Experiments on Metaheuristics: Methodological Overview and Open Issues

Marco Chiarandini, Luis Paquete, Mike Preuss, and Enda Ridge

Presentation: Chris Fawcett
EARG, October 15 2008

Outline

- Metaheuristics and measuring their performance
- Univariate Analysis
    - Characterisation and statistical testing
    - Component comparison and tuning
- Multivariate Analysis
    - Characterisation and statistical testing in two scenarios

Metaheuristics

- Creating a heuristic from a collection of other heuristics.
  - Construction heuristics
  - Local search neighbourhoods
  - Hill-climbing and memory techniques
  - etc.

Measuring Performance

- Often we don’t reach an optimal solution after a given time bound.
- Two helpful metrics:
  - Solution quality achievable given a time bound
  - Time required to find a solution with a given quality
- Both are (in general) random variables.

Measuring Performance

- The field of statistics offers:
  - A systematic framework for designing experiments.
  - A Mathematical foundation for inferring the probability of events from empirical data.

Univariate Model

- Experimenter is interested in *either* solution cost or run-time.
- In both cases, the variable not under consideration is fixed to something reasonable.

Characterisation

- Performance measure X of a metaheuristic on a single instance is described equivalently by
  - its probability distribution

$$p(x) = \Pr\left[X = x\right]$$

  - its cumulative (discrete) distribution function

$$F(x) = \Pr\left[X \leq x\right] = \sum_{x_i \leq x} p(x_i).$$

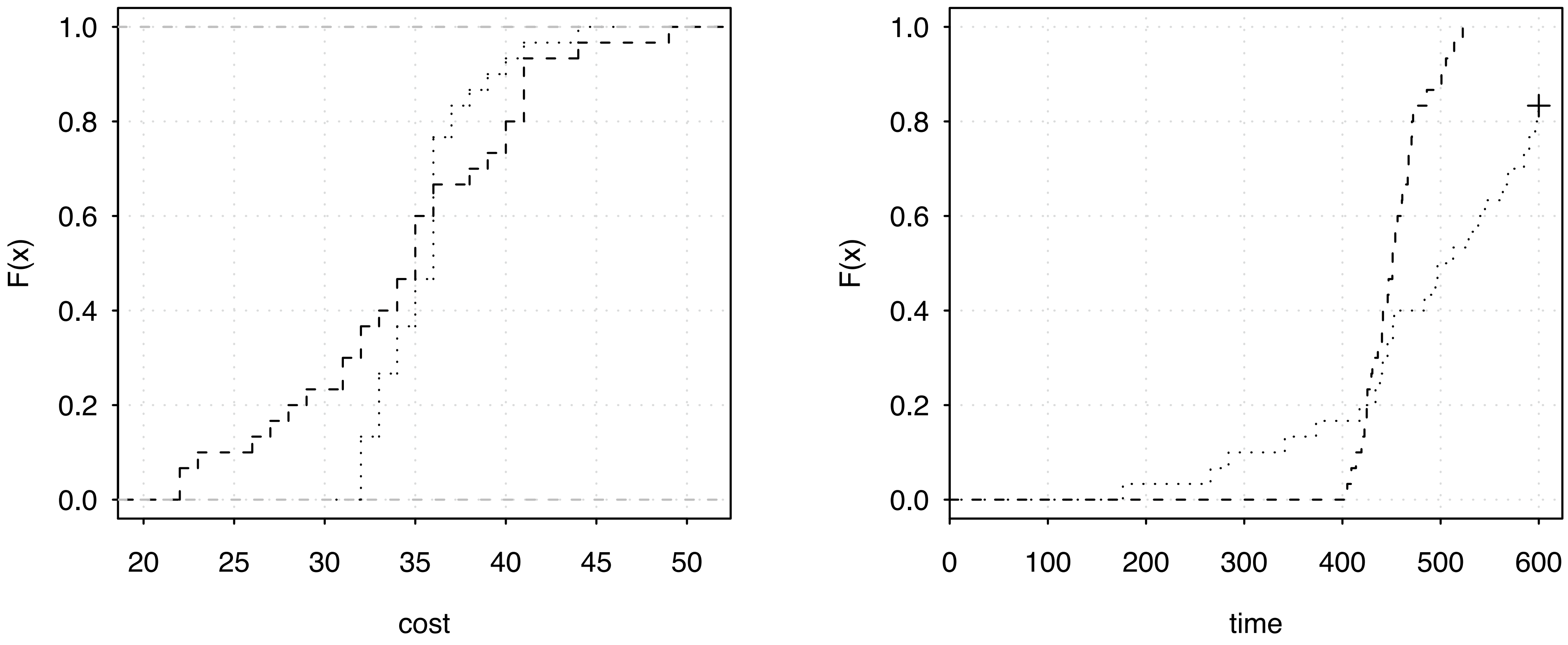
Characterisation

- Our experiments sample data X1, X2, ..., Xn from these distributions, giving an empirical cumulative distribution function (ECDF)

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

- Holds for censored and uncensored data.

