Machine Learning for Large Scale Farming

Zachariah W. Litzinger
Division of Science and Mathematics
University of Minnesota, Morris
Morris, Minnesota, USA 56267
litzi005@morris.umn.edu

ABSTRACT

Identifying which plants have not had their needs met can be difficult to do on a large farm. Using human input to identify the needs of these plants can be relatively time consuming and expensive. Even when using humans is not the issue, the number of varieties of plants often requires a number of experts for each type. Finding this many experts can be impractical for every farm. Identifying the needs of crops and sorting plants by type without human input would be a way to increase the efficiency of farming.

In this paper we will look at using machine learning to achieve these goals. Using both supervised and unsupervised techniques we can classify plants into pre-identified categories, and even sort plants without previously known classes. Looking at support vector machines and k-means clustering shows an important step to the next step towards full farm automation and improving the long term sustainability of large scale farming.

Keywords

Machine Learning, Farming, Phenotyping, Need Detection, Supervised Learning, Unsupervised Learning, Support Vector Machines, k-means

1. INTRODUCTION

Farming was an important step allowing us to sustain large human populations. Farming allowed few people to sustain many people, freeing others to specialize in the different aspects of life. However, large scale farming can quickly become inefficient by missing the needs of certain plants. Whether it's disease, nutrients, or even water, missing the needs of these plants can have large impacts on the results of production.

Farmers with enough experience eventually learn to recognize the needs of a plant just by looking at it. This type of monitoring becomes inefficient at a large scale because a single person can only monitor so many plants at a time. When monitoring by sight, a farmer may have to hire more inexperienced helpers to help with this monitoring. However, this quickly becomes expensive and runs the risk of human error coming in to play, especially since the helpers likely do not have as much experience. This can mean the crop is not reliably having its needs addressed. Without knowing the needs of an individual plant, a farm may have to pre-treat their crops for diseases and pests. This means treating every plant for something it may not need, potentially wasting a resource that could be used more efficiently and increase the environmental impact of these treatments compared to spot-treating. One of the goals of using machine learning in a farming environment is to find a method of classifying the needs of plants in a way that allows for spot-treating, the treating of plants only when needed.

> "Sustainable intensification" — producing more food from the same area of land while at the same time reducing the environmental impacts— demands innovative agronomic practices. Precision agriculture strategies (the integration of different modern technologies like sensors, information and management systems) can reduce the ecological and economic impacts in agricultural crop production.
>
> —Behmann et al. [1]

Much like the way humans learn by experience, we can use machine learning to build computer systems which can recognize the needs of crops to improve response time and maximize efficiency. At its essence, it is a method to quickly recognize patterns. Machine learning is being used to optimize search results and routing for maps while also being used to recognize features in images.

Machine learning recognizes features by looking at a large set of data and recognizing similarities between the data and categorizing it. For example, in the context of farming we may want to define different categories of similar looking plants. It can also be configured to categorize based on predefined categories. This could be used for categorizing already harvested plants or identifying ones with a specific need. Both can have further uses.

First we will cover some background of machine learning and cover some different methods we can use machine learning to speed up already necessary processes. Then we will talk more specifically about each part of machine learning and give a description of their basic implementations. Finally we will discuss the effectiveness of our techniques on the use cases with examples from the papers referenced in Section 5.

UMM CSci Senior Seminar Conference, April 2018 Morris, MN.

2. BACKGROUND

There are many different ways machine learning can be used, but the best choice for any system largely comes down to the different learning, or training, methods: supervised and unsupervised learning. Supervised learning is training based on previously labeled data with previously decided categories. For this method we split our labeled data into training data and testing data to ensure our method is generalized to future data. This can allow for identification of these specific categories as a generalized rule for future data, while keeping human input to just our training data, whereas unsupervised learning is the identification of categories based on similarities that the machine implementation finds. This can bring to light different features that may be necessary to watch out for in the future. Human input is only necessary to identify what each of these newly defined categories mean. Both methods decrease human input to a specific set of categorization needs that a human must initially interpret.

