

Evolutionary Methods on Synthetic Data

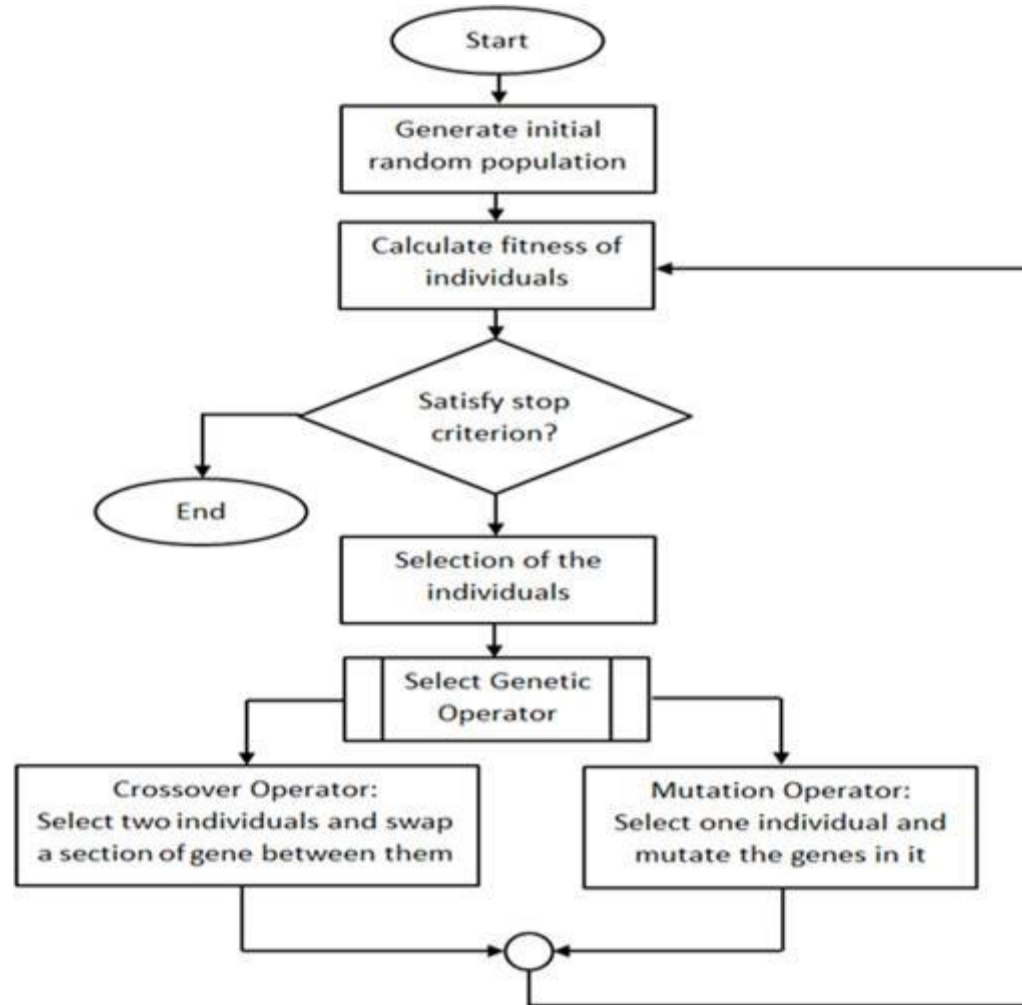
Yingrui Chen, Mark Elliot, Joe
Sakshaug

University of Manchester

Evolutionary Methods

- Artificial Intelligence
- Machine learning
- Key Features:
 - A set of criteria (or objectives) to optimise
 - A set of passible actions
 - e.g. Mutation, crossover, selection.
 - Selection mechanism
 - Iteration
 - Able to improve on the basis of an evaluation against the criteria

Genetic Algorithms



- Objectives/Criteria:
 - selected statistical properties
- Initial population:
 - A group of synthetic datasets that are generated using the univariate distributions of the original dataset.
- Selection mechanism:
 - ranking based selection scheme.