Characterisation



Characterisation

- Usually we care about performance on a class of instances.
- We use a representative sample of this class, Π, yielding the modified probability distribution

$$p(x) = \sum_{\pi \in \Pi} p(x|\pi)p(\pi).$$

- More convenient if the samples have equal probability.

Characterisation

- Summary measures for this sample data are divided into
  - measures of location (sample mean, q-quantiles)
  - measures of dispersion (sample variance)
- Summary measures by definition remove some of the information in the samples.
- Should prefer the ECDFs themselves.

Characterisation

- For run-time, there are links to a branch of statistics called survival analysis, dealing with time-to-event models.
- ECDFs for run-time are often exponentially distributed.
- ECDFs for solution cost are often well-approximated by Weibull distributions.

Statistical Analysis

- Descriptive statistics are not sufficient.
- Inferential statistics must be used to check that the sampled data are enough to generalise the results.
- Statistical testing makes these statements objective.
- Both parametric and non-parametric tests exist.

Statistical Analysis

- Parametric tests often assume normally distributed data.
  - Authors claim that this isn’t an issue because some parametric tests are robust?
- Non-parametric tests remove this normality assumption

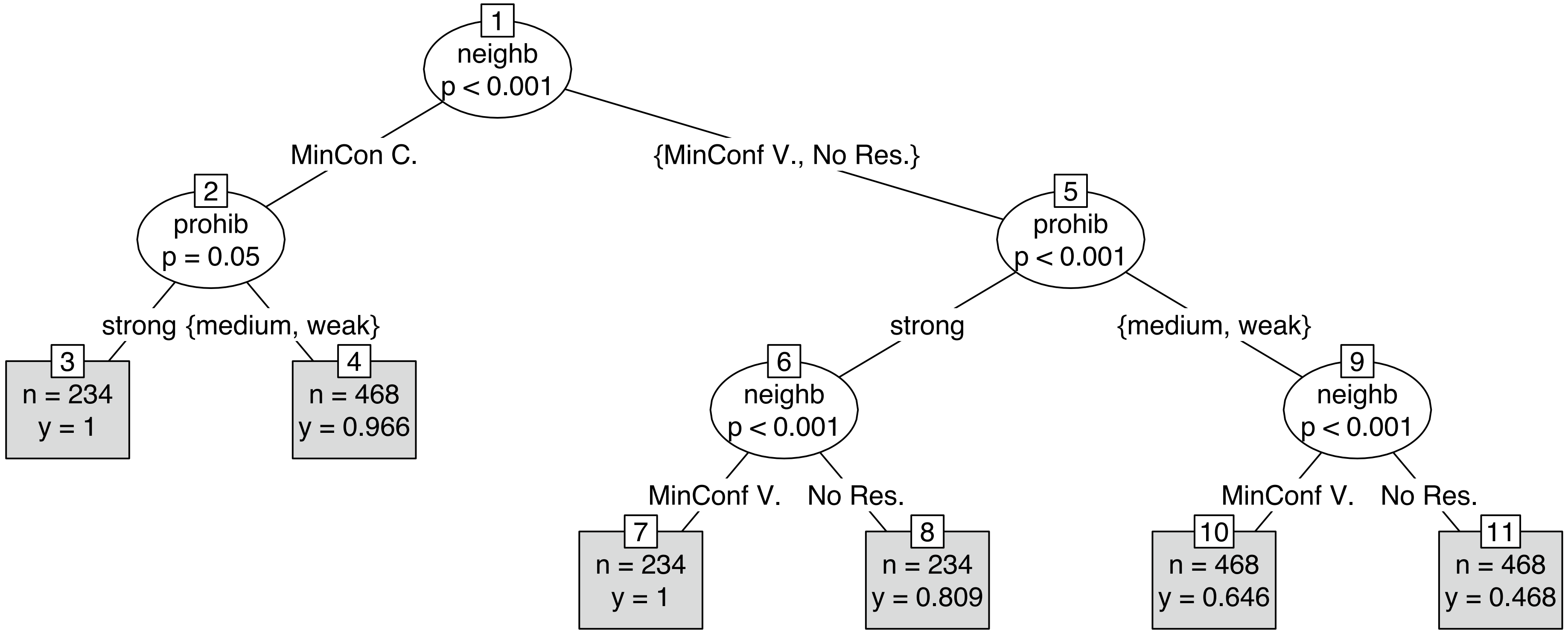
Statistical Analysis

- Two sample unreplicated tests
  - Matched pairs Welch t-test (parametric)
  - binomial test or Wilcoxon signed rank test (non-parametric)
- Replicated
  - Blocking on both instance and seed, or
  - two-way ANOVA or Kruskal-Wallis rank sum test

Statistical Analysis

