

Evolutionary Dataset Optimisation: learning algorithm quality through evolution

Henry Wilde · Vincent Knight · Jonathan Gillard

Received: date / Accepted: date

Abstract In this paper we propose a new method for learning how algorithms perform. Classically, algorithms are compared on a finite number of existing (or newly simulated) benchmark data sets based on some fixed metrics. The algorithm(s) with the smallest value of this metric are chosen to be the ‘best performing’.

We offer a new approach to flip this paradigm. We instead aim to gain a richer picture of the performance of an algorithm by generating artificial data through genetic evolution, the purpose of which is to create populations of datasets for which a particular algorithm performs well. These data sets can be studied to learn as to what attributes lead to a particular progress of a given algorithm.

Following a detailed description of the algorithm as well as a brief description of an open source implementation, a number of numeric experiments are presented to show the performance of the method which we call Evolutionary Dataset Optimisation.

Keywords Evolutionary algorithm · Optimisation · Algorithm design · Artificial data generation

1 Introduction

This work presents a novel approach to learning the quality and performance of an algorithm through the use of evolution. When an algorithm is developed to solve a given problem, the designer is presented with questions about the performance of their proposed method, and its relative performance against existing methods. This is an inherently difficult task. However, under the current paradigm, the standard response to this situation is to use a known fixed set of datasets - or simulate new data sets themselves - and a common metric amongst the proposed method and its competitors. The algorithm is then assessed based on this metric with often minimal consideration for both the appropriateness or reliability of the datasets being used, and the robustness of the method in question.

This notion is not so easily observed when travelling in the opposite direction. Suppose that, instead, the benchmark was a dataset of particular interest and a preferable algorithm was to be determined for some task. There exist a number of methods employed across

disciplines to complete this task that take into account the characteristics of the data and the context of the research problem. These methods include the use of diagnostic tests. For instance, in the case of clustering, if the data displayed an indeterminate number of non-convex blobs, then one could recommend that an appropriate clustering algorithm would be DBSCAN [4]. Otherwise, for scalability, k -means may be chosen [23].

The approach presented in this work aims to flip the paradigm described here by allowing the data itself to be unfixed. This fluidity in the data is achieved by generating data for which the algorithm performs well (or better than some other) through the use of an evolutionary algorithm. The purpose of doing so is not to simply create a bank of useful datasets but rather to allow for the subsequent studying of these datasets. In doing so, the attributes and characteristics which lead to the success (or failure) of the algorithm may be described, giving a broader understanding of the algorithm on the whole. Our framework is described in Figure 1.

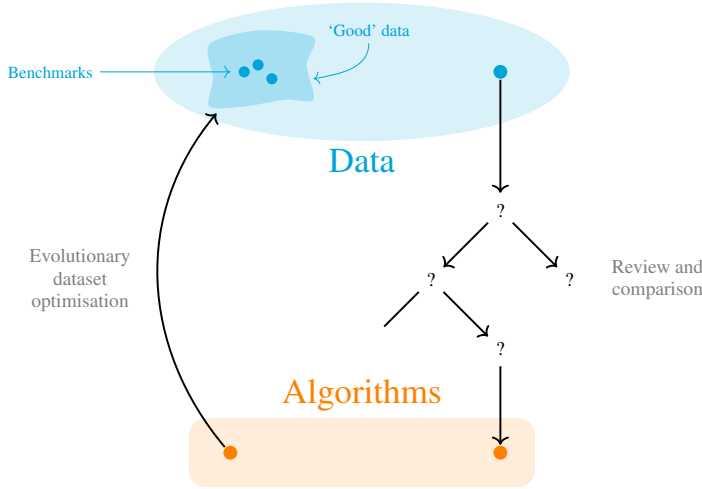


Fig. 1: On the right: the current path for selecting some algorithm(s) based on their validity and performance for a given dataset. On the left: the proposed flip to better understand the space in which ‘good’ datasets exist for an algorithm.

This proposed flip has a number of motivations, and below is a non-exhaustive list of some of the problems that are presented by the established evaluation paradigm:

1. How are these benchmark examples selected? There is no true measure of their reliability other than their frequent use. In some domains and disciplines there are well-established benchmarks so those found through literature may well be reliable, but in others less so.
2. Sometimes, when there is a lack of benchmark examples, a ‘new’ dataset is simulated to assess the algorithm. This begs the question as to how and why that simulation is created. Not only this, but the origins of existing benchmarks is often a matter of convenience rather than their merit.
3. In disciplines where there are established benchmarks, there may still be underlying problems around the true performance of an algorithm:

- (i) As an example, work by Torralba and Efros [22] showed that image classifiers trained and evaluated on a particular dataset, or datasets, did not perform reliably when evaluated using other benchmark datasets that were determined to be similar. Thus leading to a model which lacks robustness.
- (ii) The amount of learning one can gain as to the characteristics of data which lead to good (or bad) performance of an algorithm is constrained to the finite set of attributes present in the benchmark data chosen in the first place.

