Carnegie Mellon University Africa
Certificate I: Understanding AI and Machine Learning in Africa

Course AIML02: AI and Machine Learning in Africa

Module 02: Application Case Studies
Lecture 03: Agriculture

Welcome to Lecture 3 of Module 2, in which we are examining case studies of the application of AI and machine learning in Africa.

In this lecture, we will focus on a case study in agriculture, looking at the ways that computational techniques can be used to monitor crop disease in the developing world, with specific examples in Uganda.

We introduce the target article for the case study and explain its significance in the context of AI and machine learning in agriculture, highlighting the importance of crop disease monitoring in the developing world.

We then explain the advantages of using mobile devices to collect survey data and perform automated diagnosis of crop disease.

We explain how the data that is gathered can be used to construct spatial maps of the incidence of disease

We move on then to explain the advantages of performing the diagnosis and the map construction together.

Finally, we highlight the additional advantages of this approach for optimizing the limited resources available when conducting surveys and monitoring disease.

We finish up by summarizing what we have covered and identifying the articles that you should read to consolidate what you have learned.

Once you have listened to this lecture and read the commentary, you should read the target article, and listen to the lecture again.

We have four learning objectives, so that, after studying the material covered in this lecture, you should be able to do the following.

1. Identify the advantages of using mobile devices to collect survey data.

2. Explain how computer vision can be used to diagnose three types of disease or infestation.

3. Explain how samples of diagnostic data can be interpolated to generate maps showing the density of disease.

4. Explain how diagnosis and mapping can be combined to yield more effective ways of monitoring and tracking crop disease, specifically by building spatial and spatio-temporal models the incidence and spread of diseases.

5. Explain how spatial and spatio-temporal models can also be used to optimize the use of survey resources.

Slide 1 Welcome to Lecture 3 of Module 2, in which we are examining case studies of the application of AI and machine learning in Africa.

In this lecture, we will focus on a case study in agriculture, looking at the ways that computational techniques can be used to monitor crop disease in the developing world, with specific examples in Uganda.

Slide 2 This case study is based on an article written in 2013 by John Quinn, entitled: "Computational Techniques for Crop Disease Monitoring in the Developing World".

It demonstrates how data analysis techniques can be used to improve the

Speed
Accuracy
Cost-efficiency

of conducting surveys of crop disease in developing countries using cassava and banana crops in Uganda as examples.

Slide 3 The economies of many developing countries are dominated by an agricultural sector in which smallholder and subsistence farmers are responsible for most production, utilizing relatively low levels of agricultural technology.

Slide 4 Disease among staple crops presents a serious risk, with potentially devastating consequences.

Monitoring the spread of crop disease is essential as it allows for targeted interventions to be planned

and provide early warning of risk of famine.

Slide 5 To collect data on crop diseases, the standard approach in a country such as Uganda is to send teams of trained agriculturalists to farms, where they assess the health of the crops.

To accomplish this, the teams collect data from multiple site locations and fill in paper surveys.

On their return to base, this survey data is then used to infer the incidence and spread of the disease across the area of interest.

Slide 6 This process is expensive, untimely, and often inadequate

The reasons include

the scarcity of suitably trained staff,

the often-difficult logistics in arranging suitable of transport,

and the time it takes to collate the survey data and compile reports.

Slide 7 The challenge, therefore, is to identify and track the spread of viral crop diseases in a manner that is cost-effective, timely, and accurate.

The target article addresses this challenge by using data collected using digital surveys on low-cost Android phones,

performing automated disease diagnosis,

monitoring the spread of crop diseases,

and optimizing the use of scarce survey resources.

Slide 8 The collection of data in digital form

allows the symptom measurement and the diagnosis to be automated

using the results to create a spatial map of the incidence and spread of the disease

Slide 9 Android phones, costing less than $100, are used to carry out crop disease surveys.