The implementation of machine learning in a production farming environment can allow the efficiency of crops to improve just by making the recognition of these categories more accurate. It won't initially make decisions on what to do about these categories. Those decisions still require a human to understand the categorization and act accordingly. However, in the future, robots could be implemented to respond to certain categorizations. This would further remove the manpower necessary to deal with a large farm by letting the mental side of recognition and the physical reaction be put fully on some system, or set of systems of machine learning techniques combined with robotics.

The uses of machine learning this paper explores focus on the categorization of plants in a faster manner to improve the efficiency of modern farming. Farms can use these methods to prevent their crops from being impacted by diseases and pests. They can also categorize their plants into different types to identify weeds, or to identify phenotype.

2.1 Phenotyping

Defining different plants by phenotype can be useful when looking for a specific feature of a certain plant. Many social practices have caused organic food to become more popular, such as healthy eating or concern for the environment. This causes many farms to grow both organic and conventionally grown plants. Due to regulation, phenotyping these plants accurately is a necessity or the organic plants may not be permitted to label themselves as organic. Kessler et al. looked at classifying individual wheat plants to recognize whether they were organically or conventionally grown [2] using Support Vector Machines, which we will discuss in Section 3.

2.2 Identifying Stress

Plants are susceptible to a variety of diseases and damage from insects. To increase productivity of a crop, one must identify the diseased area to efficiently treat the infected plants and avoid over-treating unaffected crops. Puig et al. [3] discussed the method of using a k-means clustering algorithm, which we will discuss in Section 4, to identify damage from white grubs, which caused damage to the roots of sorghum plants. These grubs eat the roots of sorghum plants which reduces nutrient intake and causes unhealthy plants, potentially killing them. In a separate research study, Behmann et al. [1] looked at the ability to identify general biotic stress in plants by looking at the overall health of the plant to identify the plants that may need looking at. Whether the biotic stress is disease, pest, or an invasive plant, it is important to identify and act on this stress early, to prevent any loss of health to the plants.

2.3 Input Methods

Distinguishing one plant from another can be difficult for a person to do without a deep prior knowledge of the two plants. On a large scale, using humans to tell the difference between every plant is impractical. Using an external mechanical input method, known as remote sensing, can allow for a large amount of data to be collected, and then be analyzed later.

These methods of remote sensing include images of individual plants and aerial images of the crop. These images can be taken using different spectra or by identifying chlorophyll florescence levels (the specific amount of light absorption in a plant) [1].

However, machine learning is not limited to just doing image processing. This can be expanded to methods varying from using a variety of sensors to detect the differences of nutrients in the soil, to using sensors to identify the conditions of the air surrounding the plants. For instance, Kessler et al. [2] use more than "300 gas chromatography-mass spectrometry measurements" as their input data [2]. These measurements identified different metabolites that were present in each plant. For our uses, consider metabolites to be intensities of different color wavelengths of each plant being burned.

Whether looking to identify plants based on a desired characteristic or looking for warning signs that something is different than normal, both supervised and unsupervised learning can be used to improve the efficiency of these classifications.

3. SUPERVISED LEARNING AND SUPPORT VECTOR MACHINES

Supervised learning is a method of machine learning that commonly creates a classifying method for a set of data that has predefined classes. In general this is a method to match the data to the predefined classes, and then applying this trained method to classify new data.

Using supervised learning we can identify previously defined categories. This is especially important for our first goal phenotyping, as we discussed in 2.1. We can also use this method to identify different plants, such as identifying organic and conventional wheat crops [2]. Supervised learning is not limited to these categories. It can attempt to identify any set of classes as long as the data is predefined.

Support Vector Machines (SVMs) are a form of supervised learning that attempts to split two predefined classes with a linear classifier. A linear classifier is a linear function that gives the optimal separation between the data [6].