Evolutionary algorithms (EAs) have been applied successfully to solve a wide array of problems - particularly where the complexity of the problem or its domain are significant. These methods are highly adaptive and their population-based construction (displayed in Figure 2) allows for the efficient solving of problems that are otherwise beyond the scope of traditional search and optimisation methods.

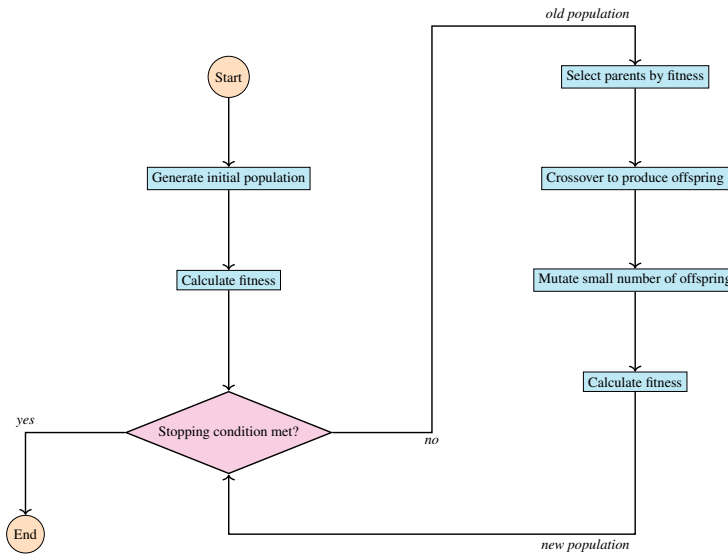


Fig. 2: A general schematic for an evolutionary algorithm.

The use of EAs to generate artificial data is not a new concept. Its applications in data generation have included developing methods for the automated testing of software [9, 15, 18] and the synthesis of existing or confidential data [3]. Such methods also have a long history in the parameter optimisation of algorithms, and recently in the automated design of convolutional neural network (CNN) architecture [19, 20].

Other methods for the generation or synthesis of artificial data include simulated annealing [13] and generative adversarial networks (GANs) [5]. The unconstrained learning style of methods such as CNNs and GANs aligns with that proposed in this work. By allowing the EA to explore and learn about the search space in an organic way, less-prejudiced insight can be established that is not necessarily reliant on any particular framework or agenda.

Note that the proposed methodology is not simply to use an EA to optimise an algorithm over a search space with fixed dimension or datatype such as those set out in [3]. The size and sample space itself is considered as a property that can be traversed through the algorithm.

2 The evolutionary algorithm

In this section, the details of an algorithm that generates data for which a given function or, equivalently, an algorithm which is well suited, is described. This algorithm is to be referred to as “Evolutionary Dataset Optimisation” (EDO).

The EDO method is built as an evolutionary algorithm which follows a traditional (generic) schema with some additional features that keep the objective of artificial data generation in mind. With that, there are a number of parameters that are passed to EDO; the typical parameters of an evolutionary algorithm are a fitness function, f , which maps from an individual to a real number, as well as a population size, N , a maximum number of iterations, M , a selection parameter, b , and a mutation probability, p_m . In addition to these, EDO takes the following parameters:

- A set of probability distribution families, \mathcal{P} . Each family in this set has some parameter limits which form a part of the overall search space. For instance, the family of normal distributions, denoted by $N(\mu, \sigma^2)$, would have limits on values for the mean, μ , and the standard deviation, σ .
- A maximum number of “subtypes” for each family in \mathcal{P} . A subtype is an independent copy of the family that progresses separate from the others. These are the actual distribution objects which are traversed in the optimisation.
- A probability vector to sample distributions from \mathcal{P} , $w = (w_1, \dots, w_{|\mathcal{P}|})$.
- Limits on the number of rows an individual dataset can have, $R \in \{(r_{\min}, r_{\max}) \in \mathbb{N}^2 \mid r_{\min} \leq r_{\max}\}$
- Limits on the number of columns a dataset can have, $C := (C_1, \dots, C_{|\mathcal{P}|})$ where $C_j \in \{(c_{\min}, c_{\max}) \in (\mathbb{N} \cup \{\infty\})^2 \mid c_{\min} \leq c_{\max}\}$ for each $j = 1, \dots, |\mathcal{P}|$. That is, C defines the minimum and maximum number of columns a dataset may have from each distribution in \mathcal{P} .
- A second selection parameter, $l \in [0, 1]$, to allow for a small proportion of ‘lucky’ individuals to be carried forward.
- A shrink factor, $s \in [0, 1]$, defining the relative size of a component of the search space to be retained after adjustment.