- Crossover operator:

$$\begin{pmatrix}
 \boxed{x_{11}^1} & x_{12}^1 & \dots & x_{1k}^1 \\
 \boxed{x_{21}^1} & \boxed{x_{22}^1} & \dots & x_{2k}^1 \\
 x_{31}^1 & x_{32}^1 & \dots & \boxed{x_{3k}^1} \\
 x_{41}^1 & \boxed{x_{42}^1} & \dots & \boxed{x_{4k}^1} \\
 x_{51}^1 & x_{52}^1 & \dots & \boxed{x_{5k}^1} \\
 x_{61}^1 & x_{62}^1 & \dots & x_{6k}^1 \\
 \vdots & \vdots & \vdots & \vdots \\
 x_{m1}^1 & x_{m2}^1 & \dots & x_{mk}^1
 \end{pmatrix}
 \begin{pmatrix}
 \boxed{x_{11}^2} & x_{12}^2 & \dots & x_{1k}^2 \\
 \boxed{x_{21}^2} & \boxed{x_{22}^2} & \dots & x_{2k}^2 \\
 x_{31}^2 & x_{32}^2 & \dots & \boxed{x_{3k}^2} \\
 x_{41}^2 & \boxed{x_{42}^2} & \dots & \boxed{x_{4k}^2} \\
 x_{51}^2 & x_{52}^2 & \dots & \boxed{x_{5k}^2} \\
 x_{61}^2 & x_{62}^2 & \dots & x_{6k}^2 \\
 \vdots & \vdots & \vdots & \vdots \\
 x_{m1}^2 & x_{m2}^2 & \dots & x_{mk}^2
 \end{pmatrix}
 \longrightarrow
 \begin{pmatrix}
 x_{11}^2 & x_{12}^1 & \dots & x_{1k}^1 \\
 x_{21}^2 & x_{22}^2 & \dots & x_{2k}^1 \\
 x_{31}^1 & x_{32}^2 & \dots & x_{3k}^2 \\
 x_{41}^1 & x_{42}^2 & \dots & x_{4k}^2 \\
 x_{51}^1 & x_{52}^2 & \dots & x_{5k}^2 \\
 x_{61}^1 & x_{62}^1 & \dots & x_{6k}^1 \\
 \vdots & \vdots & \vdots & \vdots \\
 x_{m1}^1 & x_{m2}^1 & \dots & x_{mk}^1
 \end{pmatrix}
 \begin{pmatrix}
 x_{11}^1 & x_{12}^2 & \dots & x_{1k}^2 \\
 x_{21}^1 & x_{22}^1 & \dots & x_{2k}^2 \\
 x_{31}^2 & x_{32}^1 & \dots & x_{3k}^1 \\
 x_{41}^2 & x_{42}^1 & \dots & x_{4k}^1 \\
 x_{51}^2 & x_{52}^1 & \dots & x_{5k}^1 \\
 x_{61}^2 & x_{62}^2 & \dots & x_{6k}^2 \\
 \vdots & \vdots & \vdots & \vdots \\
 x_{m1}^2 & x_{m2}^2 & \dots & x_{mk}^2
 \end{pmatrix}$$

- Mutation operator

$$\begin{pmatrix}
 x_{11}^j & x_{12}^j & \dots & x_{1k}^j \\
 \boxed{x_{21}^j} & x_{22}^j & \dots & x_{2k}^j \\
 x_{31}^j & \boxed{x_{32}^j} & \dots & \boxed{x_{3k}^j} \\
 x_{41}^j & \boxed{x_{42}^j} & \dots & \boxed{x_{4k}^j} \\
 x_{51}^j & x_{52}^j & \dots & \boxed{x_{5k}^j} \\
 x_{61}^j & x_{62}^j & \dots & \boxed{x_{6k}^j} \\
 \vdots & \vdots & \vdots & \vdots \\
 x_{m1}^j & x_{m2}^j & \dots & x_{mk}^j
 \end{pmatrix}
 \longrightarrow
 \begin{pmatrix}
 x_{11}^j & x_{12}^j & \dots & x_{1k}^j \\
 x_{21}^{j*} & x_{22}^j & \dots & x_{2k}^j \\
 x_{31}^j & x_{32}^{j*} & \dots & x_{3k}^{j*} \\
 x_{41}^j & x_{42}^{j*} & \dots & x_{4k}^{j*} \\
 x_{51}^j & x_{52}^j & \dots & x_{5k}^{j*} \\
 x_{61}^j & x_{62}^j & \dots & x_{6k}^{j*} \\
 \vdots & \vdots & \vdots & \vdots \\
 x_{m1}^j & x_{m2}^j & \dots & x_{mk}^j
 \end{pmatrix}$$

