Auburn/USDA CS Liaison Meeting 9/26/2024

Agenda

1. Statement of Work (~30mins maximum)
   1. The strange grammar/layout glitches come from PDF-docx conversion
   2. Dr. Backus’ stated feedback and discussion points
      1. Discussion of our statement that the project may not succeed.
         1. Based on the papers Dr. Backus has provided, we believe that our tool will perform similarly to the models described in the papers. We will not know what level of performance we will achieve until we have spent some time approaching the problem. There’s unfortunately no way for us to examine the dataset and estimate the difficulty or an upper performance limit, and committing in our Statement of Work to meet even a generous performance metric will put us in an uncomfortable position if it turns out not to be feasible.
         2. Please be assured that we have moved comfortably into solution exploration and are now handling the target data. This team is confident, capable, and has your various priorities in mind.
         3. Time for input from all liaisons on this point–what can we do for you with the above information in mind?
      2. CS jargon should be explained better for non-CS readers
   3. Comments from Dr. Reif?
   4. Comments from Dr. Cooper?
2. Our progress so far
   1. We are working on solutions using the mosquito data. It’s very exploratory but we’ve begun with very initial training attempts.
   2. We (especially Zach) are continually finding ways to improve cleaning and manipulation of the data for our purposes.
   3. We have found the phase transitions graph very helpful!
   4. Time for any specific progress any team members want to describe.
3. Question from Milo on transition matrix & specific transition regularity
4. Time for further questions from the team
5. Time for further questions/concerns/discussion from liaisons

Minutes

* Reviewing statement of work
  + Dr. Reif:
    - Generally okay with statement at the end regarding potential failure
    - Need more description of datasets available, reasons and pros/cons for working with each dataset
    - Justify working with aphid data even though funding is for mosquitos
    - Would be good to create document of computer science terms to include in final report
      * Append current version to each week’s agenda
    - Ticks are not insects, arthropod is a better catch-all term (although aphids/mosquitos are insects)
* Progress from the team this week
  + Milo:
    - Small issues with mosquito data, fixed now
    - Currently measuring performance based on 3 second sliding window, getting about 86% accurate, will probably decrease when issues fixed
    - Type N may be difficult to predict due to lower frequency in the data set
  + Mehrezat:
    - Milo has starting training a model on the mosquito data
    - Zach found that there are not as many errors in the aphid data as we thought, will probably be useable
    - Phase transitions graph from mosquito paper has been helpful
  + Milo:
    - Created graph of actual phase transitions in the data, compare to graph from paper
    - Dr. Cooper
      * L to NP transition is probably error in the dataset
      * Graph from paper contains both probes to repletion and incomplete probes (all probes from all files)
  + Zach:
    - We’ve split into 2 groups, Milo using convolutional neural networks, another group working on methods from existing papers
  + Dr. Backus: how are we handling files where wire starts coming off towards the end of the recording?
  + Zach:
    - It’s hard to know at this point whether we should discard them or if they will be useful — needs more testing
    - Annotation errors are a very small part of the aphid dataset
      * For example, P to E1 transition instead of P to D
* Timeline for statement of work revisions
  + Mehrezat:
    - plan on revised sent out by Monday
    - final version due Tuesday midnight
  + Dr. Backus has a deadline at the same time and will leave final revisions to Dr. Reif and Dr. Cooper
  + Dr. Reif: we should think about pros/cons of different datasets based biology of separate species
    - Aphids longer recordings, fewer individual insects
    - Mosquitos shorter recordings, many more individual insects
    - Consider what challenges each data set will provide to our program
  + Dr. Cooper: Are we splitting data into training/testing sets? Is this typical for machine learning?
    - Mehrezat: need to test on fresh data that the model has never seen before
    - Milo: We are going to use 80% of the files for training, and then 20% for testing
    - Dr. Cooper: ideally use a random sample for the 20%
  + Dr. Backus: What are our impressions of the papers on computer classification? Has science advanced since then, do we agree with the methods?
    - Zach: first paper looked at frequency components of 10 second chunks and used decision trees
      * Hard to say why this didn’t generalize well
      * Paper from the Chinese scientists generally similar, looked at some different shape-like features of 10 second chunks, also used decision trees
      * Would need different waveform types to have different frequencies for this approach to work well
  + Dr. Reif: if anyone is interested in continuing work on this over the summer there’s a possibility of funding summer work