothesis or presumption that tuple \(X\) is located in class \(C\). Specifically in the classification problem we can calculate \(P(H \mid X)\) as the probability values where the \(H\) hypothesis is correct for tuple \(X\) or on the other words \(P(H \mid X)\) is the probability where the tuple \(X\) is located in a class \(C\). Meanwhile \(P(H)\) is the prior probability that the \(H\) hypothesis is correct for each tuple regardless of its attribute values while \(P(X)\) is the prior probability of tuple \(X\). The following steps are representing how to use Naïve Bayes classifier.

1) Suppose \(D_i\) is a training set that contains tuples elements, attributes and class labels are contained in each attribute. Each tuple that has \(n\) dimension could be expressed as \(X = (x_1, x_2, \ldots x_n)\) which obtained from their attributes \(A_1, A_2, \ldots, A_N\).

2) Suppose there are \(m\) number of classes which are containing of \(C_1, C_2, \ldots, C_m\). For an input of tuple \(X\), the Naïve Bayes classifier would be predict the tuple \(X\) belongs to the \(C_i\) class if and only if satisfy the 8th equation.

\[ P(C_i \mid X) > P(C_j \mid X) \text{ for } j \leq m, j \neq i \]  

Naïve Bayes classifier works by maximizing the value of \(P(C_i \mid X)\). The class of \(C_i\) which makes \(P(C_i \mid X)\) and has a maximum value will be called as maximum posteriori. Using Naïve Bayes theorem, the values of \(P(C_i|X)\) would be estimated using a 9th equation.

\[ P(C_i \mid X) = \frac{P(X \mid C_i) P(C_i)}{P(X)} \]  

3) Given that \(P(X)\) has the same value for all classes, its means the tuple \(X\) has the same probability for entering into any class. However, only \(P(X \mid C_i)P(C_i)\) needs to be maximized. If the prior
probability for each class is unknown, then the probability of each class can vary and then the Naïve Bayes classifier must maximize the value of \( P(X|C_i) \ P(C_i) \). In addition, when dealing with data sets that have so many attributes, it is necessary to reduce the complexity calculation of \( P(X|C_i) \) with the assumption that a class has a conditional independence such as attribute values that have mutually independent. Thus Naive Bayes classifier would be maximizing the calculation as the 10th equation.

\[
P(C_i | X) = \prod_{k=1}^{n} P(X_1|C_i)P(X_2|C_i) \ldots \ P(X_n|C_i) \tag{10}
\]

Based on the tuples, we can estimate the values of \( P(X_1|C_i) \times P(X_2|C_i) \times \ldots \times P(X_n|C_i) \) depending on attribute type. For the categorical value type, \( P(X_k|C_i) \) is defined as the number of tuples in the class of dataset which has an \( X_k \) value in the \( A_k \) attribute divided by the number of all tuples in the class \( C_i \) in dataset \( D_k \) which is symbolized by \( |C_i| \). Meanwhile, for attributes of continuous value we can calculated the value of \( P(X_k|C_i) \) using Gaussian distribution method as the 11th equation.

\[
P(x_k|C_i) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu_C_i)^2}{2\sigma^2}} \tag{11}
\]

Finally to predict the class label of the \( D_x \) tuple, the probability calculation of \( P(X|C_i) \ P(C_i) \) must be calculated for each class of \( C_i \). Next, it is necessary to maximize the probability, which is to look for the \( C_i \) class that produces the maximum probability \( P(X|C_i) \ P(C_i) \) as the decision class. Mathematically, tuples \( X \) are labeled as class \( C_i \) if and only if satisfy the 12th equations.

\[
P(X | C_i) \ P(C_i) > P(X | C_j) \ P(C_j) \text{ for } 1 \leq j \leq m \text{ and } j \neq i \tag{12}
\]

### 3. Experiment

#### 3.1. Dataset & Tools

The dataset used in this study is a real-time soil moisture and soil temperature measurements for 15 locations in the Netherlands. The network monitors soil moisture in the unsaturated zone for different soil textures and land covers present in the area. The 5TM sensors are installed at depth of 5cm, 10cm, 20cm, 40cm and 80 cm. The logging time interval is set on 15 minutes. The data was collected from 5th April 2016 - 5th April 2017 (One Year) [11].

