Meeting's report for ELLEN WANG

Honour's student

08-10-2021

1. Agenda for today's meeting

- Present some more findings on readings I've completed regarding my updated understanding of the Random Forest (5 min)
- Discuss any problems with refined understanding
- Advice on writing random forest algorithm: run time, OOP or just functions?

2. Work already completed

- Looked through lots of pieces of Code on how others have been writing their random forest algorithms
- Met up with Jiyang to get a better understanding of the data
- Wrote and combined some code to run on Titanic Dataset

3. Meeting Minutes from previous week

- Literature deadline is the same as thesis.
- More important thing is to keep track of all your readings.
- Use the Jabref software to track interesting or not interesting articles, add what you think is relevant, proofs that might be reused later, application on datasets that are similar.
- Using Jabref: download the right bibtek entry from an official search engine: google scholar.
- Sometimes bibtek entry is in the article, download and export as bibtek, import into Jabref.
- Another good thing to keep track of: key words.
- Keep track of articles that reference the ones you are interested in.
- Inputs: author, name, year of publication are the columns in bibtek
- Honours tip: be as efficient as you can in every task you are doing
- Meetings tip: start each meeting with 2 minute recap of previous meeting + potential record
- Although your samples are limited, based on your POC, UNSW might preprocess more samples depending on your findings.
- Read Chapter 8 and 11 of the textbook Audrey recommended.
- 3D voxel predictions, special correlation.
- Being able to design or contain the 3D spatial idea might be a good plan, but it's better to clarify the research question with Jiyang. What do they want to interpret?
- Plan is now to create a new random forest algorithm specifically adapted to the task. Try existing tasks and then try and get a better result.
- RF understanding feedback: decisions trees had big drawbacks, bagging and other concepts resulted in the random forest. What is the history of the random forest?
- Start with the classical random forest to understand the data: what are the problems?
- Identify weaknesses of the classical random forest in order the understand the image data and adapt your new method to MRI data.

- Read a bit more about decision trees, gini impurity and entropy. Why is a single tree not enough and why was the random forest developed? Bagging was a good idea but not enough, therefore the random forest was developed.
- What parameters are involved in bias and variance control?
- There are a lot of parameters in the random forest, some are default. Which parameters are deemed more important than others?
- The random forest algorithm is famous because it is very simple to tune.
- We want large trees which will give us low bias and a high variance due to bagging. This makes the random forests not prone to overfitting.
- Process: decision trees to bagging to random forests. Create a crash course on this process.
- Task: program the basic version of the random forest. Design small pieces of code that are specifically adapted to this task.
- Details: Consider binary classification and only continuous predictors (Python or R)
- Continued: Apply your newly created decision tree on a simulated dataset, apply scikit learn and see if you get similar or better results. Simulated dataset of a fake brain, 3D cube with voxels, put values in it and classify them.
- Keep in mind: that your code will be quite slow compared to scikit learn.
- Improve the algorithm by considering and modifying the splitting criteria.
- Audrey's assumption: may have to do with the random forest resampling strategy, not just sampling voxels but also the neighbourhood around it and see if it makes a difference.
- Idea: mimic what CNN does well and incorporate that into the random forest.
- Question: what part of the original decision tree algorithm did they modify?
- For the articles you found: did they resample and change up the random forest? If so, what did they change and why? What's the justification and difference?
- Note that you are doing the exact same thing, so it will be good to understand their modifications.
- Coding: no problem coding as long as you understand it. Add comments. Start from scratch and read the original decision tree article that describes these concepts.
- Note that the naïve way is very long. Keep your first code very simple so that further versions improve each version. Don't try to put everything in the first version.
- Don't worry about speed of algorithm, when we apply real data we will work on speed.
- Do what you prefer for code: Python or R
- Keep working on the organisation of the meetings and honours project.
- Don't read the maths for now, the thesis is not a theoretical project and note that your supervisors can support you in understanding the math.
- It's better to design an algorithm that does the job.
- Audrey: send some equations that are relevant for bagging.
- Interpreting research articles: what's import to focus on and learn. Understand main ideas.

4. Planned work after this meeting

- BUild one tree for the random forest
- Revise understanding of random forest concepts (focus on metrics)
- Review Audrey's comments on RF presentation.
- Review Audrey's lecture notes.