3.1 Linear Classification

Assuming the two classes are separable by a single vector (hyperplane in any dimension) with no overlap, we can choose the two points closest to the opposing class and draw two parallel lines. The line that bisects the plane between the original lines is our main defining vector. The side vec-

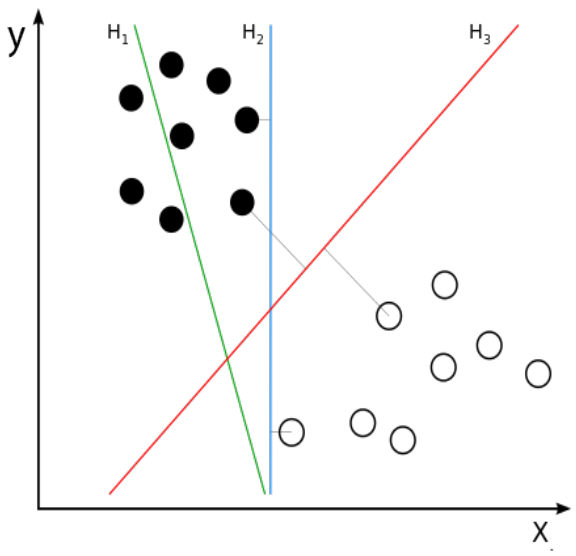

Figure 1: Three potential linear classifiers; H_3 being the best one as discussed in Section 3.1. Image from [6].

tors are our support vectors, and our classifying vectors. These support vectors are why a SVM is called a Support Vector Machine.

Looking at Figure 1 we see that there are two different classes of points: black and white. The three classifiers displayed are not all reasonable. We can see that H_1 is a bad classifier because our points are not divided by class. H_2 does divide our classes but it doesn't seem reliable because points that a human may cluster with one class could be identified as the other. Finally, H_3 is our best classifier since it separates the classes with the largest margin possible. This minimizes the possibility of a new point being misclassified.

As shown in Figure 2 we want a classifier that is the bisector of the two hyperplanes, our dotted lines in two dimensions, that separate the data as much as possible. This hyperplane can be described with the vector $\vec{w}$ where $\vec{w}$ is the normal vector to the hyperplane, or solid line in our example. Now that we have our linear classifier we can identify the class of each point with this equation:

$$\vec{w} \cdot \vec{x} - b = c$$

where $\vec{x}$ is the normal vector to our classifying hyperplane, the equivalent of a line in any other dimension, and $\frac{b}{||w||}$ is the minimum distance that our classifier is from the origin.

To classify each point we define our class by identifying which side of our line we are on. We find our class by plugging in our data point vector for $\vec{x}$ with our normal vector and offset value, b. Then we separate the classes by the sign of the output c. So In Figure 2 the black points would have a positive sign for c and the white points would have a negative sign for c. If the value of c was 0 the point would be part of our classifier and we would not know what to do with it unless we defined all zero values to have a certain sign, such as positive [6].

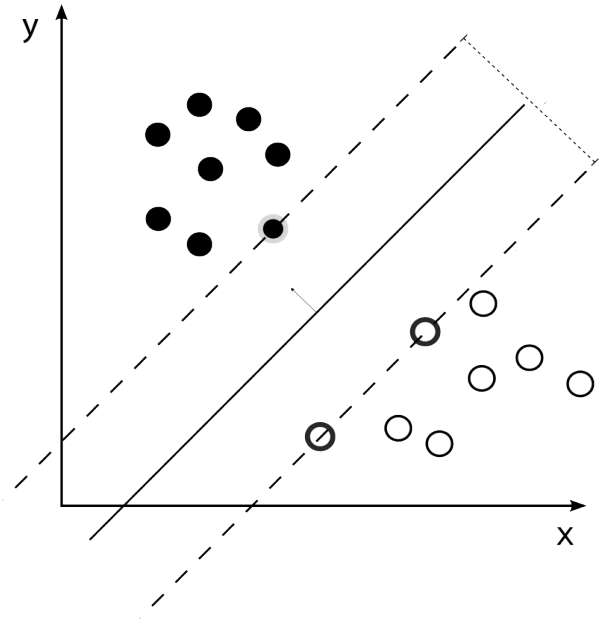

Figure 2: Our best linear classifier with support vectors displayed as dashed lines. Image from [6].

