Harrix Optimization Testing 1.0

А. Б. Сергиенко

12 сентября 2013 г.

Аннотация

Harrix Optimization Testing 1.0 — формат файлов вида ***.xml** для представления данных об исследовании эффективности алгоритмов оптимизации на тестовых функциях.

Содержание

1 Вводная информация

Описание данного формата файлов располагается по адресу https://github.com/Harrix/HarrixFileFormats.

С автором можно связаться по адресу sergienkoanton@mail.ru или http://vk.com/harrix. Сайт автора, где публикуются последние новости: http://blog.harrix.org, а проекты располагаются по адресу http://harrix.org.

В папке **Examples** располагаются примеры файлов рассматриваемого формата. В файле «**MHL_BinaryMonteCarloAlgorithm MHL_TestFunction_SumVector 50 2500.xml**», рассматривается, например, алгоритм без настраиваемых параметров.

2 Идея проведения исследований эффективности алгоритмов

У нас имеется некий алгоритм оптимизации с некоторым конечным набором параметров (если есть вещественные параметры, то дискретизируем каким-нибудь образом). Также имеется некоторая тестовая функция с фиксированной размерностью, на которой хотим провести исследование эффективности алгоритма оптимизации.

Для этого мы фиксируем набор параметров алгоритма и прогоняем его **number_of_runs** раз. В каждом запуске сравниваем найденные решения с оптимальными в результате **number_of_runs** запусков алгоритма и получаем значения ошибки по входным параметрам E_x, ошибки по значениям целевой функции E_y и надежности R.

Ошибка по входным параметрам E_x означает насколько близко найденная точка в среднем (среди запусков алгоритма в количестве **number_of_runs**) от оптимального решения в пространстве входных переменных. Чем меньше ошибка E_x, тем лучше.

Ошибка по значениям целевой функции E_y означает насколько значение целевой функции в среднем (среди запусков алгоритма в количестве **number_of_runs**) близко от значения целевой функции оптимального решения. Чем меньше ошибка E_y, тем лучше.

Надежность R показывает какая доля найденных решений от общего числа запусков алгоритма (**number_of_runs**) находится в некоторой окрестности от оптимального решения в пространстве входных переменных. Чем больше (обратите внимание!) надёжность R, тем лучше.

Для всех тестовых функций каждого класса (бинарной, вещественной и др. оптимизации) общий вид формул соответствующих показателей (E_x, E_y, R) приблизительно одинаков. Но для каждой тестовой функции имеется точная формула нахождения этих показателей. Для основного множества функций можно посмотреть тут:

https://github.com/Harrix/HarrixTestFunctions.

Для того, чтобы адекватно оценить эффективность алгоритма при разных настройках нам для каждого фиксированного набора параметров алгоритма требуется несколько точек. Поэтому мы запустили алгоритм **number_of_runs** раз, получили показатели E_x, E_y, R. И эту процедуру повторили **number_of_measuring** раз. То есть мы получили **number_of_measuring** значений параметров E_x, E_y, R.

Потом мы меняем набор параметров алгоритма и повторяем вышеописанную процедуру по получению **number_of_measuring** точек показателей E_x, E_y, R для данного набора параметров алгоритма оптимизации.

Предполагаем, что у нас было рассмотрено **number_of_experiments** вариантов настроек алгоритма. В идеальном случае мы должны рассмотреть всё множество возможных наборов настроек алгоритма оптимизации.

В итоге мы получим по **number_of_measuring**·**number_of_experiments** значений каждого показателя E_x, E_y, R.

Полученные данные и записываются в файле формата данных, описанного в данном документе.

Если решается задача условной оптимизации, то методы учета штрафов или иные способы уточнения ограничений учитываются как параметры алгоритма.

Если решается задача многокритериальной оптимизации, то специфика задачи закладывается в формулы нахождения показателей E_x, E_y, R.

3 Краткое описание формата данных

Файл формата **Harrix Optimization Testing 1.0** имеет расширение вида ***.xml**.

Файл представляет собой обычный файл формата XML. Вначале файла идет служебная информация, а потом идут непосредственно данные об эффективности алгоритма.

4 Пример файла Harrix Optimization Testing

Предложенный ниже файл не является полным исследованием алгоритма, а является лишь тестовым примером.

