Computational intelligence & modeling of crop disease data in Africa

Owomugisha, Godliver

DOI:
10.33612/diss.130773079

Publication date:
2020

Chapter 1

Introduction

In the 21st century, data has been termed as the new 'oil'. Data is the world's most valuable resource hence giving rise to a new economy according to the Economist (2017). This data comes in numerous structured, semi-structured or unstructured formats, e.g. as images, time series, spectra, clinical data, sensor measurements etc. Data also comes in different aspects: volume, velocity and variety. Often we are presented with challenges of how to turn real world data into meaningful information. Traditional data processing application methods have been found inadequate in handling big data complexities associated with diversity and massive scale. Thus advanced analytical techniques and technologies have been adopted in recent years. But is there a defined art in using the data analytic tools?. Let us consider the songwriter example from (Peng and Matsui 2016): "Imagine you were to ask a songwriter how she writes her songs. There are many tools upon which she can draw. We have a general understanding of how a good song should be structured, how long it should be, how many verses, maybe there is a verse followed by a chorus, etc. In other words, there is an abstract framework for songs in general. Similarly, we have music theory that tells us that certain combinations of notes and chords work well together and other combinations don't sound good. As good as these tools might be, ultimately, knowledge of song structure and music theory alone does not make for a good song. Something else is needed."

Just like songwriting, data analysis is an art and data science presents us with so many tools at our disposal but the challenge is always in finding the right tools that suit your problem set.

Machine learning is one broad area of Artificial Intelligence that presents us with numerous tools that give us the ability to learn from observations and make better decisions for the future, e.g. (Bishop 2006, Goodfellow et al. 2016). The primary aim here is to allow computers to learn automatically without human intervention or assistance and adjust actions accordingly. Machine learning algorithms are often categorized as supervised, unsupervised, semi-supervised and reinforcement learning. Supervised learning can be divided into regression and classification problems, mainly. Whereas regression is concerned with the prediction of continuous quantities, the outputs for classification are discrete class labels. Typically, in a classification problem, both input (X) and output (Y) variables provided and we use the algorithm to learn the mapping function (f) from the input to the output $Y = f(X)$. The ultimate goal is to approximate the mapping function so well that when you present new input data (X), you can predict the output variables (Y) for that data.

In contrast, unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled. One common application of this algorithm is in clustering challenges and methods such as K-means, mixture models and hierarchical methods (Arthur and Vassilvitskii 2007, Johnson 1967) fall under this technique.

Semi-supervised learning is a relatively new paradigm in machine learning that has been used in applications where labeled data is expensive to obtain. Typically the algorithm uses a small amount of labeled data with a large amount of unlabeled data e.g. (Peikari et al. 2018).

Lastly, in reinforcement learning, an agent learns how to behave in an environment by performing actions and seeing the results, see (Sutton and Barto 2018, Lagoudakis and Parr 2003). A good application of this method is in robotics and in recent years, a lot of improvements have been seen in this fascinating research area.

The applications of machine learning techniques in general are found in different fields ranging from medical image analysis, cyber security system, climatic and weather forecast, online marketing among others. In this thesis we present the use of supervised learning techniques in crop disease diagnosis, an area that has gained attention in recent years.

Agriculture is a core sector especially in the sub-Saharan Africa countries but the practice is hindered by so many factors with crop pests and diseases among the biggest challenges. Often, prevention measures have been put in place and agricultural experts are required to carry out surveys in different regions of the country. This process is not only labour intense, but judgements by experts may be subjective and disagree frequently.

Therefore, for an objective surveillance process or in places where experts are not available or where farmer knowledge is insufficient, other methods for carrying out field-based diagnosis are a critical need. Computational work in this area has been towards automating this process through building machine learning models that can take an image of a leaf and predict whether the plant is infected with a particular disease or not. Both the physical inspection of plants and methods that build on image data, use visual symptoms to relate to particular diseases in the plant.

