Using Machine Learning to Improve Research in the Agriculture Industry

Devin Wingfield

University of Texas at Arlington, Arlington, TX 76013 USA
dtw8917@mavs.uta.edu

Abstract

Artificial intelligence (AI) has improved significantly in recent decades, and, along with it, its applications to real-world scenarios. AI has been used within a wide variety of fields like health care and e-commerce, however, AI has yet to integrate with the agriculture industry. With the help of machine learning, AI can begin to integrate with the industry via a research assistant. The model will assist researchers conduct experiments by giving treatment methods that are best suited for the experiment rather than relying on the expertise of the researcher. This will help research within the industry to become more efficient and less error prone. To accomplish this, the model will use a Knowledge Graph created by the IDIR lab that converts the large CSV files into a graph that can be queried and then summarized by the model.

Introduction

In this study, I plan to demonstrate how artificial intelligence (AI) can improve research in the agriculture industry by utilizing machine learning. It is my intention for researchers within the field to use this AI model to minimize the amount of trial-and-error experiments they have to perform. Although researchers rely on their expertise in soil science to make educated guesses about the best treatment strategies, these predictions are not always accurate. Furthermore, with the time and cost of these experiments increasing as funding continues to diminish, the right approach becomes more critical (Heisey and Fuglie 2018). In response to these challenges, I plan to create an assistant that minimizes the possibility of wasting effort and with it, the cost to commit to these experiments. This will allow research groups to overcome the financial hurdle that seems to be growing day by day as people begin to lose interest in the field of study.

Prior Work

Currently, I am working with IDIR, Innovative Data Intelligence Research, lab's faculty, and graduate students on a project called Soil Organic Carbon Knowledge Graph. With funding from NSF's Proto-OKN program and data from the United States Department of Agriculture, we aim to maximize carbon within the soil and in doing so, reduce the amount of carbon in the atmosphere. The data analyzes around 65 different fields, each with its unique focus, and is subdivided into plots with a unique experimental unit ID (UID). These UIDs help give a structured format to the data and allow researchers to identify not only the plot of land but also the purpose by relating the UID to related metadata and measurements. For example, if the UID is associated with a specific treatment ID, that UID is undergoing a combination of treatment methods specified under the treatment ID. These treatment methods include data such as crop rotation, tillage, whether the plot is irrigated, or whether residue from the previous harvest was removed. Additionally, the measurements cover a wide range of topics such as greenhouse gas emissions, weather data, and chemicals within the soil. To build this unique project, we have created an ontology that builds on the initial structure provided by USDA and subdivides the data further to allow for easier readability. Specifically, the ontology defines three structures: Classes, Data Properties, and Object Properties. Classes generalize the data based on its purpose, for example, data focused on chemical measurements will be imported as an instance of the chemical measurements class. With each instance of a class, there are nodes with Data Properties that describe the values of the instance. Finally, Object Properties simply relate two nodes in different classes based on data with the source files. By leveraging the versatile ontology we created, complicated questions such as finding the change in soil organic carbon over time can be answered. Moreover, with our knowledge graph now containing 650,000 instances and 900,000 object properties, it presents significant potential for further research and exploration. The capabilities are far greater than what we are currently using it for, and we hope that others will be able to use this KG to find solutions to problems beyond our current scope of research.

Approach

This approach will leverage the Knowledge Graph developed by the IDIR lab to analyze data and recommend tailored treatment combinations for the user. Additionally, as long as USDA continues its experiments, the model will be up to date on the most recent data, and therefore generate improved precision and accuracy of the model's results. The model will obtain these results by first querying a Neo4j database via Cypher queries, a query language similar to

SQL but made for Neo4j databases. This step will perform two tasks. First, collect results that match the given prompt, and second, distill the output into a simplified version for the user to read. Using Cypher queries, the model can create complex calculations to produce a table with the desired values. These results can then be saved as CSV files, which the AI model can read and group for further processing. Python libraries can help convert CSV files into graphs by using the pandas and NumPy libraries to allow python code to analyze the data, then use matplotlib to visualize the data in a meaningful way. The model will then efficiently summarize the data for the reader to interpret so they only have to understand the key points of the model's analysis. By automating these processes, the model saves time and reduces the complexity involved in manually analyzing large datasets.