The Open Data Kit is used to create the digital survey,

replicating the paper-based surveys,

while also making it possible for survey teams to collect image data and GPS coordinates.

Slide 10 There are many advantages to this approach.

The time needed to do data-entry is reduced.

Results are immediately available.

Data collection can be undertaken by survey workers with only basic training

And experts are not required to travel to the field because images can be collected and assessed by them remotely or by software on the phone.

Slice 11 A typical national-scale survey of cassava disease in Uganda, for example, would require expert assessment of the status of the disease and the level of symptoms for around 20,000 plants.

For example, to assess the cassava plants for diseases, the survey teams have to examine the roots and leaves.

To assess the roots, root samples are assessed to determine the degree of necrosis.

Samples are assigned to one of five categories, ranging from completely healthy to completely necrotised.

Classifying the intermediate levels of necrosis is difficult and results vary with different surveyors.

Automating the classification process produces more accurate, standardised results.

Slide 12 To assess the leaves, surveyors have to to count the number of whiteflies on the underside of cassava leaves.

This is time-consuming and prone to error, often leading to the generation of varying and inaccurate reports from the different experts.

Slide 13 Computer vision is then used to analyze the images and identify the level of necrosis in cassava roots.

This is accomplished by using image pixel features to segment the image and label the pixels accordingly.

Slide 14 Similarly, the number of whiteflies on the underside of cassava leaves can be counted by segmenting the image into pixels that correspond to the leaf and the whiteflies.

Counting the number of whiteflies then becomes an exercise in so-called blob detection.

Slide 15 The diagnosis of viral disease from leaf images can be accomplished by training a classifer using labelled training data.

The author of the target articles achieves good results for banana and cassava plants using on color histogram features.

A color histogram is a representation of the number of pixels of each color in an image.

Slide 16 This simple approach lends itself to being implemented on mobile phones.

Automation of symptom measurement and disease diagnosis using images taken by a camera phone makes it easy for surveyors on the field to receive real-time feedback on the disease affecting a given crop.

Slide 17 Since GPS coordinates are also collected as part of survey, the diagnosis data can be used to create a risk map.

Observations are made at a small number of sample sites. Using spatial interpolation, these samples are used to infer the distribution across the entire spatial field of interest.

Slide 18 One can use different approaches to interpolate between the known sample values.

For example, Nelson et al. (1999) used Gaussian process regression to generate a spatial map showing the incidence of disease (on the left) and a confidence map of the incidence estimates (on the right). Locations closer to the sample points indicated by white dots show higher confidence values, as you would expect.

Gaussian Process Regression (GPR) is a form of non-parametic machine learning, i.e., it is not limited by a functional form. It calculates the probability distribution over all admissible functions that fit the data.

For example, when provided with geographical weather data that is incomplete, a GPR model can be used to generate weather data for unobserved locations.

The same applies for the incidence of crop disease.

We will return to Gaussian processes later in this module in Lecture 6 on the last case study.

Slide 19 Quinn et al. (2011) also used Gaussian process regression to generate a spatial model that allowed them to map out incidences of crop diseases in Uganda,

along with estimates of severity,

ranging from a healthy crop to one most affected by the disease.

Slide 20 Estimating the density of an infectious disease in space and diagnosing that disease in individual cases are generally done independently.

That is, the diagnosis is not usually formally coupled with estimates of the risk of the disease, as given by, for example, a spatial model

However, this risk estimate provides a very useful prior, that is, known information that can impact of the diagnosis.

In turn, individual diagnoses can be used to update the risk map.

Slide 21 Using the approach described in the case study, this is possible because the data collection devices are networked and their locations are known.

This combined inference of spatial disease density and diagnosis in individual cases can be done with multi-scale Bayesian models

(recall, we met the concept of probabalisitic Bayesian models in course AIML01, Module 2, Lecture 3 on statistical machine learning.)

Slide 22 This is a win-win situation: it improves the accuracy of the risk map and of individual diagnoses because the uncertainty in both tasks is jointly-modeled.

