Calculating Feature Vector from SDF file

September 1, 2020

Contents

1 Quick start

Module 1 calculates the feature vector (FV) for given chemical compound(s). One or more compounds can be given by a standard SDF file, and feature vectors are outputted in an original FV format. These formats are covered by the following sections. Here we show how to use the program.

- *Confirm the development environment*
  Any compiler compatible with ISO C++ 2011 standard should work. g++ ver 5 & 7 on Linux Mint 18 & 19 have been tested, which can be installed from command line by the next command.
  `$ sudo apt install g++`

- *Compile*
  `$ g++ -std=c++11 -o fv4_bc_in_ex fv4_in_ex.cpp`
  (The option `-std=c++11` can be omitted for g++ ver 7.)

- *Run*
  `$ ./fv4_in_ex input.sdf output.csv`
  The first argument `input.sdf` specifies the input SDF file, and the second argument `output.csv` specifies the output FV file. Both are text files. For example:
  `$ ./fv4_in_ex sample1.sdf sample1.csv`

2 SDF Format

This program uses SDF (Structure Data File), a standard format, for input. See, e.g., `https://www.chem-station.com/blog/2012/04/sdf.html` for an explanation in Japanese or the official definition (in English) `http://help.accelrysonline.com/ulm/onelab/1.0/content/ulm_pdfs/direct/reference/ctfileformats2016.pdf` for detail.

As an example, let us have a look at the attached sample1.sdf (see `https://pubchem.ncbi.nlm.nih.gov/compound/128703` for detail of this compound).

A sample input in SDF format: sample1.sdf

```
128703
  -OEChem-02061913062D

 24 23  0     1  0  0  0  0  0999 V2000
    6.0010   -1.2500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    5.1350    1.2500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    6.8671    1.2500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    4.2690   -1.2500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    7.7331   -1.2500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    2.5369   -0.2500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    8.5991    0.2500    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   12.9292   -0.2500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
```

```
   12.0632    0.2500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   13.7953    0.2500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   11.1972   -0.2500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   14.6613   -0.2500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   10.3312    0.2500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    6.0010   -0.2500    0.0000 C   0  0  1  0  0  0  0  0  0  0  0  0
   15.5273    0.2500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    5.1350    0.2500    0.0000 C   0  0  1  0  0  0  0  0  0  0  0  0
    6.8671    0.2500    0.0000 C   0  0  2  0  0  0  0  0  0  0  0  0
    9.4651   -0.2500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   16.3933   -0.2500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    4.2690   -0.2500    0.0000 C   0  0  1  0  0  0  0  0  0  0  0  0
    7.7331   -0.2500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   17.2594    0.2500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.4030    0.2500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   18.1254   -0.2500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
 14  1  1  1  0  0  0
 16  2  1  1  0  0  0
 17  3  1  1  0  0  0
 20  4  1  6  0  0  0
  5 21  2  0  0  0  0
  6 23  1  0  0  0  0
  7 18  1  0  0  0  0
  7 21  1  0  0  0  0
  8  9  1  0  0  0  0
  8 10  1  0  0  0  0
  9 11  1  0  0  0  0
 10 12  1  0  0  0  0
 11 13  1  0  0  0  0
 12 15  1  0  0  0  0
 13 18  1  0  0  0  0
 14 16  1  0  0  0  0
 14 17  1  0  0  0  0
 15 19  1  0  0  0  0
 16 20  1  0  0  0  0
 17 21  1  0  0  0  0
 19 22  1  0  0  0  0
 20 23  1  0  0  0  0
 22 24  1  0  0  0  0
M  END
```

3 FV format

The output is in an original FV (Feature Vector) format, which is just a CSV file so that can be opened by Excel and other spreadsheet software. The first line shows the components of FV

and the following lines show the values for those components of FV.

For example, let us have a look at the FV file sample1.csv for sample1.sdf.