#### 3.2. Prior Knowledge

In order to implement the machine learning framework, we use the water sprinkle monitoring system as our example. To predict the precise action of water sprinkle monitoring system in real-time, we set a prior knowledge with multiple features that are related with our dataset. Each feature is implemented as attribute in our knowledge base. Firstly, we calculate the probability of rainfall based on four attributes such as Air Temperature, Wind Speed, Air Pressure and Air Humidity. Then the rainfall attribute as a class attribute will be determined. There are four class attributes such as light rain, moderate rain, heavy rain and no rain. After we determined all classes, we calculate the probability of rainfall as the input feature to determine the action of water sprinkle system. There are three attributes and one class attributes of water sprinkle monitoring system. The classes are comprised of four values such as not active, short duration, medium duration and long duration.
| Time       | 5th April 2016 - 5th April 2017 |
|------------|----------------------------------|
| Attributes | Actual Data                     |
| Soil Moisture | 8.0692890536838E-05            |
| Soil Temperature  | 0.0157778494              |
| Air Humidity   | 0.7707157827                 |
| Air Temperature | 1.7928824762               |
| Wind Speed    | 1.5754473642                 |
| Air Pressure  | 0.8202079639                  |

3.3. Accuracy Measurement

In the experiment, accuracy is calculated by considering the smallest error value between actual data and forecasted data that was collected from 5th April 2016 - 5th April 2017 (One Year). The small error value indicates that the value is better. We use MAE (Mean Absolute Error) presented as the 13th equation.

\[
MAE = \frac{1}{N} \sum_{i=1}^{n} |F_i - A_i|
\]  

4. Result & Analysis

In this section, we use data collected from fifteen stations. Each station has a five soil temperature and soil humidity sensors set up on different depths such as 5 cm, 10 cm, 20 cm, 40 cm and 80 cm. In addition, we enrich the dataset using weather sensor comprising of four attributes such as air pressure, wind speed, air temperature and air humidity to predict a rainfall event using Naive Bayes classifier. As shown in Fig. 1-6, data was collected once per fifteen minutes. During this period, the environment attributes would be corrected in order to handle incorrect data or missing value using our algorithm. After that, we fused the result of each source that have similar feature such as sensor soil humidity and soil temperature.

Subsequently, we use forecasting method to predict data and decide water monitoring sprinkle system action. The forecasting method is conducted to give a real-time prediction approach at each time (t). By using forecasting method, we could prepare system to adapt in uncertain condition. As shown in Table 3 we get a good accuracy forecasting result for each source.
The MAE values that are shown on Table 3 present the mean of difference between forecasted and actual values. The two smallest MAE values are presented by soil moisture and soil temperature which means the moisture and temperature do not have much change within interval of 15 minutes. This information is very useful for our water sprinkle monitoring system model to decide the action in time fashion that would be effectuated to reduce the water waste and give more energy efficiency for irrigation. Overall, the other MAE results give a good prediction to farmer and give a good model to predicting rainfall. In order to evaluate our approach, Naïve Bayes classifier is conducted based on our knowledge prior. The total of event prediction is 24.187 which have a 15 minutes difference between each time phase. As shown in Table 4 we get the satisfactory result to predict the rainfall and water sprinkle monitoring system action. The true predictions are calculated by subtracting the total of actual event with forecasted event. From those tables we can see more clear differentiation between our approach and conventional ones, especially accurate rainfall prediction. Our framework gives more useful information to the water sprinkle monitoring system action because the conventional method cannot handling incorrect data problem such as -9999, NaN and empty values that are produced when sensors are faulty or the battery is turned off. In addition, the conventional method cannot handle the information fusion with precise weight because the initial weight is just set with static value.