The concepts discussed in this section form the mechanisms of the evolutionary dataset optimisation algorithm. To use the algorithm practically, these components have been implemented in Python as a library built on the scientific Python stack [14, 16]. The library is fully tested and documented (at <https://edo.readthedocs.io>) and is freely available online under the MIT license [21]. The EDO implementation was developed to be consistent with the current best practices of open source software development [8].

The statement of the EDO algorithm is presented here to lay out its general structure from a high level perspective. Lower level discussion is provided below where additional algorithms for the individual creation, evolutionary operator and shrinkage processes are given along with diagrams (where appropriate).

Note that there are no defined processes for how to stop the algorithm or adjust the mutation probability, p_m . This is down to their relevance to a particular use case. Some examples include:

- Regular decreasing in mutation probability across the available attributes [10].
- Stopping when no improvement in the best fitness is found within some K consecutive iterations [11].
- Utilising global behaviours in fitness to indicate a stopping point [12].

Algorithm 1: The evolutionary dataset optimisation algorithm**Input:** $f, N, R, C, \mathcal{P}, w, M, b, l, p_m, s$ **Output:** A full history of the populations and their fitnesses.

```

begin
  create initial population of individuals
  find fitness of each individual
  record population and its fitness
  while current iteration less than the maximum and stopping condition not met do
    select parents based on fitness and selection proportions
    use parents to create new population through crossover and mutation
    find fitness of each individual
    update population and fitness histories
    if adjusting the mutation probability then
      | update mutation probability
    end
    if using a shrink factor then
      | shrink the mutation space based on parents
    end
  end
end

```

Algorithm 2: Creating a new population**Input:** parents, $N, R, C, \mathcal{P}, w, p_m$ **Output:** A new population of size N

```

begin
  add parents to the new population
  while the size of the new population is less than N do
    sample two parents at random
    create an offspring by crossing over the two parents
    mutate the offspring according to the mutation probability
    add the mutated offspring to the population
  end
end

```

2.1 Individuals

Evolutionary algorithms operate in an iterative process on populations of individuals that each represent a solution to the problem in question. In a genetic algorithm, an individual is a solution encoded as a bit string of, typically, fixed length and treated as a chromosome-like object to be manipulated. In EDO, as the objective is to generate datasets and explore the space in which datasets exist, there is no encoding. As such the distinction is made that EDO is an evolutionary algorithm.

As is seen in Figure 3, an individual's creation is defined by the generation of its columns. A set of instructions on how to sample new values (in mutation, for instance, Section 2.4) for that column are recorded in the form of a probability distribution. These distributions are sampled and created from the families passed in \mathcal{P} . In EDO, the produced datasets and their metadata are manipulated directly so that the biological operators can be designed and be interpreted in a more meaningful way as will be seen later in this section.

However, one should not assume that the columns are a reliable representative of the distribution associated with them, or vice versa. This is particularly true of 'shorter' datasets with a small number of rows, whereas confidence in the pair could be given more liberally

for ‘longer’ datasets with a larger number of rows. In any case, appropriate methods for analysis should be employed before formal conclusions are made.

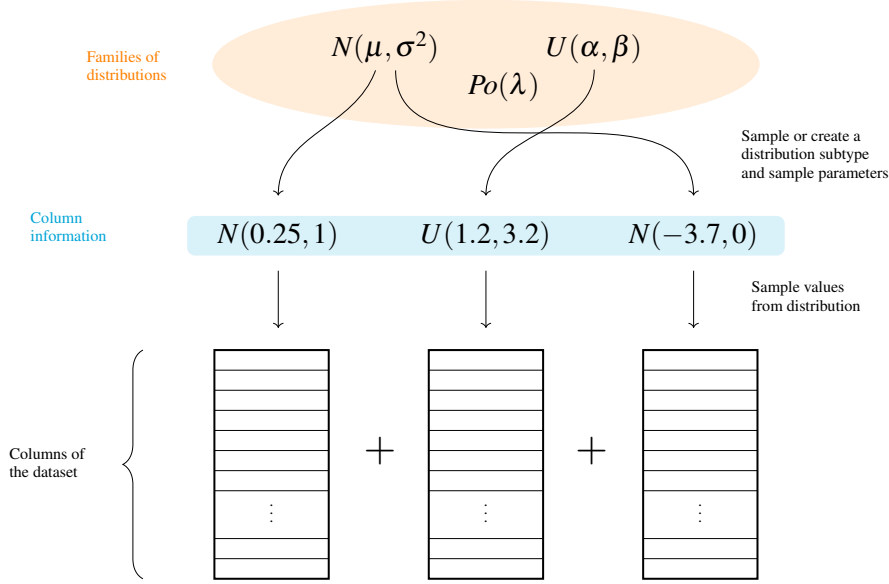


Fig. 3: An example of how an individual is first created.