Код 1. Пример части файла Harrix Optimization Testing

```
<?xml version="1.0" encoding="UTF-8"?>
<document>
<harrix_file_format>
   <format>Harrix Optimization Testing</format>
   <version>1.0</version>
   <link>https://github.com/Harrix/HarrixFileFormats</link>
</harrix_file_format>
<about>
   <author>Сергиенко Антон Борисович</author>
   <date>12.08.2013 23:17:24</date>
   <email>sergienkoanton@mail.ru</email>
</about>
<about_data>
   <!-- Обозначение алгоритма (по названию функции, которая его реализует) -->
   <name_algorithm>MHL_StandartRealGeneticAlgorithm</name_algorithm>
   <!-- Полное название алгоритма -->
   <full_name_algorithm>Стандартный генетический алгоритм на вещественных строках</
      full_name_algorithm>
   <!-- Ссылка на описание алгоритма оптимизации (если нет, то NULL) -->
```

```
   <link_algorithm>https://github.com/Harrix/HarrixOptimizationAlgorithms</
      link_algorithm>
   <!-- Название тестовой функции (по названию функции, которая его реализует) -->
   <name_test_function>MHL_TestFunction_Ackley</name_test_function>
   <!-- Полное название тестовой функции -->
   <full_name_test_function>Функция Ackley</full_name_test_function>
   <!-- Ссылка на описание тестовой функции (если нет, то NULL) -->
   <link_test_function>https://github.com/Harrix/HarrixTestFunctions</
      link_test_function>
   <!-- Размерность задачи оптимизации -->
   <dimension_test_function>5</dimension_test_function>
   <!-- Количество измерений для каждого варианта настроек алгоритма (сколько точек п
      олучим) -->
   <number_of_measuring>10</number_of_measuring>
   <!-- Количество запусков алгоритма в каждом из измерений -->
   <number_of_runs>10</number_of_runs>
   <!-- Максимальное допустимое число вычислений целевой функции -->
   <max_count_of_fitness>2500</max_count_of_fitness>
   <!-- Количество проверяемых параметров алгоритма оптимизации -->
   <number_of_parameters>5</number_of_parameters>
   <!-- Количество комбинаций вариантов настроек -->
   <number_of_experiments>1</number_of_experiments>
   <!-- Все ли комбинации вариантов настроек просмотрены -->
   <all_combinations>1</all_combinations>
</about_data>
<data>
   <experiment parameters_of_algorithm_1="Тип селекции = Пропорциональная селекция"
      parameters_of_algorithm_2="Тип скрещивания = Одноточечное скрещивание"
      parameters_of_algorithm_3="Тип мутации = Слабая мутация"
      parameters_of_algorithm_4="Тип формирования нового поколения = Только потомки"
      parameters_of_algorithm_5="Тип преобразования задачи вещественной оптимизации в
       задачу бинарной оптимизации = Стандартное представление целого числа - номер у
      зла в сетке дискретизации">
      <measuring>
         <Ex>0.102733</Ex>
         <Ey>1.40394</Ey>
         <R>0</R>
      </measuring>
      <measuring>
         <Ex>0.0840828</Ex>
         <Ey>1.4134</Ey>
         <R>0</R>
      </measuring>
      <measuring>
         <Ex>0.0674963</Ex>
         <Ey>1.20694</Ey>
         <R>0</R>
      </measuring>
      <measuring>
         <Ex>0.103118</Ex>
         <Ey>1.57915</Ey>
         <R>0</R>
      </measuring>
      <measuring>
         <Ex>0.0795264</Ex>
         <Ey>1.4047</Ey>
         <R>0</R>
      </measuring>
      <measuring>
         <Ex>0.0626839</Ex>
```

```
            <Ey>1.17213</Ey>
            <R>0</R>
        </measuring>
        <measuring>
            <Ex>0.0974347</Ex>
            <Ey>1.46336</Ey>
            <R>0</R>
        </measuring>
        <measuring>
            <Ex>0.10858</Ex>
            <Ey>1.26652</Ey>
            <R>0</R>
        </measuring>
        <measuring>
            <Ex>0.0990866</Ex>
            <Ey>1.41937</Ey>
            <R>0</R>
        </measuring>
        <measuring>
            <Ex>0.0901381</Ex>
            <Ey>1.17268</Ey>
            <R>0.1</R>
        </measuring>
    </experiment>
</data>
</document>
```

5 Подробное описание формата данных

Файл имеет строгую структуру данных, которую не следует нарушать. Все тэги являются обязательными, на те или иные параметры накладываются ограничения, которые будут ниже описаны.