- It can be more accurate to compare the ECDFs of two metaheuristics.
- Kolmogorov-Smirnov (KS) test uses the maximal difference between two ECDF curves. Can identify statistical dominance.
- Can be hard to identify a preferred metaheuristic when there is no statistical dominance.
- All of these tests assume uncensored data.

Regression Trees



Parameter Tuning

- How to determine which algorithm parameters and instance properties have effects on the response.
- What are the most important parameters?
- Factorial designs aren’t really appropriate
- Fractional Factorial Designs can be
- Authors mention desirability functions and overlay plots.

Sequential Testing

- How many runs do we need to make to identify differences between two parameter configurations?
- Racing algorithms (F-Race)
- Sequential Parameter Optimisation (SPO)
- Crossover between the two?

Multivariate Model

- A thorough understanding of metaheuristic performance should include *both* run-time and solution quality.
- Authors distinguish two scenarios of this type.

Scenario One

- We evaluate solution cost and run-time at the point where a certain termination criteria is reached.
- Each run of a metaheuristic is single data point (solution cost, run-time)

Characterisation

- If $\boldsymbol{X} \in \mathbb{R}^2$ is the bivariate performance measure, the cumulative distribution function is

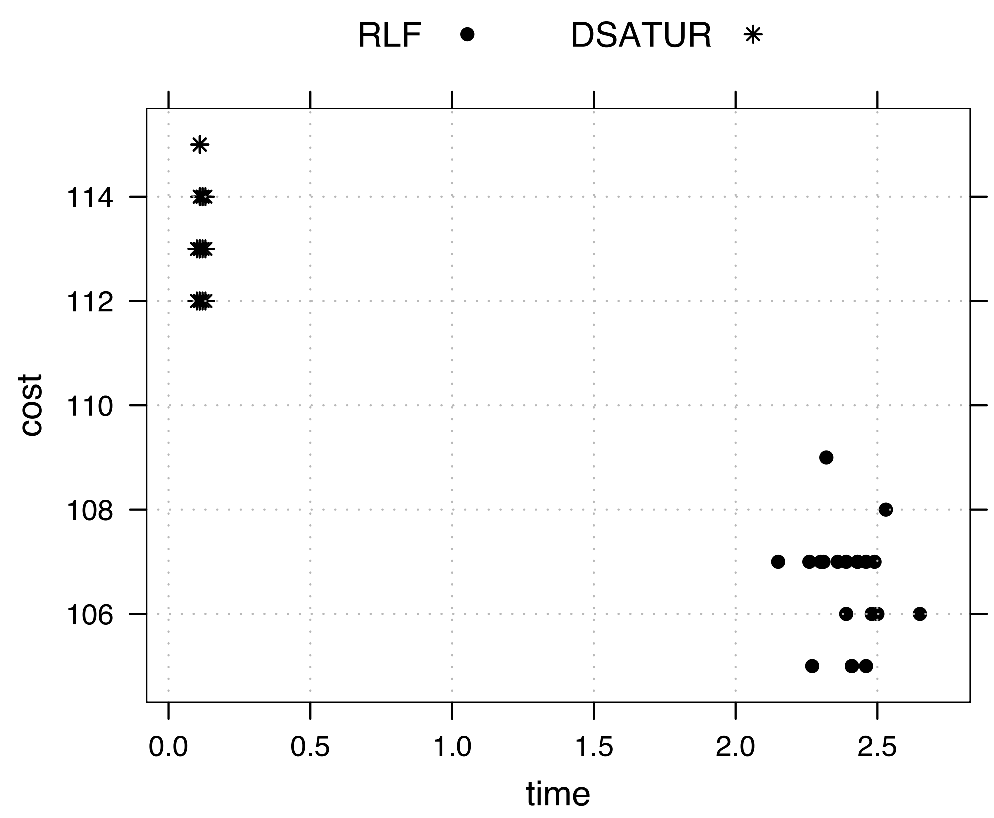
$$F(\boldsymbol{x}) = \Pr[\boldsymbol{X} \leq \boldsymbol{x}]$$

- and the corresponding ECDF:

$$F_n(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{X}_i \leq \boldsymbol{x}).$$

Characterisation

- Can compare the envelope or center of gravity of the two sets of points.