Algorithm 3: Creating an individual

Input: R, C, \mathcal{P}, w

Output: An individual defined by a dataset and some metadata

begin

 sample a number of rows and columns

 create an empty dataset

for each column in the dataset do

 sample a distribution from \mathcal{P}

 create an instance of the distribution

 fill in the column by sampling from this instance

 record the instance in the metadata

end

end

2.2 Selection

The selection operator describes the process by which individuals are chosen from the current population to generate the next. Almost always, the likelihood of an individual being selected is determined by their fitness. This is because the purpose of selection is to preserve favourable qualities and encourage some homogeneity within future generations [2].

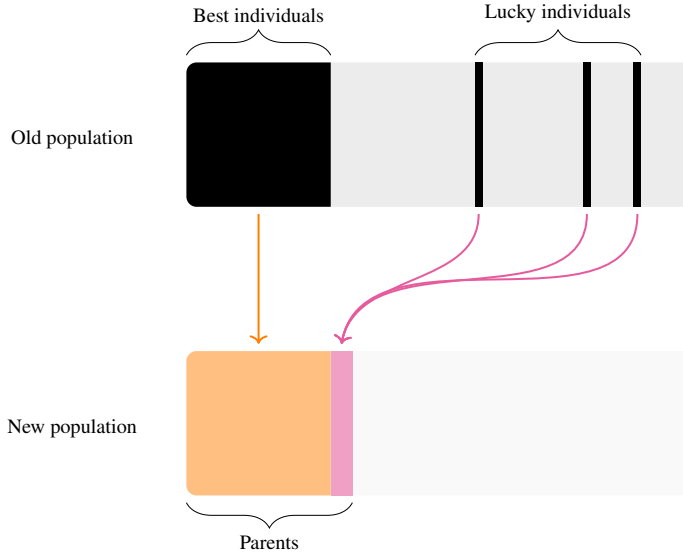


Fig. 4: The selection process with the inclusion of some lucky individuals.

Algorithm 4: The selection process

Input: population, population fitness, b , l
Output: A set of parent individuals
begin
 calculate n_b and n_l
 sort the population by the fitness of its individuals
 take the first n_b individuals and make them parents
 if there are any individuals left then
 take the next n_l individuals and make them parents
 end
end

In EDO, a modified truncation selection method is used [7], as can be seen in Figure 4. Truncation selection takes a fixed number, $n_b = \lceil bN \rceil$, of the fittest individuals in a population and makes them the ‘parents’ of the next. It has been observed that, despite its efficiency as a selection operator, truncation selection can lead to premature convergence at local optima [Tatsuya2002, 7]. The modification for EDO is an optional stage after the best individuals have been chosen: with some small l , a number, $n_l = \lceil lN \rceil$, of the remaining individuals can be selected at random to be carried forward. Hence, allowing for a small number of randomly selected individuals may encourage diversity and further exploration throughout the run of the algorithm. It should be noted that regardless of this step, an individual could potentially be present throughout the entirety of the algorithm.

After the parents have been selected, there are two adjustments made to the current search space. The first is that the subtypes for each family in \mathcal{P} are updated to only those present in the parents. The second adjustment is a process which acts on the distribution parameter limits for each subtype in \mathcal{P} . This adjustment gives the ability to ‘shrink’ the search space about the region observed in a given population. This method is based on a

power law described in [1] that relies on a shrink factor, s . At each iteration, t , every distribution subtype which is present in the parents has its parameter's limits, (l_t, u_t) , adjusted. This adjustment is such that the new limits, (l_{t+1}, u_{t+1}) are centred about the mean observed value, μ , for that parameter:

$$l_{t+1} = \max \left\{ l_t, \mu - \frac{1}{2}(u_t - l_t)s^t \right\} \quad (1)$$

$$u_{t+1} = \min \left\{ u_t, \mu + \frac{1}{2}(u_t - l_t)s^t \right\} \quad (2)$$

The shrinking process is given explicitly in Algorithm 5. Note that the behaviour of this process can produce reductive results for some use cases and is optional.

Algorithm 5: Shrinking the mutation space

Input: parents, current iteration, \mathcal{P}, M, s
Output: A new mutation space focussed around the parents

```

begin
  for each distribution subtype in  $\mathcal{P}$  do
    for each parameter of the distribution do
      get the current values for parameter over all parent columns
      find the mean of the current values
      find the new lower (1) and upper (2) bounds around the mean
      set the parameter limits
    end
  end
end

```

2.3 Crossover

Crossover is the operation of combining two individuals in order to create at least one offspring. In genetic algorithms, the term ‘crossover’ can be taken literally: two bit strings are crossed at a point to create two new bit strings. Another popular method is uniform crossover, which has been favoured for its efficiency and efficacy in combining individuals [17]. For EDO, this method is adapted to support dataset manipulation: a new individual is created by uniformly sampling each of its components (dimensions and then columns) from a set of two ‘parent’ individuals, as shown in Figure 5.

