Nature-Inspired Computing

Working with Nature-Inspired Heuristics

Dr. Şima Uyar
September 2006

Issues Considered

- experimental design
  - algorithm design
  - experiment design
- test problems
- performance criteria
- appropriate statistics

2

Experimentation

- define a set of goals / objectives
  - formulate a question or hypothesis
  - design the experiments (algorithm runs may be considered as experiments)
- collect necessary data
- analyze data
- design further experiments

Goals for Experimentation

- obtain a good solution for a given problem
- show that a specific approach is applicable in a problem domain
- show that a proposed algorithm improves a benchmark case
- show that an algorithm outperforms traditional algorithms

Goals for Experimentation

- find the best parameter setup for an algorithm
- explain an algorithm behavior
- show if an algorithm scales-up with problem size
- experiment with effect of parameter settings on performance

Different Goals

- find a very good solution at least once - design
- find a good solution at almost every run - production
- must meet scientific standards for publication

Test Problems

- benchmark problems
- real-world problems
- randomly generated problems
- choice of test problem has severe implications on
  - generalizability
  - scope of the results
  - conclusions usually depend even on the chosen problem instances

Test Problems

- using real-world data
- advantages:
  - results very relevant from the application point of view
- disadvantages
  - can be over-complicated
  - can be few available sets of real data
  - may be commercial sensitive – difficult to publish and to allow others to compare
  - results are hard to generalize

Test Problems

- use standard data sets in problem repositories, e.g.:
  - OR-Library
    http://www.ms.ic.ac.uk/info.html
  - UCI Machine Learning Repository
    www.ics.uci.edu/~mlearn/MLRepository.html
- advantage:
  - well-chosen problems and instances
  - much other work on these → results comparable
- disadvantage:
  - not real – might miss crucial aspect
  - algorithms get tuned for popular test suites

Test Problems

- random problem instance generators, e.g.:
  - GA/EA Repository of Test Problem Generators

  http://www.cs.uwyo.edu/~wspears/generators.html
- advantage:
  - allow very systematic comparisons because they:
  - can produce many instances with the same characteristics
  - enable gradual increase / decrease of hardness)
  - can be shared allowing comparisons with other researchers
- disadvantage
  - not real – might miss crucial aspects of problem
  - a generator might have hidden bias

Analysis of Results

- NIHs are stochastic, i.e.,
  - do not draw conclusions based on a single run
  - perform sufficient number of independent runs
  - use statistical measures (averages, standard deviations, ...)
  - use statistical tests

Analysis of Results

- for comparisons:
  - always do a fair competition
  - use the same amount of resources for the competitors
  - use the same performance measures

What to Measure

- average result in given time
- average time for given result
- proportion of runs within % of target
- best result over *n* runs
- amount of computing required to reach target in given time with % confidence
- …

What Time Units?

- elapsed time?
  - depends on computer, network, etc…
- CPU time?
  - depends on skill of programmer, implementation, etc…
- generations / iterations?
  - difficult to compare when parameters like population size change
- evaluations?
  - evaluation time could depend on algorithm, e.g. direct vs. indirect representation

Measures

- performance measures (offline)
  - efficiency (speed)
    - CPU time
    - no. of steps, i.e., generated points in the search space
  - effectivity (alg. quality)
    - success rate
    - solution quality at termination

Measures

- "working" measures (online)
  - population distribution (genotypic)
  - fitness distribution
  - improvements per time unit or per genetic operator
  - …

Performance Measures

- no. of generated points, i.e. no. of fitness evaluations
- AES: average no. of evaluations to solution
- SR: success rate = % of runs finding a solution (individual with acceptable quality / fitness)
- MBF: mean best fitness at termination, i.e., best per run, mean over a set of runs
- SR $\neq$ MBF
  - low SR, high MBF: good approximizer (more time helps?
  - high SR, low MBF

Fair Experimentation

- allow all algorithms the same amount of running time
- allow each NIH to compare, the same no. of evaluations, but
  - look out for hidden labour, e.g. in heuristic mutation operators
  - lookout for the possibility of fewer evaluations by smart operators

No Free Lunch Theorem - NFL

There does not exist any algorithm which is better than another over all possible instances of optimization problems.

Analysis of Algorithms

- worst-case analysis
- average-case analysis
- experimental analysis

Experimental Research

- experimental design
- experimental analysis

Most Common Errors

- reporting result of 1 run is sufficient
- reporting best result of several runs is sufficient
- using plots is sufficient; no statistics needed
  - obvious from plots !
- reporting averages of several runs is sufficient

Why Need Statistics?

- need to draw strongest possible conclusions from limited data
  - two problems:
    - important differences may be obscured by experimental imprecision
      - hard to distinguish between real differences and random variation
    - tendency to over-generalize from limited data

Why Need Statistics?

- statistics allow general conclusions
  - extrapolating from SAMPLE to POPULATION
    - sample: data collected from experiments
    - population: data from all possible experiments

What can Statistics Do?

- statistical estimation
  - estimate population mean from sample mean
- statistical hypothesis testing
  - decide whether observed difference is likely to be caused by chance
- statistical modeling
  - test how well experimental data fit a model
    - e.g. linear regression

Why are Averages not Sufficient?

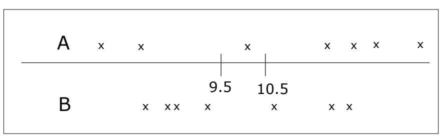
- assume two methods: A and B
- want to show method A better than method B
  - is the difference between means greater than 0?

Why are Averages not Sufficient?

Assume 7 samples:
Average of A: 10.5
Average of B: 9.5

Average of A **>** Average of B

BUT



Why are Averages not Sufficient?