Первая строчка семантической нагрузки не несет, и нужна только для объявления формата XML для парсеров XML файлов. В общем, эта строчка должна быть, и ее не надо трогать.

Код 2. Первая строчка файла Harrix Optimization Testing

```
<?xml version="1.0" encoding="UTF-8"?>
```

Далее идет строчка с тэгом **<document>**, а закрывающимся тэгом **</document>** заканчивается весь документ. Внутри этого блока располагается вся информация.

Код 3. Блок <document> в файле Harrix Optimization Testing

```
<?xml version="1.0" encoding="UTF-8"?>
<document>
...
</document>
```

Внутри блока **<document>** располагаются 4 подблока (их порядок не менять):

- **<harrix_file_format>** — информация о формате данных для распознавания типа документа;
- **<about>** — информация о самом файле: авторе документа и времени создания;

- **<about_data>** — информация о исследовании, которое проводилось: алгоритм оптимизации, тестовая функция и так далее;
- **<data>** — непосредственно данные, полученные во время исследования.

Код 4. Подблоки в файле Harrix Optimization Testing

```
<?xml version="1.0" encoding="UTF-8"?>
   <document>
   <harrix_file_format>
   ...
   </harrix_file_format>
   <about>
   ...
   </about>
   <about_data>
   ...
   </about_data>
   <data>
   ...
   </data>
</document>
```

Рассмотрим каждый подблок в отдельности.

5.1 Подблок <harrix_file_format>

Код 5. Подблок в файле Harrix Optimization Testing

```
<harrix_file_format>
   <format>Harrix Optimization Testing</format>
   <version>1.0</version>
   <link>https://github.com/Harrix/HarrixFileFormats</link>
</harrix_file_format>
```

В данном подблоке всего три тэга:

- **<format>** — здесь содержится название формата данных;
- **<version>** — версия формата данных;
- **<link>** — ссылка, где находится описание данного формата, то есть данный документ.

Содержимое данных тэгов приведено выше в примере кода. Все три тэга обязательны и должны содержать именно эту информацию. То есть нужно просто скопировать этот код.

5.2 Подблок <about>

Код 6. Подблок в файле Harrix Optimization Testing

```
<about>
   <author>Сергиенко Антон Борисович</author>
   <date>12.08.2013 23:17:24</date>
   <email>sergienkoanton@mail.ru</email>
</about>
```

В данном подблоке тэги:

- **<author>** — имя автора исследования;
- **<date>** — время проведения исследования или время формирования файла;
- **<email>** — электронная почта автора исследования, чтобы с ним можно было связаться для уточнения вопросов. Если автор не хочет по каким-то причинам выставлять свой e-mail, то в качестве значения параметра надо вставить **NULL**.

Содержимое этих двух тэгов произвольное, например, дату можно записывать как хочется — никаких требований нет.