Observe that there is no requirement on the dimensions of the parents to be of similar or equal shapes. This is because the driving aim of the proposed method is to explore the space of all possible datasets. In the case where there is incongruence in the lengths of the two parents, missing values may appear in a shorter column that is sampled. To resolve this, values are sampled from the probability distribution associated with that column to fill in these gaps.

2.4 Mutation

Mutation is used in evolutionary algorithms to encourage a broader exploration of the search space at each generation. Under this framework, the mutation process manipulates the phe-

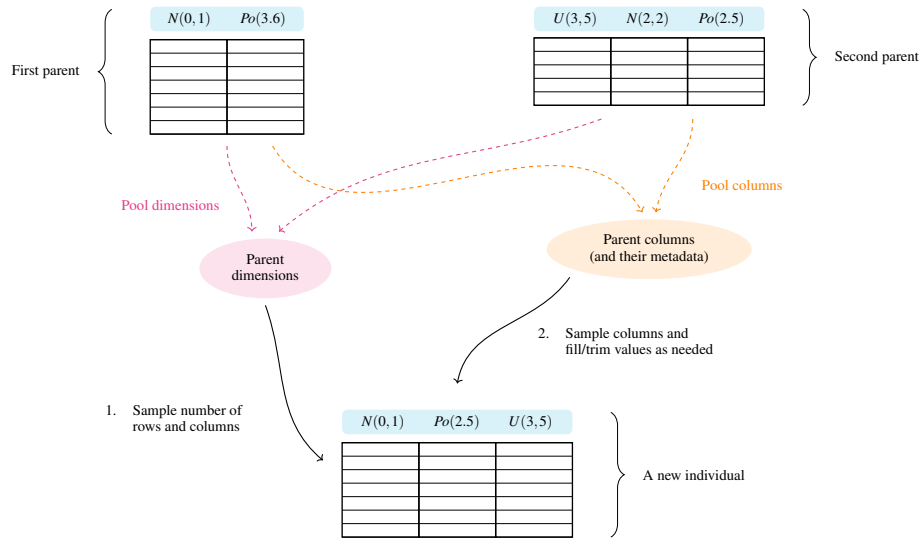


Fig. 5: The crossover process between two individuals with different dimensions.

Algorithm 6: The crossover process**Input:** Two parents**Output:** An offspring made from the parents ready for mutation**begin**

collate the columns and metadata from each parent in a pool
 sample each dimension from between the parents uniformly
 form an empty dataset with these dimensions

for each column in the dataset do

sample a column (and its corresponding metadata) from the pool

if this column is longer than required then

randomly select entries and delete them as needed

end**if this column is shorter than required then**

sample new values from the metadata and append them to the column as needed

end

add this column to the dataset and record its metadata

end**end**

notype of an individual where numerous things need to be modified including an individual's dimensions, column metadata and the entries themselves. This process is described in Figure 6.

As shown in Figure 6, each of the potential mutations occur with the same probability, p_m . However, the way in which columns are maintained assure that (assuming appropriate choices for f and \mathcal{P}) many mutations in the metadata and the dataset itself will only result in some incremental change in the individual's fitness relative to, say, a completely new individual.