Another aspect that could be implemented in the model is how it responds to queries. Since our target audience is soil scientists, we should train the model to use field-specific terminology rather than providing generalized responses. The most effective way to achieve this is training the model on academic papers in soil science, which define and apply vast amounts of terminology. This approach will enable our target audience to gain a deeper understanding of the proposed solutions and allow them to adjust the outcomes based on their knowledge and expertise.

Evaluation

Like any AI model, the model can create hallucinations, so two tasks must be validated. First, the clarity of the model's responses must make logical sense while being concise. Because the model is designed to mimic the communication style of soil scientists, its responses must be reviewed by domain experts. This will take time to create a well-defined system of how the model should respond but in the long run, this will be highly beneficial. Scientists will also need training to formulate effective queries and minimize the risk of query failures. The second task we need to evaluate is the accuracy of the model's responses. We cannot simply ask the model to give us solutions to problems we ourselves do not know the answer to. The reason why we cannot do this is simply because this would take too long to verify and is precisely the scenario we are trying to avoid with this project. To efficiently evaluate the accuracy of the model, we first need to test it on questions that are already known. Questions like, are soil organic carbon and crop yield related? The answer should be yes as there has already been a proven correlation between the two (Oldfield, Bradford, and Wood 2019). Once the model has gone through its supervised training, then we can confidently attempt to ask for theoretical treatment strategies.

Discussion

If trained properly, the ability of an AI model to assist researchers in their experiments should drastically reduce the time and money needed to conduct their experiments. By integrating AI into this process, researchers can leverage the model's power to simulate and predict outcomes before physical experiments are conducted. I expect the efficiency of research in the agriculture field to increase as researchers will have the power of an AI model with decades of data to help their research. AI can identify patterns, correlations, and insights that would take researchers months or even years to uncover manually. Furthermore, the enhanced capabilities provided by the model will empower researchers to take bolder, more ambitious approaches in their work. With the significant reduction in time, cost, and effort required to conduct experiments, researchers will have greater freedom to explore more innovative and high-risk ideas that may have previously been deemed too resource-intensive or uncertain. This shift in focus away from logistical constraints will enable scientists to pursue bigger, more complex research questions, pushing the boundaries of agricultural science. For example, by accelerating discoveries in agrarian science, AI can help researchers develop more sustainable and efficient farming practices. This will become a more vital pursuit in the future as the world population is predicted to reach nearly 11 billion people by 2100 (Cilluffo and Ruiz 2019).

Conclusion

Agricultural research, like any field, carries the inherent risk of having a hypothesis proven wrong. However, this risk is amplified by limited funding and lower interest compared to other disciplines, where securing resources often involves stringent scrutiny. A failed hypothesis in agriculture can mean months or even years of lost work. An AI assistant can help mitigate this challenge by enabling researchers to test bold hypotheses with greater confidence, reducing the fear of failure, and accelerating research output and discoveries. This is made possible by leveraging the Knowledge Graph developed by the IDIR lab, which encompasses over 60 fields and 600,000 data instances. Using Neo4j and Cypher queries, the model can efficiently analyze this wealth of information and produce insightful results. To fully realize the potential of this approach, the model's performance will be evaluated in two key areas: the relevance and clarity of its responses to queries and the accuracy of the information provided. With input and oversight from soil scientists, the AI can be effectively refined to transform agricultural research, driving innovation and progress in this vital field.

Acknowledgments

This work would not have been possible without the mentorship of Dr. Chengkai Li and the support of the USDA team, including Dr. Catherine Stewart, Dr. Virginia Jin, Timothy Propst, and Dr. Jennifer Woodraw. Their invaluable contributions to the SOCKG project have been instrumental, and we at UTA could not have achieved this milestone without their guidance. I would also like to acknowledge my colleagues who have worked alongside me on this project, as their efforts have been essential in building such a versatile and promising Knowledge Graph.

References

Cilluffo, A.; and Ruiz, N. G. 2019. World's population is projected to nearly stop growing by the end of the century.

Heisey, P.; and Fuglie, K. 2018. Agricultural Research in High-Income Countries Faces New Challenges as Public Funding Stalls. *Amber Waves Magizine*.

Oldfield, E. E.; Bradford, M. A.; and Wood, S. A. 2019. Global meta-analysis of the relationship between soil organic matter and crop yields. *European Geosciences Union*, 5(1): 15–32.