



Effective early disease risk assessment with matrix factorization on a large-scale medical database [↗](#)

PLOS One

Chu-Yu Chin¹, Sun-Yuan Hsieh¹, Vincent S. Tseng²

¹National Cheng Kung University, ²National Chiao Tung University

[dx.doi.org/10.17504/protocols.io.rv2d68e](https://doi.org/10.17504/protocols.io.rv2d68e)

Chu-Yu Chin

ABSTRACT

The early assessment of disease risk is an emerging topic in medical informatics. If diseases are detected at an early stage, prognosis can be improved and medical resources can be used more efficiently. A number of recent studies have considered risk factor analysis approaches, such as association rule mining, sequential rule mining, regression, and medical expert advice.

In this study, for improving disease risk assessment, non-negative matrix factorization and support vector machine (SVM) were integrated to discover important and implicit risk factors.

To make the method easy to follow, here we provide an experimental protocol. This experimental protocol comprises three main stages: data preprocessing, risk factor optimization, and early disease risk assessment. To discover the optimized risk factors, the NMF algorithm with parameter optimization was used for constructing the NMF-based matrix. In the assessment model learning and early disease risk assessment stages, the machine learning classifier SVM was used for disease modeling with the NMF-based matrix, yielding the final disease risk assessment, which serves as an excellent reference for physicians and patients.

EXTERNAL LINK

<https://doi.org/10.1371/journal.pone.0207579>

THIS PROTOCOL ACCOMPANIES THE FOLLOWING PUBLICATION

Chin C, Hsieh S, Tseng VS (2018) *eDRAM*: Effective early disease risk assessment with matrix factorization on a large-scale medical database: A case study on rheumatoid arthritis. PLoS ONE 13(11): e0207579. doi: [10.1371/journal.pone.0207579](https://doi.org/10.1371/journal.pone.0207579)

PROTOCOL STATUS

Working

Install Prerequisite software and libraries

1 This protocol requires:

- Matlab 2016a
- LibSVM version 3.22

Each instruction can be verified on property websites.

The following steps include: dataset preparation, executing NMF multiplicative update algorithm, parameter settings for SVM and effectiveness evaluation on the SVM-based disease risk assessment.

Prepare and load a Patient-Diagnosis Disease Matrix by the LibSVM format with label

2 The medical diagnostic record data to be analyzed is converted into a patient-diagnosis disease matrix in advance.

Given an $N \times M$ matrix, each row represents the medical history of a diagnosed patient across all diseases or symptoms (DS). Each column indicates the diagnostic record status of all patients for a single DS.

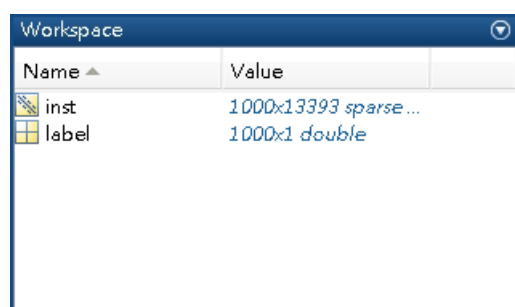
Running the command below in the Matlab to load patient-diagnosis disease matrix by the libSVM Format.

```
>> [label, inst] = libsvmread('C:\SVM_data\200.libsvm.txt');
```

An example of patient-diagnosis disease matrix with the LibSVM format.

EXPECTED RESULT

The instance data matrix and label will be show in the Matlab workspace.



Executing the NMF multiplicative update algorithm

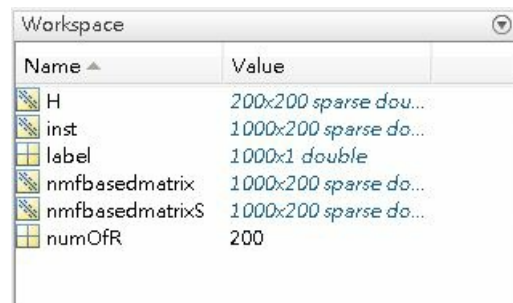
- The step comprises three phases. First, a number of dimension should be set. The appropriate values can be selected based on the training data by performing step 3-5. Second, to generate the NMF-Based matrix, the NMF multiplicative update algorithm is performed. Third, the NMF-Based matrix is stored in the LibSVM file format.

```
>> numOfR=200 >> rng(1) >> [nmfbasedmatrix,H] = nnmf(inst, numOfR, 'algorithm','mult'); >> %Create sparse matrix >>
```

```
nmfbasedmatrixS = sparse(nmfbasedmatrix) >> %write the matrix to the file by the libSVM format. >>
libsvmwrite('C:\NMF_Based_matrix\200.libsvm',label,nmfbasedmatrixS);
```