Table 4. Naive Bayes Classifier Accuracy

| No. | Event                                | True Prediction (Our Framework) | True Prediction (Conventional) | %     | %     |
|-----|--------------------------------------|--------------------------------|--------------------------------|-------|-------|
| 1.  | Rainfall Prediction                   | 23,218                         | 21,417                         | 91.85%| 88.55%|
| 2.  | Water Sprinkle Monitoring System     | 22,134                         | 21.630                         | 91.51%| 89.43%|

Table 5. Computational Speed

| No. | Programming Language | Computational Speed (s) (Our Framework) | Computational Speed (s) (Conventional) |
|-----|----------------------|-----------------------------------------|----------------------------------------|
| 1.  | C++                  | 0.228                                   | 0.117                                  |
| 2.  | Python               | 0.521                                   | 0.432                                  |
| 3.  | Java                 | 0.627                                   | 0.511                                  |

Table 5 shows the computational speed using three different programming languages that commonly used as Internet of Things programming and data manipulation. The purpose of computational speed
measurement is to find the best tool to implement in real-time condition. The code is designed with explicit iterative architecture and not using vector. The environment to calculate time is based on the same machine with specification processor Intel-i7 with ram size 16 GB. Overall the C++ programming language has fastest computation since its ability to work intrinsically stingy with memory and small portion of caching.

5. Conclusion
In this paper, we have proposed an initial machine learning framework that might be applicable in Agricultural Cyber-Physical Systems application. The results shows that the forecasting result of every sensor is indicating the good accuracy for farmers and it could predicts the future event such as rainfall and water monitoring sprinkle system behavior in real-time. In the future, we will present the algorithm that could decide efficient time scheduling to irrigate soil and crops using multi-sensor and multiple datasets in dynamic environment.

6. References
[1] Delenogare LS, Benitez GB, Ayala NF, Frank AG. The expected contribution of Industry 4.0 technologies for industrial performance. International Journal of Production Economics. 2018; 204: p. 383-394.
[2] Lee J, Davari H, Singh J, Pandhare V. Industrial Artificial Intelligence for Industry 4.0-based Manufacturing Systems. Manufacturing Letters. 2018; 18: p. 20-23.
[3] An W, Wu D, Ci S, Luo H, Adamchuk V, Xu Z. Chapter 25 - Agriculture Cyber-Physical Systems, in: H. Song, D.B. Rawat, S. Jeschke, C. Brecher (Eds.). In Cyber-Physical Systems, Intelligent Data-Centric Systems. Boston: Academic Press; 2017. p. 399-417.
[4] Chen RY. An intelligent value stream-based approach to collaboration of food traceability cyber physical system by fog computing. Food Control. 2017; 71: p. 124-136.
[5] Ramundo L, Taisch M, Terzi S. State of the art of technology in the Food sector value chain towards the IoT. In IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a Better Tomorrow (RTSI); 2016; Bologna, Italy.
[6] Huuskonen J, Oksanen T. Soil sampling with drones and augmented reality in precision agriculture. Computers and Electronics in Agriculture. 2018; 154: p. 25-34.
[7] Capmourteres V, Adams J, Berg A, Fraser E, Swanton C, Anand M. Precision conservation meets precision agriculture: A case study from southern Ontario. Agricultural Systems. 2018; 167: p. 176-185.
[8] Vuran MC, Salam A, Wong R, Irmak S. Internet of underground things in precision agriculture: Architecture and technology aspects. Ad Hoc Networks. 2018; 81: p. 160-173.
[9] Tzounis A, Katsoulas N, Bartzanas T, Kittas C. Review: Internet of Things in agriculture, recent advances and future challenges. Biosystem Engineering. 2017; 164: p. 31-48.
[10] Zikria YB, Yu H, Afzal MK, Rehmani MH, Hahn O. Internet of Things (IoT): Operating System, Applications and Protocols Design, and Validation Techniques. Future Generation Computer Systems. 2018; 88: p. 699-706.
[11] Benninga HJF, Carranza CDU, Pezic M, Santen Pv, Ploeg MJVD, Augustijn DCM, et al. The Raam regional soil moisture monitoring network in the Netherlands. Earth System Science Data. 2018; 10: p. 61-79.