User's manual of Self-organizing clustering

Kou AMANO
RIKEN BioResource Center

September 24, 2010

1 Outline

Self-organizing clustering (SOC) is a vector clustering machine which is based on k-means. It provides functions of: 1. conversion from DNA sequences to vectors, 2. initialization of cluster configuration, 3. clustering with an improved learning process. These three functions are implemented as three commands (`fasta2matrix`, `soc-init` and `soc-lm`).

2 Install

First, you have to obtain the tar ball of SOC. To obtain the source code, please contact `<kamano@affrc.go.jp>` or `<amano@brc.riken.jp>` . To install SOC, please execute the following steps:

```
[user]$ tar -zxvf SOC_0.5.3-00.tar.gz
[user]$ cd SOC_0.5.3-00
[user]$ ./configure.pl
[user]$ make
[user]$ su
[root]# mkdir -p /home/pub/bin
[root]# make install
[root]# exit
[user]$ PATH=$PATH:/home/pub/bin
[user]$ export PATH
```

3 Commands

The latest version of SOC is 0.5.9-00.

3.1 fasta2matrix

3.1.1 Description

The `fasta2matrix` command execute conversion from a set of multiple DNA sequence data to a float matrix data. The `fasta2matrix` needs administrative information such as memory size. Users have to specify memory sizes: `<number of samples>` $\times$ (`<header size>` + `<body size>`) .

3.1.2 Usage

```
fasta2matrix [if=<file name>] [s=<segment size>] [g=<step size>]
   [n=<number of samples>] [hs=<header (comment) size>]
   [bs=<body (sequence) size>] [--help|-h] [--check|-c] [-H|+H]
   [-m|+m] [-p<total frequency of oligonucleotides for normalization>|+p]
```

Options:

- `<file name>`
  To specify the input file name.
  Default value: None.
- `<segment size>`
  To specify the length of oligonucleotides.
  Default value: 2.
- `<step size>`
  To specify the window slide size.
  Default value: 1.
- `<number of samples>`
  For memory allocation.
  The `fasta2matrix` stores all of multiple fasta sequence data on memory.
  Default value: 15000.
- `<header size>`
  For memory allocation.
  Default value: 128.
- `<body size>`
  For memory allocation.
  Default value: 4000.
- `[--check|-c]`
  To print command arguments.
  Default value: None.
- `[-H|+H]`
  Where '-H' is specified, the program prints headers to the output file.
  Default value: '-H'.

- `[-m|+m]`
  Where '-m' is specified, the program prints matrix size to the output file.
  Default value: '-m'.

- `<total frequency of oligonucleotides for normalization>`
  If users need the mormalized oligonucleotide frequencies, the users can specify '-p' option.
  Default value: None.

3.1.3 Examples

To print an oligonucleotide matrix with 100×64 of size from file `test.fasta`:
```
fasta2matrix if=test.fasta s=3 n=100
```

To print an oligonucleotide matrix with 100×64 of size without header from file `test.fasta`:
```
fasta2matrix if=test.fasta s=3 n=100 +H
```

To print an oligonucleotide matrix with 100×64 of size without matrix size from file `test.fasta`:
```
fasta2matrix if=test.fasta s=3 n=100 +m
```

3.2 soc-init

3.2.1 Description

The `soc-init` provides several primitive types of initial cluster configuration.

3.2.2 Usage

```
soc-init if=<sample data file> of=<output data file>
   [frac=<number of cluster nodes>] [<FORMAT>]
```

Options:

- `<sample data file>`
  To specify the sample file (input data) which is generated by `fasta2matrix`.

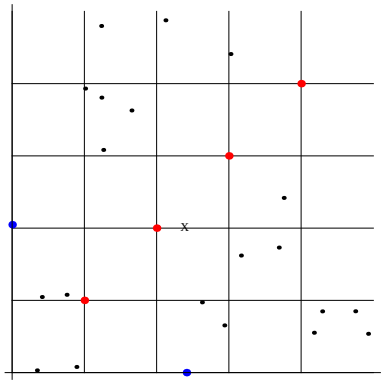
- `<output data file>`
  To specify the file of cluster-nodes data to output.

- `<number of cluster nodes>`
  To specify the number of initial cluster-nodes.

- `<FORMAT>`
  See below.