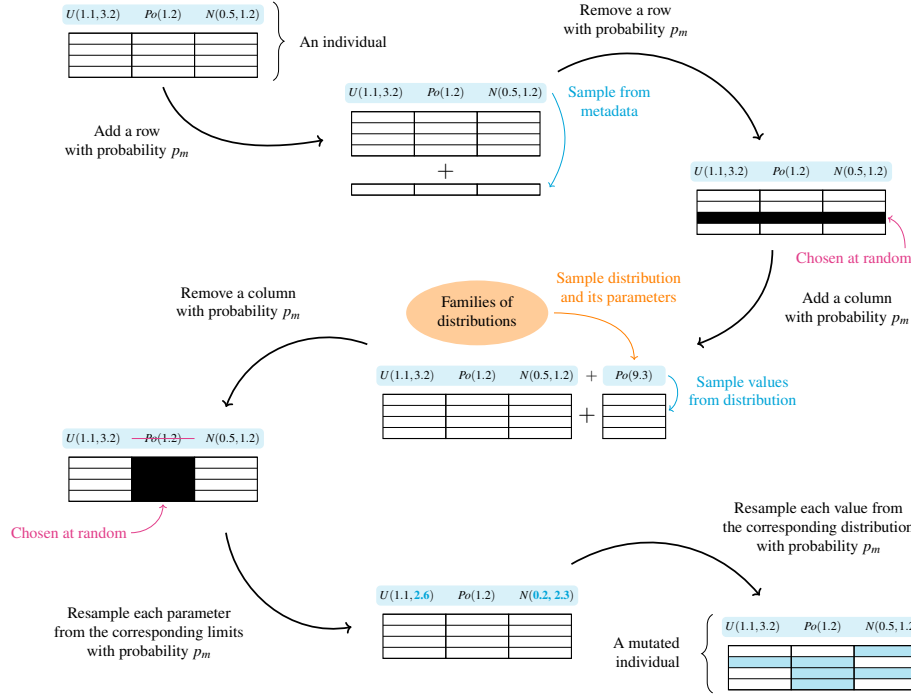


Fig. 6: The mutation process.

3 Examples

3.1 k -means clustering

The following examples act as a form of validation for EDO, and also highlight some of the nuances in its use. The examples will be focused around the clustering of data and, in particular, the k -means (Lloyd's) algorithm. Clustering was chosen as it is a well-understood problem that is easily accessible - especially when restricted to two dimensions. The k -means algorithm is an iterative, centroid-based method that aims to minimise the 'inertia' of the current partition, $Z = \{Z_1, \dots, Z_k\}$, of some dataset X :

$$I(Z, X) := \frac{1}{|X|} \sum_{j=1}^k \sum_{x \in Z_j} d(x, z_j)^2 \quad (3)$$

A full statement of the algorithm to minimise (3) is given in ??.

This inertia function is taken as the objective of the k -means algorithm, and is used for evaluating the final clustering. This is particularly true when the algorithm is not being considered as an unsupervised classifier where accuracy may be used [6]. With that, the first example is to use this inertia as the fitness function in EDO. That is, to find datasets which minimise I .

For the purposes of visualisation, in this example EDO is restricted to only two-dimensional datasets, i.e. $C = ((2, 2))$. In addition to this, all columns are formed from uniform distributions where the bounds are sampled from the unit interval. Thus, the only family in \mathcal{P}

Algorithm 7: The mutation process**Input:** An individual, p_m , R , C , \mathcal{P} , w **Output:** A mutated individual**begin** sample a random number $r \in [0, 1]$ **if** $r < p_m$ *and adding a row would not violate R* **then**

sample a value from each distribution in the metadata

append these values as a row to the end of the dataset

end sample a new $r \in [0, 1]$ **if** $r < p_m$ *and removing a row would not violate R* **then**

remove a row at random from the dataset

end sample a new $r \in [0, 1]$ **if** $r < p_m$ *and adding a new column would not violate C* **then** create a new column using \mathcal{P} and w

append this column to the end of the dataset

end sample a new $r \in [0, 1]$ **if** $r < p_m$ *and removing a column would not violate C* **then**

remove a column (and its associated metadata) at random from the dataset

end **for each distribution in the metadata do** **for each parameter of the distribution do** sample a random number $r \in [0, 1]$ **if** $r < p_m$ **then**

sample a new value from within the distribution parameter limits

update the parameter value with this new value

end **end** **end** **for each entry in the dataset do** sample a random number $r \in [0, 1]$ **if** $r < p_m$ **then**

sample a new value from the associated column distribution

update the entry with this new value

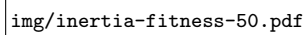
end **end****end**

is:

$$\mathcal{U} := \{U(a, b) \mid a, b \in [0, 1]\} \quad (4)$$

The remaining parameters are as follows: $N = 100$, $R = (3, 100)$, $M = 1000$, $b = 0.2$, $l = 0$, $p_m = 0.01$, and shrinkage excluded. Figure ?? shows an example of the fitness (above) and dimension (below) progression of the evolutionary algorithm under these conditions up until the 50th epoch.

