Guide

1 About the program

This program provides a subpopulation-based evolutionary algorithm to take advantage of parallel architectures involving a multicore CPU and multiple GPUs for accelerating an electroencephalogram (EEG) feature selection problem. The procedure has been implemented using two scheduling alternatives for evaluation of individuals according to the number of subpopulations (one or more). It uses OpenMP to distribute dynamically either subpopulations or individuals among devices and OpenCL to evaluate the individuals taking into account the devices characteristics, providing two parallelism levels in CPU and up to three levels in each GPU.

2 Program compilation and use of parameters

There is a *Makefile* file to build the project. Running the following order in a Unix shell the program should be compiled:

```
Make -j N_FEATURES=NF
```

Where *NF* is the number of features to use (columns) of the database. This number must be between 4 and the total number of features of the database. The executable file, named *hpmoon* will be generated in the "*bin*" folder. For running it, in the shell the next order must be executed:

```
./bin/hpmoon –conf config.xml
```

Where *config.xml* is the necessary configuration file for the correct performance of the program, specified by the *–conf* option and located in the root folder of the project.

In addition, the user can indicate separately through line arguments each setting of the XML file. Table 1 summarizes the list of parameters and their possible values, and how to use them in the line of arguments. In any case, the special option *-h* displays the available options and examples of use.

Also, the *Makefile* file contains a rule to generate *Doxygen* documentation in the *"doc/html"* folder. This can be done by running the following command:

```
make documentation
```

On the other hand, it is possible to build only the Fonseca library *fpli_hv.a* that contains the function which calculates the Hypervolume indicator:

```
make fonseca
```

Finally, the files and documents generated when compiling the project can be deleted. There are two types of cleaning depending on the content to be deleted. The command:

```
make clean
```

Deletes the following contents:

- **Binary files.**
- **.o files.**
- **~ files.**
- **Hypervolume project of Fonseca.** This project generates the necessary library for calculating the hypervolume.

For a complete cleaning, run the following command:

```
make eraseAll
```

Which will remove the same content as the previous command and also the following content:

- ***gnuplot* files.**
- **Documentation files** generated by *Doxygen*.

gnuplot files contain the fitness of the individuals in the first Pareto front and the necessary source code for the *gnuplot* program. If the user would generate a graph using *gnuplot* and the source code generated by the program, it will also be deleted when using this command.

3 The XML configuration file

The XML configuration file is required to run the program as well as the parameters used in the *make* command. The parameters of the XML file are read and used at runtime while the parameters used in the *make* command are read and used at compile time to avoid dynamic memory. The parameters are:

- ***NSubpopulations*** is the total number of subpopulations (only for islands-based model).
- ***SubpopulationSize*** is the number of individuals of the subpopulation.
- ***NInstances*** is the maximum number of instances (rows) to be taken from the database.

- ***NExecutions*** is the number of executions of the complete genetic algorithm. The measures of Hypervolume and times are the average obtained from all executions. This is useful for benchmark.
- ***DataBaseFileName*** is the name of the file containing the database.
- ***NMigrations*** is the number of individuals migrations between subpopulations.
- ***NGenerations*** is the number of evolves of a subpopulation (generations of individuals).
- ***MaxFeatures*** is the maximum number of features initially set to "1".
- ***TournamentSize*** is the number of individuals competing in the tournament.
- ***DataFileName*** is the name of the file which will contain the fitness of the individuals in the first Pareto front.
- ***PlotFileName*** is the name of the file which will contain the *gnuplot* code for data display.
- ***ImageFileName*** is the name of the file which will contain the image data (graphic) after using the *gnuplot* command to generate it.
- ***NDevices*** is the number of *OpenCL* devices that will run the program. Set to "0" to run in sequential mode.
- ***Devices*** specify the names of the *OpenCL* devices that will run the program. The values must be separated by commas.
- ***ComputeUnits*** specify the compute units for each *OpenCL* device that will run the program. The values must be separated by commas too and in the same order than their corresponding devices.
- ***WiLocal*** specify the number of work-items (threads) per compute unit for each *OpenCL* device that will run the program. The values must be separated by commas too and in the same order than their corresponding devices.
- ***MaxIndividualsOnGpuKernel*** is the maximum number of individuals to be processed in a single execution of the kernel. This is only applicable for GPU kernels.
- ***KernelsFileName*** is the name of the file containing the kernels with the *OpenCL* code.