5.3 Подблок <about_data>

Код 7. Подблок в файле Harrix Optimization Testing

```
<about_data>
   <!-- Обозначение алгоритма (по названию функции, которая его реализует) -->
   <name_algorithm>MHL_StandartRealGeneticAlgorithm</name_algorithm>
   <!-- Полное название алгоритма -->
   <full_name_algorithm>Стандартный генетический алгоритм на вещественных строках</
      full_name_algorithm>
   <!-- Ссылка на описание алгоритма оптимизации (если нет, то NULL) -->
   <link_algorithm>https://github.com/Harrix/HarrixOptimizationAlgorithms</
      link_algorithm>
   <!-- Название тестовой функции (по названию функции, которая его реализует) -->
   <name_test_function>MHL_TestFunction_Ackley</name_test_function>
   <!-- Полное название тестовой функции -->
   <full_name_test_function>Функция Ackley</full_name_test_function>
   <!-- Ссылка на описание тестовой функции (если нет, то NULL) -->
   <link_test_function>https://github.com/Harrix/HarrixTestFunctions</
      link_test_function>
   <!-- Размерность задачи оптимизации -->
   <dimension_test_function>5</dimension_test_function>
   <!-- Количество измерений для каждого варианта настроек алгоритма (сколько точек п
      олучим) -->
   <number_of_measuring>10</number_of_measuring>
   <!-- Количество запусков алгоритма в каждом из измерений -->
   <number_of_runs>10</number_of_runs>
   <!-- Максимальное допустимое число вычислений целевой функции -->
   <max_count_of_fitness>2500</max_count_of_fitness>
   <!-- Количество проверяемых параметров алгоритма оптимизации -->
   <number_of_parameters>5</number_of_parameters>
   <!-- Количество комбинаций вариантов настроек -->
   <number_of_experiments>1</number_of_experiments>
   <!-- Все ли комбинации вариантов настроек просмотрены -->
   <all_combinations>1</all_combinations>
</about_data>
```

Строчки вида **<!– –>**, например:

Код 8. Комментарий в файле Harrix Optimization Testing

```
<!-- Количество комбинаций вариантов настроек -->
```

являются комментариями и могут быть удалены без ущерба для файла.

Код 9. Подблок без комментариев в файле Harrix Optimization Testing

```
<about_data>
   <name_algorithm>MHL_StandartRealGeneticAlgorithm</name_algorithm>
   <full_name_algorithm>Стандартный генетический алгоритм на вещественных строках</
      full_name_algorithm>
   <link_algorithm>https://github.com/Harrix/HarrixOptimizationAlgorithms</
      link_algorithm>
   <name_test_function>MHL_TestFunction_Ackley</name_test_function>
   <full_name_test_function>Функция Ackley</full_name_test_function>
   <link_test_function>https://github.com/Harrix/HarrixTestFunctions</
      link_test_function>
   <dimension_test_function>5</dimension_test_function>
   <number_of_measuring>10</number_of_measuring>
   <number_of_runs>10</number_of_runs>
   <max_count_of_fitness>2500</max_count_of_fitness>
   <number_of_parameters>5</number_of_parameters>
   <number_of_experiments>1</number_of_experiments>
   <all_combinations>1</all_combinations>
</about_data>
```

В данном подблоке 13 тэгов, и каждый из них обязателен:

- **<name_algorithm>** — обозначение алгоритма (по названию функции, класса, которая его реализует, например: **MHL_StandartRealGeneticAlgorithm**).

  Например, в автор данного формата свои алгоритмы прописывает в библиотеке https://github.com/Harrix/MathHarrixLibrary, файлы с описанием алгоритмов на https://github.com/Harrix/HarrixOptimizationAlgorithms, где содержатся одноименные файлы с описанием алгоритмов. Если вы исследуете какой-то уже существующий алгоритм, то используйте уже существующий идентификатор, чтобы в дальнейшем можно было сравнивать алгоритмы. Если предлагаете свой алгоритм, то придумайте свой идентификатор (обязательно без пробелов);

- **<full_name_algorithm>** — полное название алгоритма оптимизации;

- **<link_algorithm>** — ссылка на описание алгоритма, где можно прочитать о нем подробно. Если же такого описания нет, или оно не выложено в сети, то в качестве значения в данном тэге должно быть слово «**NULL**», например:

  Код 10. У алгоритма нет ссылки в файле Harrix Optimization Testing

  ```
  <link_algorithm>NULL</link_algorithm>
  ```

- **<name_test_function>** — обозначение тестовой функции (по названию функции, которая ее реализует, например: **MHL_TestFunction_Ackley**).