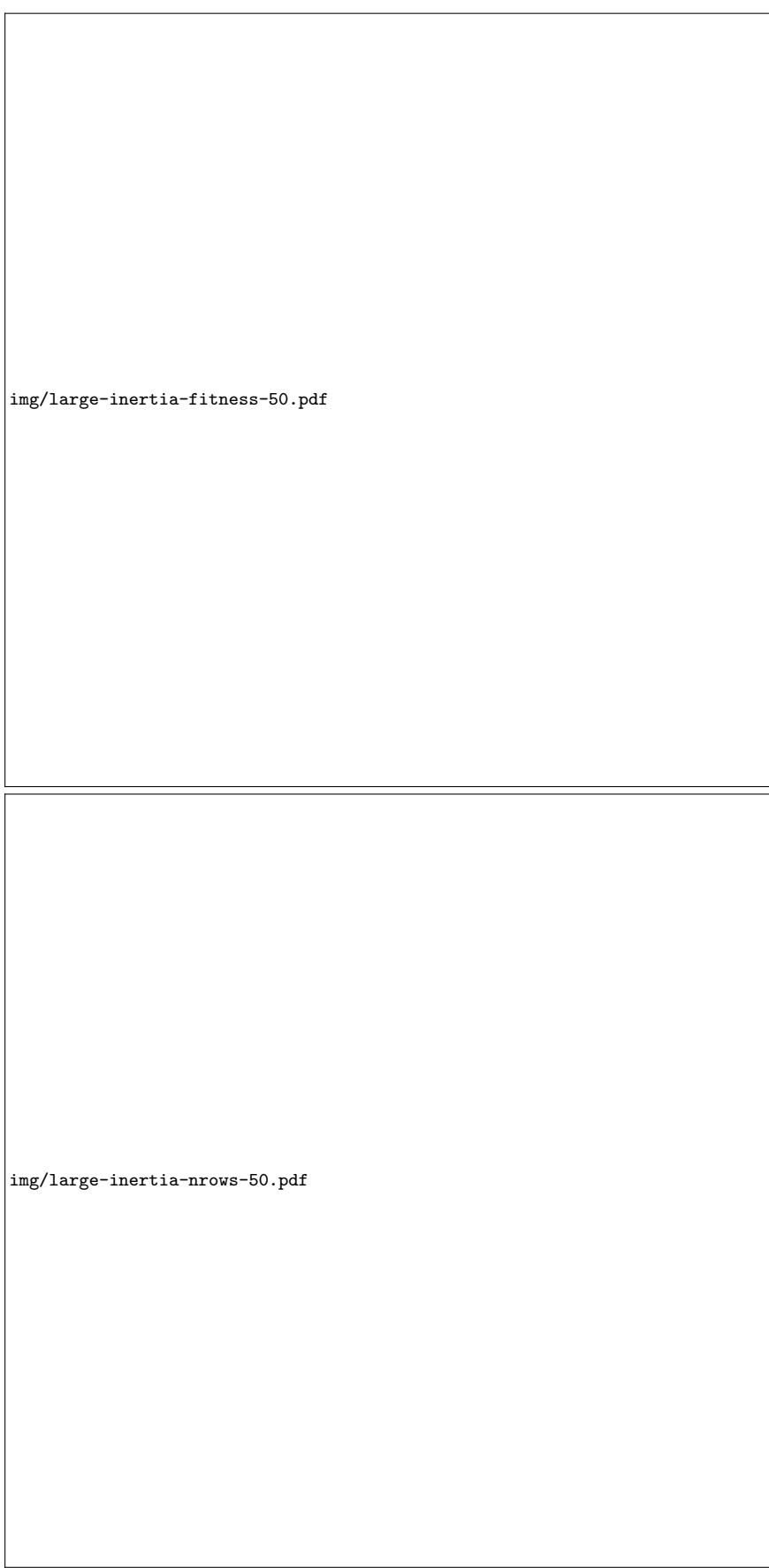
There is a steep learning curve here; within the first 50 generations an individual is found with a fitness of roughly 10^{-10} which could not be improved on for a further 900 epochs. The same quick convergence is seen in the number of rows. This behaviour is quickly recognised as preferable and was dominant across all the trials conducted in this work. This preference for datasets with fewer rows makes sense given that I is the sum of the mean error from each cluster centre. With that, when k is fixed *a priori*, reducing the number of points in each cluster (i.e. the terms of the second summation) quickly reduces the mean error of that cluster and thus the value of I .



img/inertia-fitness-50.pdf

img/inertia-nrows-50.pdf

Fig. 7a: Progressions for final inertia and dimension across the first 50 epochs with $R = (3, 100)$.



img/large-inertia-fitness-50.pdf

img/large-inertia-nrows-50.pdf

Fig. 7b: Progressions for final inertia and dimension across the first 50 epochs with $R = (50, 100)$.

Something that may be seen as unwanted is a compaction of the cluster centres. Referring to Figure ??, the best and median individuals show two clusters that are essentially the same point whereas the worst is a random cloud across the whole of \mathcal{U} which was found in the initial population. The kind of behaviour exhibited by the best performing individuals occurs in part because it is allowed. There are two immediate ways in which this allowed: first, that the ‘trivial’ case is included in R and, secondly, that the fitness function does nothing to penalise the proximity of the inter-cluster means, as well as aiming to reduce the intra-cluster means. This kind of unwanted behaviour highlights a subtlety in how EDO should be used; that experimentation and rigour are required to properly understand an algorithm’s quality.

Hence, consider Figure ?? where the individuals have been generated with the same parameters as previously except with adjusted row limits, $R = (50, 100)$, so as to exclude this trivial case. In these trials, the results are equivalent: the worst performing individuals are without structure whilst the best-performing individuals display clusters that are dense about a single point despite the minimum number of rows being increased. Perhaps then, this compacted clustering is ‘optimal’.