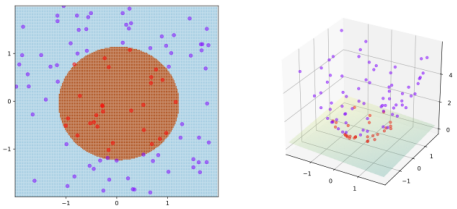

Figure 3: Example of a quadratic kernel function altering data to allow for linear classification by an SVM on data that is non-linearizable (a circle). Image from [5].

3.2 Nonlinear Classification

The data provided may not be possible to separate with a linear classifier. This means that the linear classifier must be defined in a higher dimension than the data is naturally in. This requires the definition of a kernel function. A kernel function is a function used to map data that is non-linear to a higher dimension to allow for linear separation, as seen in Figure 3. This is why it was important to think of our vectors before as hyperplanes, it allows us to work in any dimensional space if our kernel function forces it.

The general form of a two dimensional kernel function that takes in and returns a pair of values is:

$$\Phi(x, y) = \langle \phi(x), \phi(y) \rangle$$

However, if data is non-linearizable naturally, like in Figure 3 we must use a function that expands our data into another dimension to allow a linear classifier to work. If this is the case, we can use a quadratic kernel function to expand this data into three dimensions. This is of form:

$$\Phi(x, y) = \langle x, y, x^2 + y^2 \rangle$$

Note that the output has three dimensions, rather than just two. As you can see in Figure 3, the output data allows a linear classifier in the form of a plane to separate our data into the classes we wanted.

Kernels are not easy to find and require a lot of trial and error. A more complicated function can be used to transform the data such as a Gaussian function, which will not be explained in depth in this paper.

4. UNSUPERVISED LEARNING AND K-MEANS CLUSTERING

Unsupervised learning is a method of machine learning that creates its own classes rather than using predefined classes. In general it is a method to create classes out of a set of data to see how the machine classifies things we may not have previously thought belonged together.

This is a very important method for identifying features of a plant that were previously unnoticed. Using unsupervised learning in a farming environment can allow a farm to identify sources of biotic stress by looking at each of the classifications given by the algorithm, and identifying the key features the system created and interpreting them for best use. This may also be more helpful for researchers wanting to identify new similarities, or used by users that want to simply identify when leaves are a different color to identify health.

As an unsupervised method, k-means clustering is very similar to Support Vector Machines but creates divisions entirely algorithmically and defines its own clusters rather than being based on desired classifications.

To initialize a k-means analysis, the algorithm chooses k random points within the domain. These points will be the initial means that we want to adjust to get a more accurate model. The number of points we choose will correspond to the number of clusters we want to organize our data into. To define our clusters, the mean closest to our data point is the class it resides within. Next, we update our classification. This requires us to find the centroid of each class and choosing that to be our new mean to evaluate at. In our context the centroid is defined as the average point within our class. You can find our centroid, C, by adding all of our points, x, together and divide by the number of points, p in every cluster:

$$C = \frac{\sum_{n=1}^{p} x_n}{p}$$

So in each iteration we find the centroid of our current class structure and redefine our mean to be each new centroid. Then we re-classify the points by assigning to them the class of the closest new centroid. This will bring us closer and closer to separating into categories until the changes of the mean values between iterations approach zero. This is when we finish the iteration, and use those means as our model for categorization.

Looking at Figure 4 we can see the initialization of the k-means algorithm on the left. These points were three randomly selected and do not represent a good set of means. To test if it is our final set of means, we use the centroid equation on each region, and redefine our means as that centroid as shown in the right picture. Then we would progress forward until the centroid does not change. Note that although the example in Figure 4 is in two dimensions, finding these clusters can be done in any dimension.

