Constructive AI

Evaluating CAI Systems

Matthew Lewis

m.lewis4@herts.ac.uk

Overview

In this lecture, we will:

- Introduce qualitative & quantitative evaluation in the context of CAI systems.
- Give examples of questions to ask in qualitative evaluation.
- Give examples of metrics to use for a multi-resource problem.
- Go through some worked examples of quantitative evaluation for multi-resource problems.
- This will be relevant to part (b) of coursework CW2

Evaluation in CAI

- In a robotic system we can evaluate the components individually, e.g. object detection, distance sensing.
- However, in embodied cognition, we are considering the *system as a whole* interacting with its environment. Therefore, rather than evaluating components, we want to evaluate the whole system.
- It might be that distance detection is not accurate, but it is good enough for the robot's purposes (and improving it wouldn't improve the robot's performance).
- In embodied cognition, it might be the case that there is no explicit distance sensing (e.g. Braitenberg Vehicles).

Why Evaluate?

- Useful to ask “why are you evaluating?” – helps to choose appropriate methods
- E.g. Evaluation of robots (or robot controllers) when investigating synthetic approach to understanding life (understanding by making). Answering the question: Is this a good solution to problems of living?
- In evolutionary algorithms, evaluation/comparison can be used to choose which robots to “breed” to produce the next generation.

Qualitative vs. Quantitative Evaluation

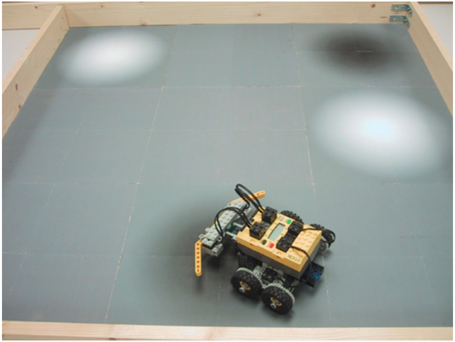
- Quantitative: Evaluation in terms of numerical "quantities" (metrics)
- Qualitative: Not quantitative! Examples on next slide.
- Quantitative analysis (if done rigorously) can give reliable results – often regarded as "gold standard"
- However, be careful not to be misled by numbers – think about what the numbers mean
- Qualitative evaluation can help to understand quantitative metrics
- Qualitative evaluation can highlight areas that can be investigated more thoroughly by quantitative methods.

Qualitative Evaluation Examples

- Do you observe any emergent behaviours?
- How does a robot move about its environment?
  - Does it explore all areas of the environment?
  - Does it hit obstacles? Which ones?
  - Are there any places where it gets stuck?
  - Does it exhibit repetitive patterns of behaviour?
- If line following is a task…
  - Does it find a line when it crosses it?
  - Does it keep following a thin line?
  - Does it keep following a line that turns sharply?
- For a “predator” robot…
  - What is its hunting “strategy”?

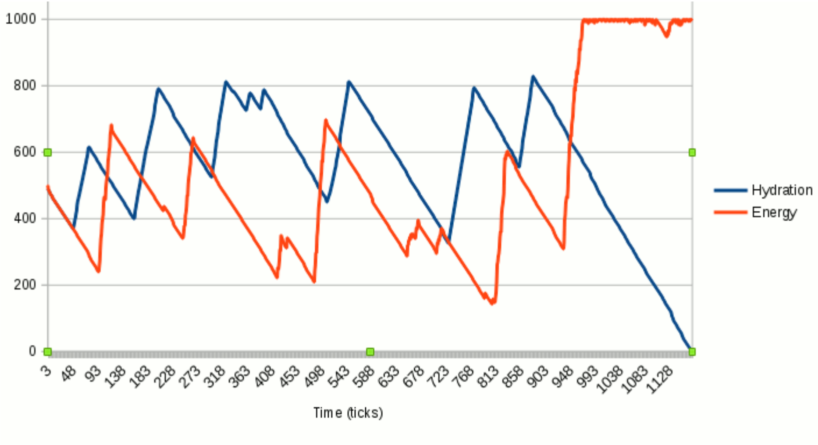
Quantitative Evaluation: Robot Survival

- We will look at an example using multi-resource problems (2 resource and 3 resource problems)
- A robot is designed to survive in an environment, managing its 2 (or 3) internal variables using "resources" found in the environment.



Quantitative Evaluation: Robot Survival

Collect & plot data: values of the physiological variables



Metric: Survival Time

- Definition: The time (in some units, e.g. seconds, time ticks) from the start of the experiment until the robot “dies”
- Easy to understand, to calculate and directly relevant.
- Disadvantage: The time of each run may be limited, and if the robot survives then the survival time is the maximum time of the run.
  - Not useful for comparing robots in an unchallenging environment

Metric: “Wellbeing” (or “comfort”) at time *t*

- Indication of state of the robot, based on the essential variables
- Calculated as the *mean* of the distance of the distance of each essential variable (at time *t*) from its fatal limit.
- High values indicate the robot is “doing well”.
- To get the overall wellbeing, take the mean over the entire run.
- We may need to scale values (e.g. if one variable runs from 0–100, but another runs from 0–10).