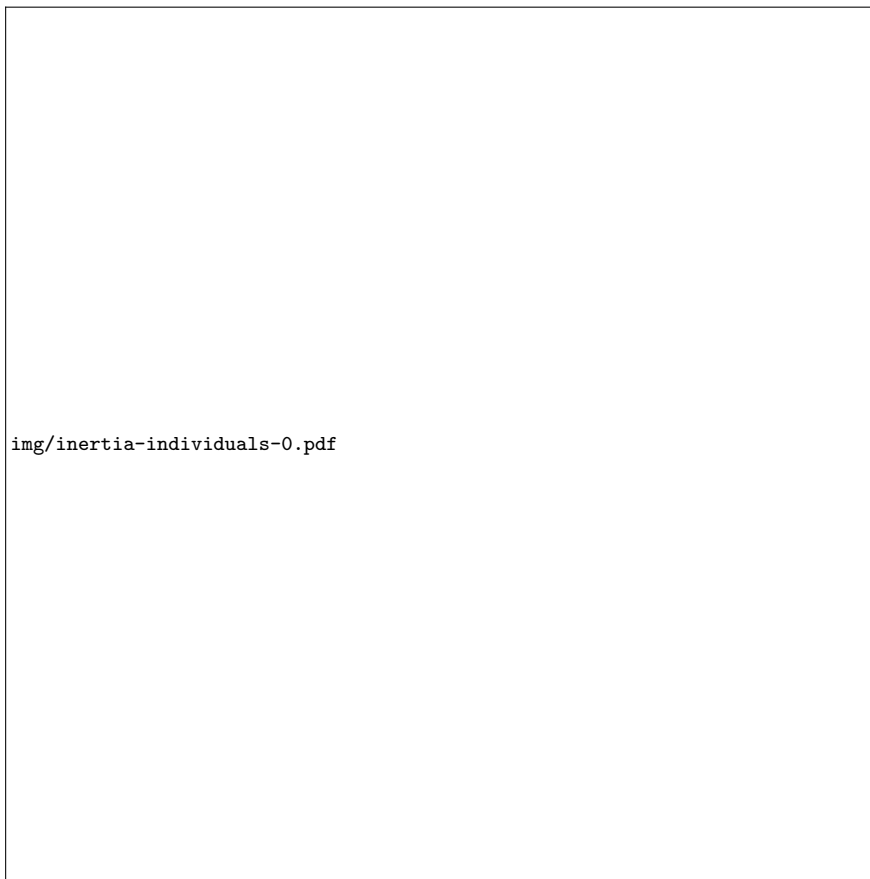
However, more extensive studying may be done. That is, the defined fitness function may require further attention. Indeed, the final inertia could be considered a flawed or fragile fitness function if it is supposed to evaluate the appropriateness or efficacy of the k -means algorithm. Incorporating the inter-cluster spread to the fitness of an individual dataset can reduce this observed compaction. The silhouette coefficient is a metric used to evaluate the appropriateness of a clustering to a dataset, and is given by the mean of the silhouette value, $S(x)$, of each point $x \in Z_j$ in each cluster:

$$\begin{aligned} A(x) &:= \frac{1}{|Z_j| - 1} \sum_{y \in Z_j \setminus \{x\}} d(x, y), \\ B(x) &:= \min_{k \neq j} \frac{1}{|Z_k|} \sum_{w \in Z_k} d(x, w), \\ S(x) &:= \begin{cases} \frac{B(x) - A(x)}{\max\{A(x), B(x)\}} & \text{if } |Z_j| > 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (5)$$

The optimisation of the silhouette coefficient is analogous to finding a dataset which increases both the intra-cluster cohesion (the inverse of A) and inter-cluster separation (B). Hence, the inertia is addressed by maximising cohesion. Meanwhile, the spread of the clusters themselves is considered by maximising separation.

Repeating the trials with the same parameters as with inertia, the silhouette fitness function yields the results summarised in Figures ?? and ?. Irrespective of row limits, the datasets produced show increased separation from one another whilst maintaining low values in the final inertia of the clustering as shown in Figure ?. Again, the form of the individual clusters is much the same. The low values of inertia correspond to tight clusters, and the tightest clusters are those with a minimal number of points, i.e. a single point. As with the previous example, albeit at a much slower rate, the preferable individuals are those leading toward this case. That this gradual reduction in the dimension of the individuals occurs after the improvement of the fitness function bolsters the claim that the base case is also optimal.

However, due to the nature of the implementation, any individual from any generation may be retrieved and studied should the final results be too concentrated on any given case.



(a)



(b)

Fig. 8: Representative individuals based on inertia with: $R = (3, 100)$; $R = (50, 100)$. Centroids displayed as crosses.