- interested in distribution of mean of n samples
  - as n gets bigger, distribution approaches true mean
- want to be able to say:

In x% of all possible experiments, the true mean of this distribution will lie within a specified interval.

Confidence Intervals

- confidence interval of a proportion
- confidence interval of a mean

Confidence Interval of a Mean

- CI is a range of values
  - e.g. 95%CI: can be 95% sure that CI includes true population mean
    - no uncertainty about sample mean!!!!
  - commonly shown as
    - 20.0 to 32.0
    - [20.0, 32.0]

Confidence Interval of a Mean

- assumption: population distributed according to Gaussian distribution
  - not too important if large samples used
  - central limit theorem
- to calculate CI:
  - sample mean
  - sample SD
  - sample size
  - how much confidence?
    - typically 95% (sometimes 99%: wider interval)

Central Limit Theorem

- CLT: sum of independent, identically distributed (IID) random variables approaches a Gaussian
- CLT: regardless of the distribution of values in population, for large sample sizes, the distribution of means from independently chosen samples will approximate a Gaussian distribution
- how large?
  - depends on definition of “approximately” and the distribution of population
    - even if weird distribution, 100 samples enough
    - if approximately symmetrical and unimodal, 10 – 20 samples enough

Comparing Groups with Confidence Intervals

- CI of a difference between means
  - can be x% sure that value of true difference between populations lies within CI
    (implicit t-test)

Comparing Groups with Confidence Intervals

- Are CI sufficient for a comparison?

Assume two approaches A and B:

Case 1:

95%CI for A: 125 to 750

95%CI for B: 900 to 1800

Case 2:

95%CI for A: 125 to 1300

95%CI for B: 900 to 1800

p-Values

- for comparing two groups
  - CI of difference between means
    - question: How large is the difference in the overall population?
  - using p-values
    - question: How sure are we that there is a difference between the populations?
      - observed difference may be due to coincidence or random sampling
    - tells you how rare such a coincidence is

p-Values

- p-value : if both are from the same distribution, probability that difference the between the means of randomly selected samples will be larger than or equal to observed
- null hypothesis (H_0): distributions in the two populations are the same
- t-test
  - t-test assumes Gaussian distribution and equal SDs

p-Values

Example:

- assume p value is 0.034
  - 3.4% of all experiments will result in a difference ≥ observed
- two possible interpretations:
  - they have different means
  - they have identical means and observed difference is a coincidence
- can't say if H_0 is correct or not !

Statistical Significance & Hypothesis Testing

Hypothesis Testing:

1. assume samples randomly selected from populations
2. state null hypothesis: distribution of values in two populations same
3. define threshold for declaring p value significant (*significance level of test* : α)
   - usually α chosen as 0.05
4. select test and calculate p
5. if $p < \alpha \rightarrow$ difference is *statistically significant* and *reject null hypothesis*

Significance

- if α=0.05: is p=0.04 more significant then p=0.004?
  - based on definition: no
  - sometimes "very significant" and "extremely significant" used
    - commonly:
      - p<0.05 : significant
      - p<0.01 : highly significant
      - p<0.001: extremely significant

Significance

- if α=0.05:
  - p = 0.049 shows a significant difference
  - p = 0.051 shows a not significant difference
  - look at p value itself
- if p is slightly greater than α
  - sometimes "marginally significant" or "almost significant" is used
  - or add a third category: significant, not significant and *inconclusive*

Significance

- if not significant
  - can not say null hypothesis is true !
  - means data not strong enough to reject null hypothesis

Non-Parametric Tests

- does not assume Gaussian distribution
- Mann-Whitney rank sum test
- Wilcoxon rank sum test
  - usually called Mann-Whitney-Wilcoxon test
  - works on ranks of values
  - rank data points regardless of group
    - if a tie occurs, give average of the ranks
  - add up ranks in each group
  - question: if distribution of ranks between two groups were random, what is the probability that the difference between the sums would be so large?
  - use p-value
- alternately use t-test on ranks

Parametric x Non-Parametric Tests

- use non-parametric if:
  - definitely sure that no Gaussian distribution
  - data has very large outliers
- if parametric tests used with non-Gaussian distributions
  - OK if large sample sizes
  - what is large?

Comparing Three or More Means

- ANOVA : one-way analysis of variance
  - analyzes variance among values
  - tests null hypothesis that all populations have identical means
  - calculates p-value
  - if null hypothesis were true, what is the probability that the means of randomly selected samples will vary as much as or more than what has occurred?
  - has same assumptions as t-test
  - can't say which is better

Multiple Comparison Post Tests

- many tests
  - compare control group mean to all others? – Dunnett's test
  - compare all – Bonferroni, Tukey, Student-Newman-Keuls
    - Bonferoni: easiest and most common but too large CIs (do not use if ≥ 5 groups)
- for non-parametric testing
  - Kruskal-Wallis test

MANOVA

- use MANOVA if comparing
  - multiple groups and
  - effects of multiple factors

References for Presentation

- Bartz-Beielstein T., Experimental Research in Evolutionary Computation, Springer, 2006.
- Eiben A., Smith J. E., Introduction to Evolutionary Computing, Springer, 2003.
- Eiben A., Jelasity M., "A Critical Note on Experimental Research Methodology in EC"
- Johnson D. S., "A Theoreticians Guide to the Experimental Analysis of Algorithms", DIMACS 2002.
- Motulsky H., Intuitive Biostatistics, Oxford University Press, 1995.
- Wineberg M., Christensen S., "Using Appropriate Statistics – Statistics for Artificial Intelligence", Tutorial Slides, GECCO 2004 Tutorial Program, 2004. **(http://www.scs.carleton.ca/~schriste/tamale/UsingAppropriateStatistics.pdf)**