Example of a FV file: sample1.csv (Note: \\ shows that there is NO line break.)

```
CID,n,M,C_in,C_ex,S_in,S_ex,N_in,N_ex,H,S1S_in,S1S_ex,C1S_in,C1S_ex,\\
C2S_in,C2S_ex,C1N_in,C1N_ex,#degree1_in,#degree1_ex,#degree2_in,\\
#degree2_ex,#degree3_in,#degree3_ex,#degree4_in,#degree4_ex,\\
#double_bond_in,#double_bond_ex,#triple_bond_in,#triple_bond_ex,\\
Diameter,Bc_121_in,Bc_121_ex,Bc_122_in,Bc_122_ex,Bc_123_in,Bc_123_ex,\\
Bc_131_in,Bc_131_ex,Bc_132_in,Bc_132_ex,Bc_141_in,Bc_141_ex,\\
Bc_221_in,Bc_221_ex,Bc_222_in,Bc_222_ex,Bc_223_in,Bc_223_ex,\\
Bc_231_in,Bc_231_ex,Bc_232_in,Bc_232_ex,Bc_241_in,Bc_241_ex,\\
Bc_331_in,Bc_331_ex,Bc_332_in,Bc_332_ex,Bc_341_in,Bc_341_ex,\\
Bc_441_in,Bc_441_ex,2-branch_height,2-branch_leaf_number
128703,24,130.833,14,3,0,6,1,0,0,0,5,0,1,2,0,12,3,0,7,10,2,5,0,0,0,0,1,0,0,\\
0.75,0,2,0,0,0,0,0,4,0,1,0,0,9,1,0,0,0,0,1,1,0,0,0,0,4,0,0,0,0,0,0,0,1,2
```

Detail of the components

- `CID`
  CID in PubChem (`https://pubchem.ncbi.nlm.nih.gov/`). For example, the compound in sample1.sdf is `https://pubchem.ncbi.nlm.nih.gov/compound/128703`.
- `n`
  Number of atoms except for the hydrogen
- `M`
  Average molecular mass defined by $M = \frac{1}{n}\sum_{\mathtt{a}} \lfloor 10 \cdot \mathrm{mass}(\mathtt{a}) \rfloor$
- `C_in,O_in,N_in`
  Numbers of internal atoms
- `C_ex,O_ex,N_ex`
  Numbers of external atoms
- `H`
  Numbers of atoms
- `C1O_in, C2O_in, C1N_in, C1C_in`
  Numbers of internal paths. For example, `C1O_in` shows the number of internal single bonds by `C` and `O`; `C2O_in` shows the number of internal double bonds by `C` and `O`.
- `C1O_ex, C2O_ex, C1N_ex, C1C_ex`
  Numbers of external paths. For example, `C1O_ex` shows the number of external single bonds by `C` and `O`; `C2O_ex` shows the number of external double bonds by `C` and `O`.
- `#degree1_in,#degree2_in,#degree3_in,#degree4_in`
  Numbers of internal atoms with the given degrees (valences)

- `#degree1_ex,#degree2_ex,#degree3_ex,#degree4_ex`
  Numbers of external atoms with the given degrees (valences)
- `#double_bond_in,#triple_bond_in`
  Numbers of internal double bonds and internal triple bonds
- `#double_bond_ex,#triple_bond_ex`
  Numbers of external double bonds and external triple bonds
- `Diameter`
  Diameter of the graph divided by n
- `Bc_xyz_in`
  Internal degree configuration (x, y, z), where $x \leq y$ are the degrees of the end-vertices of a internal bond, and z is its multiplicity.
- `Bc_xyz_ex`
  External degree configuration (x, y, z), where $x \leq y$ are the degrees of the end-vertices of a external bond, and z is its multiplicity.
- `2-branch_height`
  2-branch-height bh_2: see the paper for more information on this decriptor.
- `2-branch_leaf_number`
  2-branch-leaf-number bl_2 see the for more information on this decriptor.

4 Program notes

The program `fv4_in_ex` calculates and outputs FV for a given SDF file. Here some notes on using it are provided.

1. The mass of each atom is hard-coded in the program. At the point of writing this document, there are 15 atoms in total. If they are not enough, please add the new atoms in the same manner.

Mass data of atoms used in the program (`function init_MassMap()`)

```
M["B"]  = 108;
M["C"]  = 120;
M["O"]  = 160;
M["N"]  = 140;
M["F"]  = 190;
M["Si"] = 280;
M["P"]  = 310;
M["S"]  = 320;
M["Cl"] = 355;
M["V"]  = 510;
M["Br"] = 800;
M["Cd"] = 1124;
```

```
M["I"]  = 1270;
M["Hg"] = 2006;
M["Pb"] = 2072;
M["Al"] = 269;
```

2. Change `bool debug` to `true` in line 27 of the source file `fv4_in_ex.cpp` to enable verbose output.