The main novelty of this thesis project lies in how we identify diseases in plants before symptoms are visibly seen with our naked eye. The investigations in this area included the use of spectrometry for early disease detection. In a natural world, all objects have a degree of reflection and absorption when light strikes on them. Here we explored the characteristic changes that can be observed in a leaf as a result of disease. However, spectral data comes in higher dimensionality. Another core aspect of this research is feature selection and dimensionality reduction for a technology that is aimed to be deployed on weaker powered devices such as a mobile phone. Integrating computational techniques such as LVQ where a lot of success has been shown in feature selection was important to this particular problem. Therefore, the work presented in this thesis is inspired by previous research (Mwebaze et al. 2011, Mwebaze et al. 2015, Mwebaze and Biehl 2016) that investigated on diseases that affect Cassava (Manihot esculenta), the second most important crop in the Sub-Saharan Africa, thus our focus on the same crop. Over the different chapters in this thesis, we present the contribution of this study aiming at early diagnosis of cassava diseases which allows for early action measures.

1.1 Scope of this thesis

We begin the thesis by giving a general introduction of the algorithms that have been employed. To a larger extent, most of our research questions were answered under a prototype based scheme with GMLVQ. The classification algorithm has previously displayed favorable performance in related studies, e.g (Melchert et al. 2016a, Melchert et al. 2016b).

The rest of the thesis is presented in two parts. In part one, we reviewed and developed algorithms to do the automated diagnosis of diseases in cassava plant based on plant images.

This part was built on previous research done in (Mwebaze et al. 2011) to classify between healthy and diseased crops based on their images taken with a phone camera. Guided by a research question: Can we reliably distinguish between diseased plants that are symptomatic from healthy plants?. The work presented in this part of the thesis extends the previous research in two areas: (i). We transfer the method from a two-class problem (healthy vs. diseased) to a multi-class classification problem representing four diseases common in cassava.

(ii). We enable the classification to handle different severity levels for each disease as well as expanding on the dataset to contain images from the different diseases and their corresponding severity levels.

The second part of this thesis focuses on diagnosis of disease in crops before they become symptomatic by use of spectroscopy. Spectroscopy is one of the most widely used technologies in a variety of scientific fields. Here, we use visible and near-infrared spectral analysis to do non-invasive diagnosis of plants before they manifest disease, thus allowing early intervention measure. The research area is guided by our hypothesis that crop diseases cause several metabolic changes in the biology of the leaf that can be detected at an early stage using spectrometry. Core research questions in the area include: (i). Can we distinguish between healthy plants and plants that are diseased when no symptoms are yet visible?. (ii). Is there a different spectral distribution for the different diseases we are looking at?. (iii). Is there a range of wavelengths that is most sensitive to each of the different diseases?. We answer the question of early detection by time (in weeks) before symptoms are visible to the human eye. In the end, we present the design of a low cost 3-D printed smartphone add-on spectrometer targeting small holder farmers to be able to do field diagnosis. The next section presents the thesis outline.

1.2 Outline

Chapter 2 provides a general introduction to Learning Vector Quantization and relevance learning. Chapter 3 is an extension of our previous studies where we used a dataset of photographic images. This part generally extends the disease classification to a multi-class problem. We use standard techniques to classify different severity levels for the four (4) cassava diseases for a system that can implemented on a smartphone to be used by farmers in remote places. The chapter also presents results on different feature extraction techniques for image data.

Chapter 4 presents a preliminary study on spectral data for disease diagnosis. Here we collected data from mature plants aged 6 - 9 months. Our analysis compares the two types of data for disease diagnosis. Our dataset is composed of cassava leaf images captured with a smartphone camera and their corresponding spectrogram using a spectrometer. The chapter also gives a general introduction to spectral data, the pre-processing techniques and a pipeline of these methods which we use in other chapters.

In Chapter 5, we investigate the spectral properties associated with different diseases aiming at understanding the sensitivity of the wavelength bands in relation to the diseases identification. The aspects of matrix relevance learning, feature selection and dimensionality reduction are discussed here.

The novel part of this research is presented in Chapter 6. We discuss a scientific experiment that was conducted in a controlled environment guided by the biochemists. The main objective here was to diagnose diseases before symptoms can be seen by the human eye. We therefore present the experimental setup, the tools, methods and results of the study.

In Chapter 7, we discuss the construction of a low-cost 3-D printed smartphone add-on spectrometer. Another objective is to transform all this knowledge back to a mobile phone, a tool that can be used by farmers in their gardens. We present the architecture of our designed prototype and findings of our investigation.

Finally, Chapter 8 presents a brief summary of the entire research and a collection of ideas for future work and investigation.