Characterization of Session Traces

Discussion led by Steve Marron

4/14/00, 4/28/00 & 5/5/00

Goals: Analyze "population of traces":

1. Summarize by low dim'al (10-20 d?) "feature vectors"
2. Analyze resulting population (e.g. clustering)
3. Use as basis to study change

3 Approaches to Summarization

1. Simple Summaries
2. Markov Chain methods
3. Conservative Cascades

1. Simple Summaries

a. “Size” Summaries

Show SessionData3p21d1s.ps

i. Total Time

ii. Sign On Time

iii. Sign Off Time

iv. Total Packet Size

v. Total Number of Packets

vi. Proportion of Big Packets

Then “remove sizes” to get [0,1] scales

Show bottom of SessionData3p21d1s.ps and show SessionData3p21d2s.ps

1. Simple Summaries

b. “Shape” Summaries

Show CombineSessionData1p31.pd

vii. x – pt. “farthest from diagonal”

viii. y – pt. “farthest from diagonal”

Note: a crude “shape measure”

ix. RSS for linear fit

Note: “very large” for OP5, P2
“very small” for P3, P4, P5

1b. Shape Summaries (cont.)

x. slope of left linear fit

Note: small for “slow start”?

xi. slope of right linear fit

Need both this and above?
Could replace by ratio.

xii. Proportion of RSS explained by break in line.

Large when close to “broken line”, P3, 4, 5, OP1, 2, 3, 5
Small when “more structure”

Simple Summary Questions

1. Transformations?

    - Logs

    - Differences and ratios

2. Overparametrized?

    - (x,y) farthest contains most of 1st slope and 2nd slope info?

    - Really need both of 1st slope and 2nd slope?

Simple Summary Questions (cont.)

3. What's missing? Measures of:

   - TCP "steppiness" (how important is this?)

Look at Peak 1 and Peak 2

   - Size of "big flat bits".

Look at Off Peak 1, 2, 5

   - Number of "big flat bits"

   - ?

Some Approaches

- interarrival time distributions

Show right columns in

- interarrival time series

Show middle left pic in CombineSessionData1p22.pd

- log interarrival time series

Show lower left pic in CombineSessionData1p22.pd

- clustering of log interarrival times

Show 1st pages of CombineSessionData1p25.pd

- generation from summaries

Show 2nd pages of CombineSessionData1p25.pd

- deeper clustering?

Problems

- Want too many things?
- No longer “simple”?

2. 2nd Generation Simple Summaries

a. Size Summaries:

Show SessionData3p41d1s.ps

i. log10(Total Time)

ii. log10(Sign On Time)

iii. log10(Sign Off Time)

iv. log10(Total Packet Size)

v. log10(Total # of Packets)

vi. log10(# of Big Packets)

Note: logs allow representing “ratios” and “proportions” as linear functions, e.g. % sign on time, Avg. Packet size

2. 2^{nd} Generation Simple Summaries

b. Shape Summaries

- after removing "size", note no sign on – off, and [0,1] axes

i. Successive Broken Line Fits

Show CombineSessionData1p42.pd

- choose next knot by largest residual
- Projection $\Rightarrow$ RSS $\downarrow$
- Doesn't "track corners"
- Knots can "bunch up"

2. 2^{nd} Generation Simple Summaries

ii. Broken Line Interpolations

Show CombineSessionData1p44.pd

- choose knots from max resid’s
- tracks “corners” and “flat spots”
- good “visual summary”
- RSS can increase
- Number of knots?

2. 2^{nd} Generation Simple Summaries

iii. Broken Line Summaries

Show CombineSessionData1p45.pd

For each linear piece:

a. log10(slope)

- ratios “contain locations”

b. log10(RSS)

- log scale is “natural”

c. bias

- log version which keeps signs?

2. 2^{nd} Generation Simple Summaries

iv. Number of knots?

4 ???

- Captures “most”
- Allows 2 flat pieces

6 ???

v. Resulting parameters?

4 knots $\Rightarrow$ $6+3*(4+1) = 21$ total

6 knots $\Rightarrow$ $6+3*(6+1) = 27$ total

3. 3rd Generation Simple Summaries

a. Size Summaries, same as 2nd Gen. (above)

3. 3rd Generation Simple Summaries

b. Shape Summaries:

Show CombineSessionData1p51.pd

For each linear piece:

i. log10(width)

- log10(time between knots)
- allows reconstruction

ii. log10(slope)

- log scale naturally includes ratios
- also allows reconstruction

3. 3rd Generation Simple Summaries

c. Shape Summaries (cont.):

iii. Area, absolute Residuals

- Records "how far off"
  (trace is from linear fit)

- On "visual scale"

iv. Area, Residuals

- shows "direction" of deviation

- reflects "bias" component of error

- "difference" not "ratio", so no log