`<FORMAT>`:

- `Diagonal`
  To array cluster-nodes on a diagonal line across the sample vector space.
- `node=Central`
  To place cluster-nodes at the sample-nodes near to the centroid.
- `Grid=<tensor size>`
  To array cluster-nodes on the lattice points in the sample vector space.
- `Axis-mean=<axis1,axis2,...>`
  To place cluster-nodes on coordinate axes of the sample vector space.



Where `Diagonal` option is specified, the cluster-nodes are arrayed as red dots.
Where `node=Central` option is specified, the cluster-nodes are placed on sample-nodes nearest to centroid (x).
Where `Grid` option is specified, the cluster-nodes are placed on the user-specified lattice points .
Where `Axis-mean` option is specified, the cluster-nodes are placed as blue dots.

Fig: Initialization patterns of cluster-nodes

3.2.3 Examples

To locate 10 cluster-nodes around the centroid and write the coordinates to the file "test.fasta.mat.cl":

```
soc-init if=test.fasta.mat of=test.fasta.mat.cl frac=10 node=Central
```

To locate 16 (4×4) cluster-nodes on the lattice points and write the coordinates to the file "test.fasta.mat.cl":

```
soc-init if=test.fasta.mat of=test.fasta.mat.cl Grid=4,4
```

3.3 `soc-lm`

3.3.1 Description

The `soc-lm` learns to move cluster-nodes, which represent clusters, toward the centroids of clusters. It also dynamically generates/merges the cluster-nodes in the learning stage.

3.3.2 Usage

```
soc-lm sf=<sample file> cf=<cluster file> [loop=<loop>]
   [rate=<learning rate>] [unify=<n1>,<n2>,<n3>,<n4>]
```

[generate=<n1>,<n2>,<n3>,<n4>] [cutoff=<n1>,<n2>,<n3>]
[cutoff-gen=<n1>,<n2>,<n3>,<n4>] [delete=<n1>,<n2>,<n3>]
[lf=<log file>] [of=<output file>] [of_type=<output file type>]
[rf=<report file>] [dist_func=<distance function>]

Options:

- <sample file>
  To specify the sample file (input data) which is generated by fasta2matrix).
- <cluster file>
  To specify the cluster file (input data) which is generated by soc-init.
- <loop>
  To specify learning iterations.
- <learning rate>
  To specify the ratio of moving distance to remaining distance. Let the distance between given cluster-node and given centroid is l, and the learning rate is $r(0 <= r <= 1)$, the remaining distance is $l(1 - r)$.
- unify=<n1>,<n2>,<n3>,<n4>
  To specify the thresholds for unifying cluster-nodes.
  IF n1 is specified AND distance between given cluster-node pair is lower than n2 AND number of clusters is higher than n3, THEN the cluster-node pair whose distance is shortest is unified with interval n4.
- generate=<n1>,<n2>,<n3>,<n4>
  To specify the thresholds for generating cluster-node.
  IF n1 is specified AND cluster-radius is higher than n2 AND number of clusters is lower than n3, THEN the cluster-nodes whose radius is longest is divided with interval n4.
- delete=<n1>,<n2>,<n3>
  To specify the threshold to delete cluster-node(s).
  IF n1 is specified AND distances between cluster-node pairs is higher than n2, THEN ones of the pairs are deleted until all distances of cluster-node pair is shorter than n2 with interval n3.
  -n1 がセットされており、かつ、n2 を上回るクラスタノード間の距離がある場合、n2 を上回る距離をもつクラスタノードペアがなくなるまで、当該のクラスタノードペアの片方を delete する。-
- <log file>
  To specify log file.
- <output file>
  To specify output file.
- <output file type>
  To specify output file type.

- `<report file>`
  To specify another (reduced) output file.

- `<distance function>`
  To specify distance function ($\cos\theta$ or euclidean).

3.3.3 Examples

To classify sample data (`test.fasta.mat`) with default of option:
`soc-lm sf=test.fasta.mat cf=t.fasta.mat.cl`

To classify sample data (`test.fasta.mat`) under the condition - loop is 100, cluster deletion is on, cluster deletion threshold is 4 and learning rate is 0.9:
`soc-lm sf=test.fasta.mat cf=test.fasta.mat.cl delete=1,4,2 rate=0.9`