The following table summarizes the restrictions of input parameters. The parameters passed to the *make* command are shown in uppercase. In lowercase, the parameters found in the XML configuration file.

PARAMETER	**RANGE**	**OPTION**
N_FEATURES	4 <= ***NF*** <= Number of features of the DB	-
NSubpopulations	1 <= ***NP***	-ns
SubpopulationSize	4 <= ***PS***	-ss
NInstances	4 <= ***NI*** <= Number of instances	-ni

	of the DB	
NExecutions	1 <= ***NE***	-nexec
DataBaseFileName	-	-db
NMigrations	1 <= ***NM***	-nm
NGenerations	0 <= ***NG***	-g
MaxFeatures	1 <= ***MaxF***	-maxf
TournamentSize	2 <= ***TS***	-ts
DataFileName	-	-plotdata
PlotFileName	-	-plotsrc
ImageFileName	-	-plotimg
NDevices	0 <= ***ND***	-nd
Devices	-	-devn
ComputeUnits	1 <= ***CU***	-cu
WiLocal	1 <= ***WL*** <= Maximum number of local work-item of the device	-wl
MaxIndividualsOnGpuKernel	1 <= ***MaxIndGPUKernel***	-maxind
KernelsFileName	-	-ke

Table 1. It shows the range of values of the input parameters and how to use them from the arguments line.

4 OpenCL optimization and limitations. OpenMP use.

The following points should be considered to obtain good performance when running the program:

1) The evaluation function for each individual has been parallelized with *OpenCL*. A compute unit is formed by *WiLocal* work-items and evaluates only one individual. Therefore, in *GPUs*, *WiLocal* should be a multiple of 32 or 64 according to the device for improve the performance. The user can approximate the optimal value of *WiLocal* and *ComputeUnits*. The value is calculated as the number of stream processor or *CUDA* cores divided by the number of compute units. For example, the *Nvidia GeForce GTX 770* has 1536 *CUDA* cores and 8 compute units, so 1536/8 = 192 local work-items, but sometimes it is better to increase this value, for example, 256, 512 or 1024 according to special cases. In the case of 256 work-items, *WiLocal* = 256 and *ComputeUnits* = 8, comprising in total of 256 * 8 = 2048 work-items. The best combination is determined by the characteristics of the problem.
   In *CPUs*, *ComputeUnits* should have a value equal to the number of compute units (physical cores) and *WiLocal* must be set to "1". If another value is specified, it will be ignored.

2) The sort function according to the Pareto front, *nonDominationSort* contains one loop of quadratic order and is related to the number of individuals. For good quality results it is not necessary to increase the number of individuals too. It's better to increase the number of iterations of the program (number of generations).
3) On *GPU*, the program gets better performance with values of *N_FEATURES* and *NInstances* higher than the number of local work-items. However, the database is stored in local memory and their capacity is very limited (approximately 49 KB depending on the device). So the program will abort if the database is too big.
4) The evaluation of all individuals is executed in a single kernel in the case of the *CPUs*. However, on *Nvidia GPUs*, a kernel has a certain time limit for its execution. If the number of individuals is high, the execution will fail. The *MaxIndividualsOnGpuKernel* parameter allows to indicate the quantity of individuals which will be evaluated in a single execution of the kernel. Just reduce this parameter if the execution fails for this reason. The kernel is called iteratively to evaluate all individuals.
5) If multiple devices are specified, the evaluation of the individuals is distributed dynamically among the *OpenCL* devices using *OpenMP*. Each *OpenMP* thread handles one device. This way, each device is independent and compute chunks of individuals equals to its number of compute units until all individuals are evaluated.