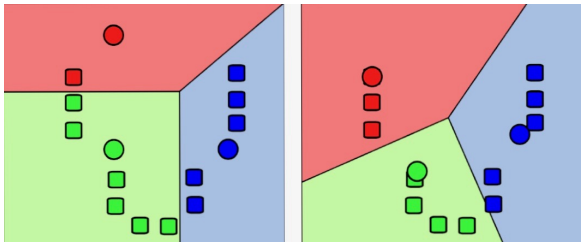

Figure 4: First step of a k-means iteration. From [4]

SVM Processing Conventional/Organical Grown Wheat		
Year Trained On	Year Tested On	Accuracy
2007	2007	0.97
2010	2010	0.88
2007	2010	0.55
2010	2007	0.56
2007, 2009, 2010	2007, 2009, 2010	0.90

Table 1: Kessler et al. Results [2]

It can be noted that there is no promise that this algorithm will create clusters in any way that makes sense. However, frequently it can display new features that one did not previously expect [4].

5. RESULTS

This section will be a discussion of the success of the two methods we talked about in this paper. We will look at the results of more specific studies that used these methods and what they were trying to do.

5.1 Effectiveness of Supervised Learning

5.1.1 Phenotyping

Kessler et al. [2] analyzed metabolite values for the input data instead of image data. In previous work they had used metabolic profiling techniques. Using the same input data allowed them to more easily check whether their new method was as accurate as their older method. The benefit of using a machine learning approach is to not rely on only a single, previously known metabolite as a marker of a given plant. Instead it can use the entire metabolite measurements of any plant to more accurately define a phenotype. We also note that their method was destructive to the plant. This was offset by using samples of presorted plants and testing a small subset of them to classify the identity of that set. Note that Behmann et al. [1] discussed the feasibility that image data would be a reasonable way to find similar results.

Kessler et al. [2] analyzed 313 samples of wheat based on whether they were conventionally or organically grown. Their input data was not linearly separable originally, so a Gaussian kernel (a kernel using a gaussian function) was used on their 36 metabolites, creating a linearizable 35 dimensional plane. This was shown to be accurate up to 90% of the time within a single year as shown in Table 1. Considering the small number of samples but large dimensional space, this is considered to be very accurate. This is ex-

citing because conventional and organic wheat can then be identified and processed separately with this method.

However, it would be ideal if we could use the training data from one year to the next. Kessler et al. [2] noted that with their input data the accuracy between other years was only 55 percent. This is not very impressive. This means the training data will need to be sourced from the same years. The changes between years are likely due to factors like the weather changes within years. However, it is possible that a larger data set encompassing many more years than just three may be able to classify the different plants into more subsets, giving more accurate general classifications. In the future, it may be helpful to note that these methods are more accurate when using more specific data points that are already verified to be unique by biologists. With enough data from separate years, there may be a generalized method that can classify future years accurately.

Adding more data to a system will most likely create a more accurate categorization. However, other ways we could make this more accurate for general data would require more research to identify consistent identifiers that can be used in this method. This will allow our system to train using more generalizable data and it may become more applicable on all new sets of data.

5.1.2 *Biotic Stress Identification*

Support vector machines can also be used to identify disease. Behmann et. al. [1] analyzed the method of using SVMs for weed detection and nitrogen deficiency in corn. Their method started with image processing. They used the scaled addition of RGB color values over all pixels in each image, giving a general summary of the image. The linear classifier was a plane since the three color values cause the SVM to operate in three dimensional space. They stated that SVMs were generally 97% accurate. This was a high enough accuracy to be considered correct. They also discovered that when a shape classification was put into place (another machine classified the shape and added that to the data set as another dimension) the classification accuracy would increase with smaller grain plants, which were only accurate 85% of the time. Using these methods Behmann et al. [1] found that there was an increase of weed control by 85-98%. This showed that "the amount of herbicides applied was reduced by between 8-81%."

This was wide variability based on the needs for pesticides of each type of plant. However, it shows great improvement in the ability to spot treat for weeds on a large farm rather than treating full crops.