Wellbeing: Example

Three essential variables (all with arbitrary units)

	min	max	fatal limit(s)	ideal value
energy	0	100	0	100
damage	0	100	100	0
temperature	-100	$+100$	± 100	0

We run the robot in the environment and collect data:

time (ticks)	0	1	2	3	4	...	3000
energy	50.0	49.5	49.0	48.5	48.0	...	63.8
damage	37.3	37.2	37.1	37.0	36.9	...	21.5
temperature	17.4	17.3	17.4	17.4	17.3	...	-29.1
energy$-$0	50.0	49.5	49.0	48.5	48.0	...	63.8
100$-$damage	62.7	62.8	62.9	63.0	63.1	...	78.5
100$-$|temperature|	82.6	82.7	82.6	82.6	82.7	...	70.9
wellbeing	65.1	65.0	64.9	64.7	64.6	...	71.1

Wellbeing: Disadvantages

- Remember to scale values if required.
- What happens if the robot dies?
  - We could take mean wellbeing during life of robot
  - or we could find a way to calculate wellbeing for a dead robot
- May not be indicative of “closeness to death”.
  - Example: two robots – which one is doing better?

	Robot 1	Robot 2
energy	30.0	79.5
damage	60.0	99.5
wellbeing	35.0	40.0

  - We can have a dead robot with wellbeing 50
  - Problem gets worse with more essential variables
- May depend on the starting value (particularly if the run-time is short)

Metric: Physiological Balance

- Indication of how “balanced” the robot has managed its physiological needs, based on the essential variables.
- Calculated as the *variance* of the distance of the distance of each essential variable (at time t) from its fatal limit.
- Low values indicate the robot currently has balanced needs.
- High values indicate the robot has some needs significantly larger than others.
- As with wellbeing, we typically take a mean over the run, to give an overall indication of the balance of the robot.
- Helps to compare different strategies, but not necessarily saying one is better than another.

Example: Comparing robots quantitatively

- We run an experiment, with a robot trying to survive in an environment trying to satisfy its two physiological needs.
- We compare two different controllers for the robot.
- Hypotheses: The two controllers differ as measured by the robot's survival time, overall wellbeing, overall wellbeing (extended), physiological balance.
  - Null hypotheses: The two controllers are the same as measured by the robot's survival time, overall wellbeing, . . .
- We run each robot ten times, up to a maximum of 10 minutes for each run.
- We record the values of the two physiological each time-step (0.1s).

Example: Comparing robots quantitatively

From the recorded data, we calculate the survival time, overall wellbeing, and physiological balance:

Run	S.T.(s)	Ov.Wb.L	Ov.Wb.E	P.B.
1	275	40.9	18.8	164.1
2	355	28.9	17.1	40.2
3	491	27.7	22.7	44.8
4	411	30.6	21	18.8
5	319	38.0	20.2	48.0
6	600	36.1	36.1	16.0
7	563	34.1	32.0	58.3
8	600	35.2	35.2	25.5
9	319	26.4	14.0	13.6
10	600	33.2	33.2	20.8
mean:	453.3	33.1	25.0	45.0

Run	S.T.(s)	Ov.Wb.L	Ov.Wb.E	P.B.
1	600	49.1	49.1	36.3
2	567	34.6	32.7	75.6
3	600	36.9	36.9	50.9
4	507	41.5	35.1	246.2
5	499	38.1	31.7	298.1
6	600	37.5	37.5	9.9
7	575	24.7	23.7	25.4
8	600	35.5	35.5	22.7
9	595	31.0	30.8	53.6
10	600	44.9	44.9	49.5
mean:	574.3	37.4	35.8	86.8

We calculate two different values for the overall wellbeing:

- The lifetime overall wellbeing (Ov.Wb.L) taking the mean over the time when the robot is alive
- The extended overall wellbeing (Ov.Wb.E) taking the mean over the full 10 minutes (when we define the wellbeing to be zero)

Example: Comparing robots quantitatively

- We can compare the metrics using a *t*-test.
  - This can be done, for example, using Excel (T.TEST() function), R (statistics language), SPSS (statistics package), or Python (with stats.ttest_ind() from SciPy).
  - For more than two robots, use ANOVA.
- We use a two-tailed test, not assuming equal variances. This gives the following p-values:

	S.T.(s)	Ov.Wb.L	Ov.Wb.E	P.B.
p-value	0.0187	0.1232	0.0061	0.2504

Example: Comparing robots quantitatively

	S.T.(s)	Ov.Wb.L	Ov.Wb.E	P.B.
p-value	0.0187	0.1232	0.0061	0.2504
	< 0.05		< 0.01	

- Hence, we accept the hypothesis that the controllers differ under the survival time metric (significance level < 0.05), and under our extended mean wellbeing metric (significance level < 0.01).
- If we used only the mean wellbeing over the life of the robot, then we would retain (keep) the null hypothesis (since $p > 0.05$).
- We also retain the null hypothesis that the controllers are the same in terms of their physiological balance. Intuitively: the two controllers maintain the same balance between satisfying the two variables.