Statistical Analysis

- Can compare bivariate means using
  - Hotelling's T^2 test (parametric)
  - multivariate analysis of variance (MANOVA)
- Rank ordering doesn't extend to multiple dimensions, so non-parametric testing is unclear.
- Can also extend Kolmogorov-Smirnov and Birnbaum-Hall to the ECDFs.

Scenario Two

- The experimenter is interested in solution cost during the run of a metaheuristic.
- A single run is now a set of (solution cost, run-time) points.
- Analysis and characterisation from random-set theory.

Characterisation

- Given that we care about improvements over the course of a run, we can use the set of non-dominated points:

$$\mathcal{X} = \{\boldsymbol{X}_j \in \mathbb{R}^2, j = 1, \ldots, m\}$$

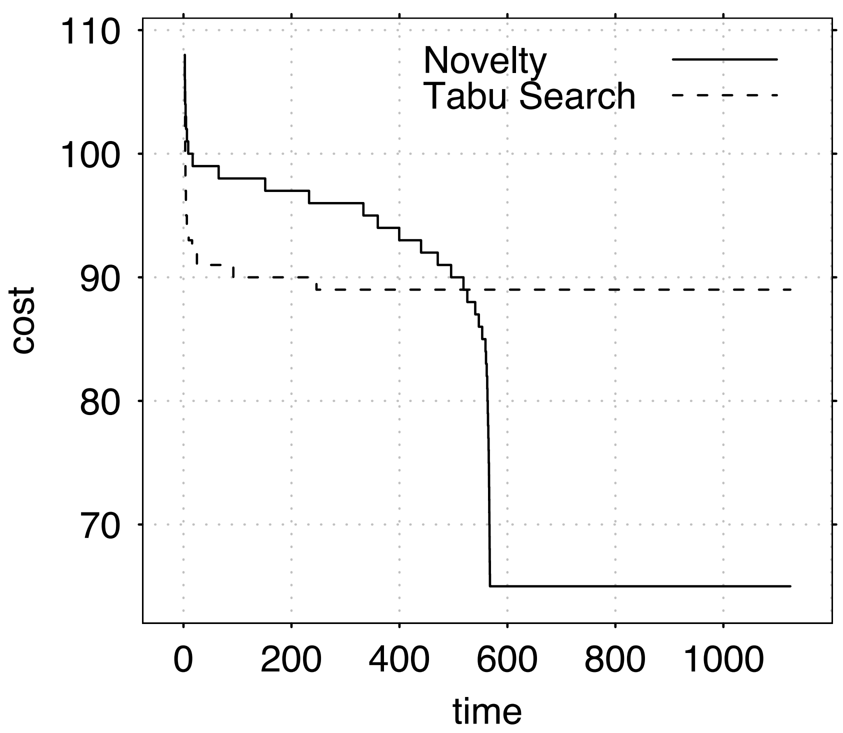
$$F(\boldsymbol{x}) = \Pr[\mathcal{X} \trianglelefteq \boldsymbol{x}]$$

- ECDF defined as usual:

$$F_n(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} I(\mathcal{X}_i \trianglelefteq \boldsymbol{x})$$

Characterisation

- We can slice this bivariate ECDF in any of the three axes to create interesting graphs.
  - The authors only mention probability quantiles.



Statistical Analysis

- Fairly rare to see statistical dominance of these ECDFs in practice.
- Perhaps finding the best performance for some specific run-times, etc.

Statistical Analysis

- Taillard has used a Mann-Whitney test for comparing the solution costs of a set of algorithms each time any of them improve.
- Can also use KS or Birnbaum-Hall analogues to test inequality of ECDFs.
- Note also that these bivariate ECDFs do not capture dependence between solution cost and run-time.

Nutshell Summary?

- Univariate case is well studied and principled analysis is (relatively) straightforward.
- Multivariate case gets hairy quickly, is still an active area of research.
- Advanced methods for the multivariate case haven’t really been explored.
- Try to be as principled as possible, perhaps?