Slide 23 But that's not all.

A probabilistic spatial or spatio-temporal model can also be used to determine which locations would be most informative for collecting new data.

This is not possible in the traditional paper-based survey system because data entry happens after the surveyors return home from their trip to the farms.

The approach in the target article also allows models to be learned in real-time, as data is collected.

Slide 24 This problem is essentially active learning

in which we collect data from locations in which the model has the lowest confidence

or, alternatively, the locations that provide the most information,

where information means reduction in uncertainty, just as it is defined in information theory

Slide 25 This would suit situations where phones are given to agricultural workers across the country

Rather than by experts travelling in the field,

Effectively, crowd-sourcing the survey

Slide 26 What about the experts that are conducting a survey out in the field?

They have a limited amount of time, limited budget, and limited number of survey visits.

The model can also be used to guide surveyors in the field so that they collect more valuable data while keeping fixed their budgeted number of samples or visits.

Slide 27 If the only constraints we have are that the experts are able to travel any paths in a given road network with some given budget, this would be a difficult optimisation problem.

However, the road network is often sparse in rural parts of the developing world.

This makes it reasonable to assume that survey teams will follow a given route.

Slide 28 This route corresponds to a one-dimensional manifold *R* within the spatial field.

A manifold is a mathematical - topological - way of representing some high-dimensional structure in a lower-dimensional space, thereby simplifying the analysis of that structure.

With a fixed survey budget allowing k stops, we would like to identify the points along R that maximise the informativeness of the survey.

Under this constraint, optimisation is tractable with a Monte Carlo algorithm

A Monte Carlo algorithm is a mathematical technique based on statistical random sampling to estimate the optimal solution to some problem.

Slide 29 The optimal location for the next sample is recomputed after each stop,

given on the spatial model

given the most recent observation.

Slide 30 The use of AI to automate disease diagnosis and tracking has the following advantages

1. The use of Android phones to collect data is inexpensive compared to the use of paper-based surveys.

2. By automating the disease diagnosis process, experts no longer needed to be on site for crop evaluation.

3. Using the risk map, probable diagnosis of plants with known locations could be accomplished more easily and more reliably.

4. Data collection in the field is optimized by having the survey teams collect data from locations which would be informative for the model.

Slide 31 A spatial model is not able to provide disease forecasts that could be used to inform decision-making to implement disease prevention measures

However, the approach can be extended to include temporal information, yielding a spatio-temporal model.

A spatio-temporal model is one which contains data elements of time and space, making it possible to describe an event as having occurred at a given time t and in location x.

To summarize:

1. The ability to track the spread of viral crop diseases in developing countries can help governments to put in place measures to mitigate disasters such as famine, and to do so effectively and in a timely manner.

2. The approach described in this case study makes use of digital surveys administered using Android phones to collect data which is then analyzed using novel data analysis techniques to improve the speed, accuracy, and cost efficiency of crop disease surveys.

3. The model makes it possible to generate real-time diagnosis of crop diseases, map out the location of crop disease outbreaks, and make optimal use of fixed budgets given to survey teams by identifying optimal survey locations.

Here is the target article used for the case study. Please read it carefully.

Quinn, J. (2013). Computational techniques for crop disease monitoring in the developing world. In Advances in Intelligent Data Analysis XII (pp. 13–18). Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41398-8_2

And here are the references cited to support the main points in what we covered in this lecture.

Nelson, M. R., Orum, T. V., Jaime-Garcia, R., & Nadeem, A. (1999). Applications of geographic information systems and geostatistics in plant disease epidemiology and management. Plant Disease, 83(4), 308–319. https://apsjournals.apsnet.org/doi/10.1094/PDIS.1999.83.4.308

Quinn, J. A., Leyton-Brown, K., & Mwebaze, E. (2011). Modeling and monitoring crop disease in developing countries. https://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/viewFile/3777/4083