5.2 Effectiveness of Unsupervised Learning

5.2.1 *Biotic Stress Identification*

Recognizing insect damage on crops by hand can be a long and arduous process if the input is large enough. So using k-means is a great way to separate the data into easier to handle pieces which can be interpreted by hand. After taking in the image data, Puig et al. [3] use the k-means method to identify where there are damaged crops in images. The differences from pixel to pixel are categorized from the k-means computation and then mapped back onto their crops to visually show the location of where there seem to be problems as seen in Figure 5. In these images it is shown that out of the 6.09 hectacres of land there are 3.25 hectacres of healthy crops, 1.71 hectacres of dead crops, and 1.13 hectacres of crops that are damaged (transition areas).

> The location and distribution of transition areas is relevant information in order to design a site-specific control strategy.
>
> —Puig et al. [3]

Behmann et al. [1] also discussed this method. Since k-means don't require much training data but can form its own classifications easily, it is very useful for creating models for data that aren't known about.

5.2.2 *Phenotyping*

Although the discussion of phenotyping wheat with Kessler et al. focused on Support Vector machines [2], with our previous discussion of k-means, we can see how we might be able to extend the work that Behmann et al. [1] method to future implementations of the research Kessler et al. [2] did. If the years were not given initially, then use another machine to identify phenotype. We can see how this may be practical in Figure 6.

By looking at the data we can easily see how k-means may be able to classify these three different data sets. Then the data appears to be seperable again by a second clustering algorithm on the initial clusters. This is just speculation however and definitely requires further investigation.

6. CONCLUSION

Improving the efficiency of manpower will always be a discussion in any job. For farming, machine learning could be a route to lessening the amount of manpower to identify the needs of plants and allowing a person to more efficiently know the information they want to know. Whether much information is known about our data or not, machine learning has different methods to address the different classification needs in farming.

Using Supervised Learning to phenotype your corn, distinguishing the weeds from your crops, or even identifying a specific type of damage is possible as long as you have enough previously identified data to ensure that you have a more accurate model.

If you want to categorize some plants that you don't know the mix of or recognize a problem you haven't seen before, unsupervised learning will categorize your data without predefined categories. This will only require you to look at similar characteristics of each set, rather than identify the individual characteristics of each plant.

Machine learning methods can be continuously improved and fine tuned to continue to recognize the needs plants need and eventually recognize these needs without much, or any, human input.

In the future, there could be further investigation into the possibility of using a Support Vector Machine for the wheat identification on data that is both trained on many years of data, and tested on data that is outside of that scope. I believe that I would be interesting to see if a few years of data could accurately identify the data in the future.

In addition, it would be interesting to see Support Vector Machines for many classes at once. It may be interesting to see it looking for specific diseases and identify between the diseases. This would likely help with more efficient, and precise, ability to identify what is wrong with the plant while