  Например, в автор данного формата основные тестовые функции прописывает в библиотеке https://github.com/Harrix/MathHarrixLibrary, файл с описанием тестовых функций на https://github.com/Harrix/HarrixTestFunctions, где содержится подробное описание тестовых функций. Если вы исследуете какую-то уже существующую функцию, то используйте уже существующий идентификатор (посмотрите по ссылке выше), чтобы в дальнейшем можно было сравнивать алгоритмы. Если предлагаете свою тестовую функцию, то придумайте свой идентификатор (обязательно без пробелов);

- **<full_name_test_function>** — полное название тестовой функции;

- **<link_test_function>** — ссылка на описание тестовой функции, где можно прочитать о ней подробно. Если же такого описания нет, или оно не выложено в сети, то в качестве значения в данном тэге должно быть слово «**NULL**», например:

Код 11. У алгоритма нет ссылки в файле Harrix Optimization Testing

```
<link_test_function>NULL</link_test_function>
```

- **<dimension_test_function>** — размерность тестовой задачи, то есть это количество входных параметров у тестовой функции. Прошу обратить внимание, что это количество входных переменных тестовой функции, а не размерность объекта, с которым непосредственно работает алгоритм оптимизации (например, стандартный генетический алгоритм работает с бинарными строками, но оптимизирует вещественную тестовую функцию);

- **<number_of_measuring>** — количество измерений для каждого варианта настроек алгоритма (сколько точек получим);

- **<number_of_runs>** — количество запусков алгоритма в каждом из измерений;

- **<max_count_of_fitness>** — максимальное допустимое число вычислений целевой функции (алгоритм может использовать меньше вычислений целевой функции, но не более).

  Используется именно максимальное допустимое число вычислений целевой функции, а не просто число вычислений целевой функции по той причине, что структура разных алгоритмов оптимизации не всегда способна использовать конкретное число вычислений целевой функции;

- **<number_of_parameters>** — количество проверяемых параметров алгоритма оптимизации.

  Если алгоритм оптимизации не имеет настраиваемых параметров, то число параметров ставим равным нулю.

  Если алгоритм имеет переменное число настроек (например, параметр «Размер турнира» в генетическом алгоритме будет появляться только при использовании турнирной селекции), то ставим общее число параметров алгоритма, которое вообще может быть. Например, алгоритм имеет первый параметр «Селекция», который может принимать значения 0 и 1. Если значение равно 0, то алгоритм имеет второй параметр «Полярность», который может принимать какие-то значения. Если значение равно 1, то алгоритм имеет второй параметр «Вязкость», который может принимать какие-то значения. В итоге мы получаем 3 параметра, которое мы и записываем в тэге **<number_of_parameters>**, хотя каждая настройка имеет по два параметра. При этом помните, что потом в записи данных, вы должны записывать параметры под постоянными номерами. И если вы в каком-то измерении записали параметр «Полярность» под номером 2, то другие параметры не могут под таким же номером находиться, даже если в текущей комбинации настроек такого параметра нет, например:

Код 12. Когда в разных комбинациях присутствуют разные настройки Harrix Optimization Testing

```
<data>
   <experiment parameters_of_algorithm_1="Селекция = 0"
      parameters_of_algorithm_2="Полярность = 0.52">
   ...
```

```
    </experiment>

    <experiment parameters_of_algorithm_1="Селекция = 0"
       parameters_of_algorithm_2="Полярность = 0.01">
    ...
    </experiment>

    <experiment parameters_of_algorithm_1="Селекция = 1"
       parameters_of_algorithm_3="Вязкость = 1.3">
    ...
    </experiment>
</data>
```

- **<number_of_experiments>** — количество комбинаций вариантов настроек алгоритма оптимизации, которые были рассмотрены в данном исследовании.

- **<all_combinations>** — все ли возможные комбинации настроек алгоритма оптимизации были просмотрены? Подсчитать возможное количество настроек алгоритма, например, как произведение всех возможных настроек каждого параметра не всегда возможно, так как в некоторых алгоритмах какой-нибудь параметр может быть использован только при каком-то значении другого параметра (например, размер турнира без использования турнирной селекции в генетическом алгоритме не используется). Поэтому предлагается автору исследования самому определиться: или это полное исследование алгоритма или же частичное, где рассмотрено только некотрое неполное подмножество всех возможных настроек (например, автор проводил предварительный анализ работы алгоритма на паре настроек). Какие принимает значения? Если всё множество комбинаций настроек просмотрено, то ставим **1**, иначе **0**.