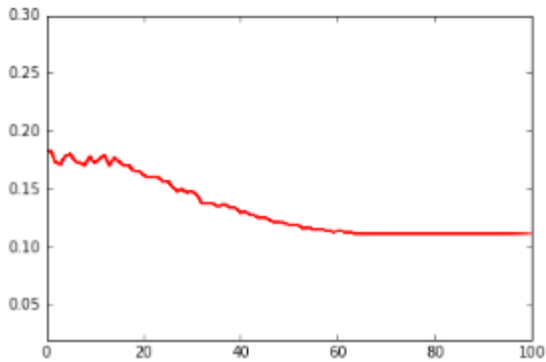
Pilot Experiments

- Dataset: 1000 cases, 4 variables, categorical, single-level microdata
- Single objective: Mean of differences between synthetic and original datasets in values of Cramer's V for all bivariate cross-classifications:

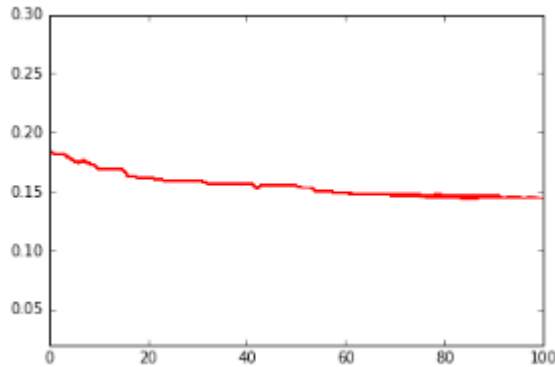
$$F = \binom{4}{2}^{-1} \sum_{i=1}^3 \sum_{j=i+1}^4 |\phi_c(x'_i, x'_j) - \phi_c(x_i, x_j)|$$

Results

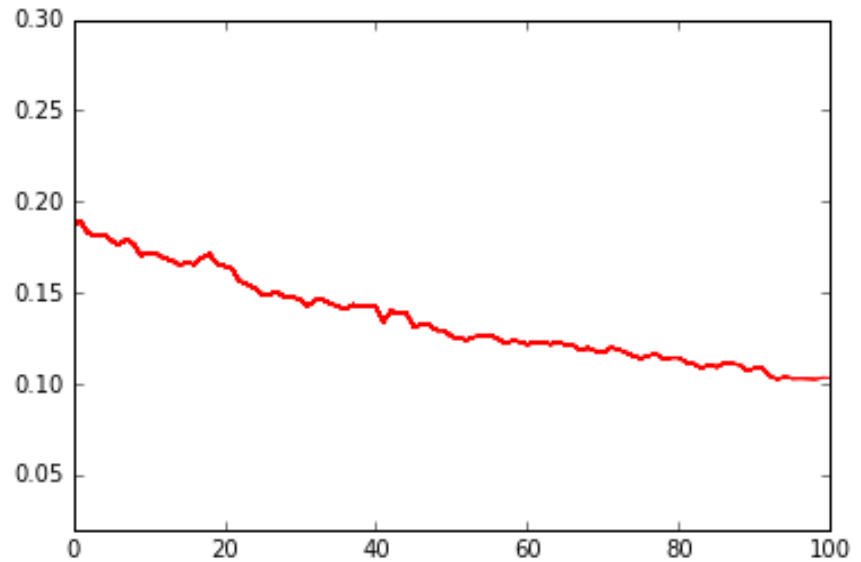
Minimum divergence (best fitness) values
in current iteration vs. number of iterations



crossover only



mutation (mutation rate=0.1) only



General GA model

Next steps

- More crossover/mutation mechanisms
 - Randomly select attributes to crossover and mutate.
 - Adaptive crossover/mutation.
 - Targets: Not forcing all candidates into evolution, trying to keep more features from good parents.
- Multi-objectives
 - Other analytical properties (e.g. the full hierarchical equivalence structure).
 - Disclosure risk measures.