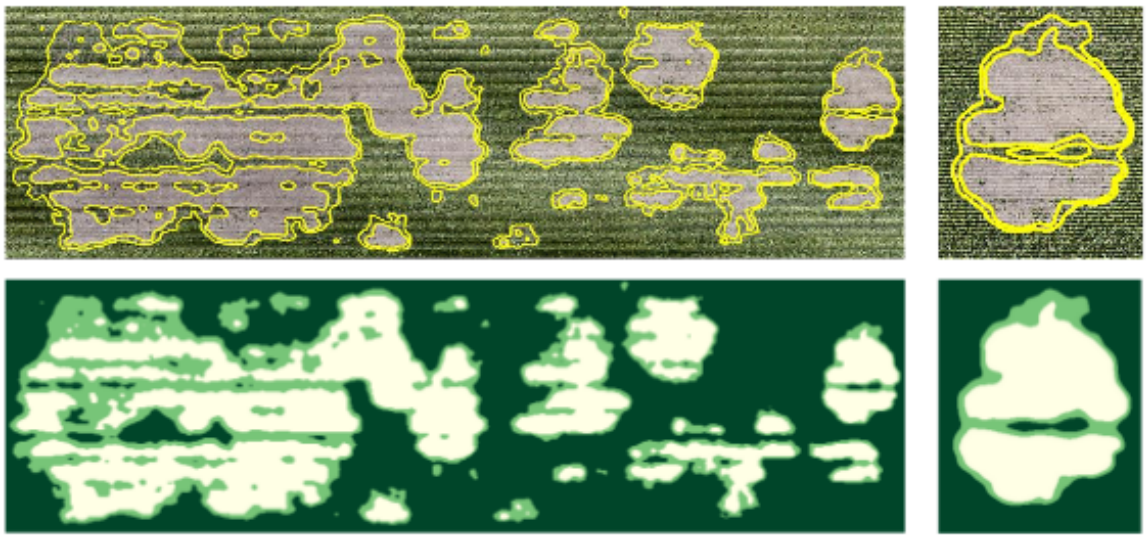

Figure 5: Above: Overlay of three identified k-means clustering classes on an areal image of damaged crops in Queensland, Australia [3]. The lower image is a display of the three different classes identified, but not an overlay. This method classified the red green and blue values of each pixel as the input data.

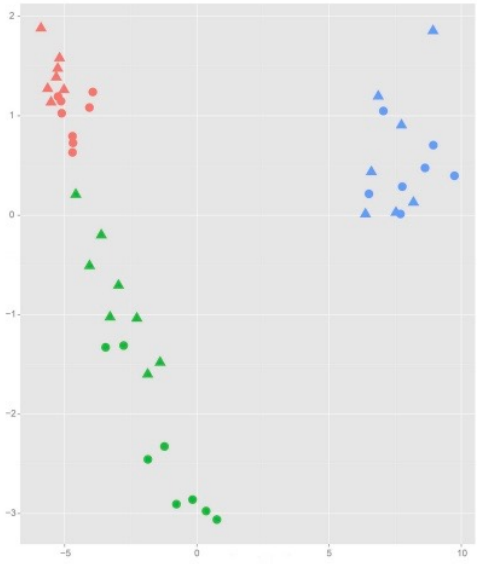

Figure 6: Comparison of amount of two metabolites of wheat referenced as PC1 and PC2 in Kessler et al. [2]. Circles: Organic, Triangles: Conventional. Red: 2007, Green: 2009, Blue: 2010

looking forward to the future for possible robotic responses. This would further reduce the human response necessary down the line. It may be possible that humans are only necessary for identifying the results of unsupervised classification in the future, if enough data is run through supervised learning machines.

7. ACKNOWLEDGMENTS

I thank my advisor, Nic McPhee, and seminar professor, Elena Machkasova, for their advice and support through this paper. In addition, a huge thanks to Kirbie Dramdahl for the review and comments as my alumni reviewer.

8. REFERENCES

[1] J. Behmann, A.-K. Mahlein, T. Rumpf, C. Römer, and L. Plümer. A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. *Precision Agriculture*, 16(3):239–260, June 2015.

[2] N. Kessler, A. Bonte, S. P. Albaum, P. Mäder, M. Messmer, A. Goesmann, K. Niehaus, G. Langenkämper, and T. W. Nattkemper. Learning to Classify Organic and Conventional Wheat - A Machine Learning Driven Approach Using the MeltDB 2.0 Metabolomics Analysis Platform. *Frontiers in Bioengineering and Biotechnology*, 3:35, 2015.

[3] E. Puig, F. Gonzalez, G. Hamilton, and P. Grundy. Assessment of crop insect damage using unmanned aerial systems: A machine learning approach. In *21st International Congress on Modelling and Simulation (MODSIM2015)*, Gold Coast, Qld, December 2015.

[4] Wikipedia. k-means clustering — Wikipedia, The Free Encyclopedia, 2018. [Online; accessed 22-February-2018].

[5] Wikipedia. Kernel methods for vector output — Wikipedia, The Free Encyclopedia, 2018. [Online; accessed 22-February-2018].

[6] Wikipedia. Support vector machine — Wikipedia, The Free Encyclopedia, 2018. [Online; accessed 22-February-2018].