5.4 Подблок <data>

Код 13. Подблок в файле Harrix Optimization Testing

```
<data>
   <experiment parameters_of_algorithm_1="Тип селекции = Пропорциональная селекция"
      parameters_of_algorithm_2="Тип скрещивания = Одноточечное скрещивание"
      parameters_of_algorithm_3="Тип мутации = Слабая мутация"
      parameters_of_algorithm_4="Тип формирования нового поколения = Только потомки"
      parameters_of_algorithm_5="Тип преобразования задачи вещественной оптимизации в
       задачу бинарной оптимизации = Стандартное представление целого числа - номер у
      зла в сетке дискретизации">
      <measuring>
         <Ex>0.102733</Ex>
         <Ey>1.40394</Ey>
         <R>0</R>
      </measuring>
      <measuring>
         <Ex>0.0840828</Ex>
         <Ey>1.4134</Ey>
         <R>0</R>
      </measuring>
      <measuring>
         <Ex>0.0674963</Ex>
         <Ey>1.20694</Ey>
         <R>0</R>
      </measuring>
      <measuring>
```

```
            <Ex>0.103118</Ex>
            <Ey>1.57915</Ey>
            <R>0</R>
        </measuring>
        <measuring>
            <Ex>0.0795264</Ex>
            <Ey>1.4047</Ey>
            <R>0</R>
        </measuring>
        <measuring>
            <Ex>0.0626839</Ex>
            <Ey>1.17213</Ey>
            <R>0</R>
        </measuring>
        <measuring>
            <Ex>0.0974347</Ex>
            <Ey>1.46336</Ey>
            <R>0</R>
        </measuring>
        <measuring>
            <Ex>0.10858</Ex>
            <Ey>1.26652</Ey>
            <R>0</R>
        </measuring>
        <measuring>
            <Ex>0.0990866</Ex>
            <Ey>1.41937</Ey>
            <R>0</R>
        </measuring>
        <measuring>
            <Ex>0.0901381</Ex>
            <Ey>1.17268</Ey>
            <R>0.1</R>
        </measuring>
    </experiment>
</data>
```

Данный подблок состоит из множества подблоков **<experiment>**, каждый из которых соответствует одному набору настроек алгоритма. Поэтому число блоков **<experiment>** должно совпадает с числом из тэга **<number_of_experiments>**.

Каждый подблок **<experiment>** состоит из множества подблоков **<measuring>**, каждый из которых соответствует одному измерению показателей E_x, E_y, R. Поэтому число подблоков **<measuring>** должно совпадать с числом из тэга **<number_of_measuring>**.

Каждый подблок **<measuring>** содержит по три тэга:

- **Ex** —значение показателя E_x (ошибка по входным параметрам);
- **Ey** —значение показателя E_y (ошибка по значениям целевой функции);
- **R** —значение показателя R (надёжность).

Обратите внимание, что в качестве разделителя в вещественных числах используется точка, а не запятая.

О том, что значат показатели читайте в начале документа.

Также подблок **<experiment>** содержит атрибуты, обозначающие набор параметров алгоритма. Атрибуты построены по следущим правилам:

- каждый параметр алгоритма располагается в своем атрибуте, то есть число атрибутов равно числу настраиваемых параметров алгоритма оптимизации;
- каждый атрибут имеет имя вида: «**parameters_of_algorithm_**» + номер атрибута.
- в каждом блоке **<experiment>** соответствующий параметр алгоритма должен находится в соответствующем атрибуте, а не скакать по атрибутам;
- значение атрибута строится следующим образом: **Имя параметра алгоритма** + « = » + **Значение параметра алгоритма** (например, **parameters_of_algorithm_1="Тип селекции = Пропорциональная селекция"**);
- если параметр алгоритма в каких-то настройках отсутствует, то всё равно заполняем, например, словом «**Отсутствует**» (например, **parameters_of_algorithm_3="Дополнительный усилитель = Отсутствует"**);
- Если алгоритм не содержит параметров, то атрибутов у подблока **<experiment>** не будет.

Вот и всё. Описание самой структуры закончено.

6 Функции, которые обрабатывают данный формат файлов

В библиотеке https://github.com/Harrix/DataOfHarrixOptimizationTesting на языке C++ имеется класс **DataOfHarrixOptimizationTesting**, который парсит и анализирует данный формат файлов с среде Qt.