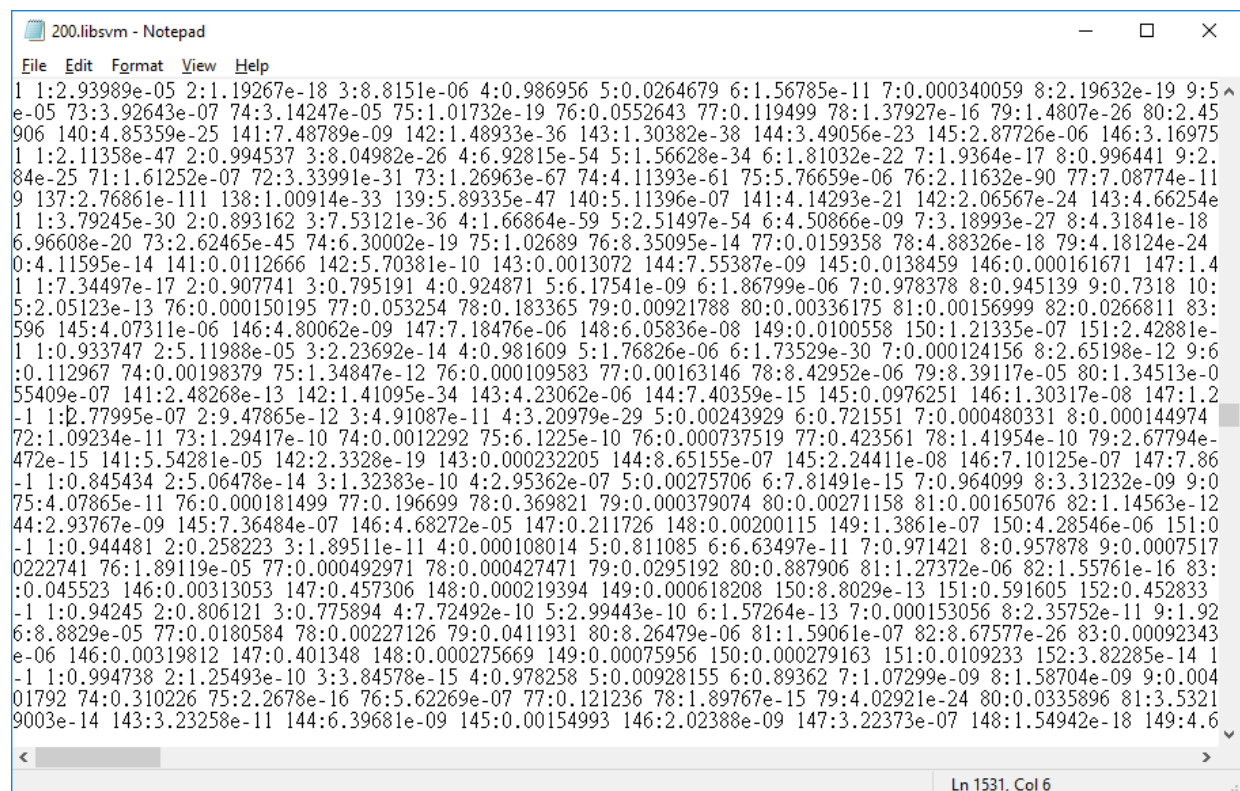
EXPECTED RESULT

A new matrix called the NMF-based matrix has been generated and displayed in the matlab workspace.



Name	Value
H	200x200 sparse dou...
inst	1000x200 sparse do...
label	1000x1 double
nmfbasedmatrix	1000x200 sparse do...
nmfbasedmatrixS	1000x200 sparse do...
numOfr	200

The NMF-based matrix is stored in the SVM file format.



```
200.libsvm - Notepad
File Edit Format View Help
1 1:2.93989e-05 2:1.19267e-18 3:8.8151e-06 4:0.986956 5:0.0264679 6:1.56785e-11 7:0.000340059 8:2.19632e-19 9:5
e-05 73:3.92643e-07 74:3.14247e-05 75:1.01732e-19 76:0.0552643 77:0.119499 78:1.37927e-16 79:1.4807e-26 80:2.45
906 140:4.85359e-25 141:7.48789e-09 142:1.48933e-36 143:1.30382e-38 144:3.49056e-23 145:2.87726e-06 146:3.16975
1 1:2.11358e-47 2:0.994537 3:8.04982e-26 4:6.92815e-54 5:1.56628e-34 6:1.81032e-22 7:1.9364e-17 8:0.996441 9:2.
84e-25 71:1.61252e-07 72:3.33991e-31 73:1.26963e-67 74:4.11393e-61 75:5.76659e-06 76:2.11632e-90 77:7.08774e-11
9 137:2.76861e-111 138:1.00914e-33 139:5.89335e-47 140:5.11396e-07 141:4.14293e-21 142:2.06567e-24 143:4.66254e
1 1:3.79245e-30 2:0.893162 3:7.53121e-36 4:1.66864e-59 5:2.51497e-54 6:4.50866e-09 7:3.18993e-27 8:4.31841e-18
6.96608e-20 73:2.62465e-45 74:6.30002e-19 75:1.02689 76:8.35095e-14 77:0.0159358 78:4.88326e-18 79:4.18124e-24
0:4.11595e-14 141:0.0112666 142:5.70381e-10 143:0.0013072 144:7.55387e-09 145:0.0138459 146:0.000161671 147:1.4
1 1:7.34497e-17 2:0.907741 3:0.795191 4:0.924871 5:6.17541e-09 6:1.86799e-06 7:0.978378 8:0.945139 9:0.7318 10:
5:2.05123e-13 76:0.000150195 77:0.053254 78:0.183365 79:0.00921788 80:0.00336175 81:0.00156999 82:0.0266811 83:
596 145:4.07311e-06 146:4.80062e-09 147:7.18476e-06 148:6.05836e-08 149:0.0100558 150:1.21335e-07 151:2.42881e-
1 1:0.933747 2:5.11988e-05 3:2.23692e-14 4:0.981609 5:1.76826e-06 6:1.73529e-30 7:0.000124156 8:2.65198e-12 9:6
:0.112967 74:0.00198379 75:1.34847e-12 76:0.000109583 77:0.00163146 78:8.42952e-06 79:8.39117e-05 80:1.34513e-0
55409e-07 141:2.48268e-13 142:1.41095e-34 143:4.23062e-06 144:7.40359e-15 145:0.0976251 146:1.30317e-08 147:1.2
-1 1:2.77995e-07 2:9.47865e-12 3:4.91087e-11 4:3.20979e-29 5:0.00243929 6:0.721551 7:0.000480331 8:0.000144974
72:1.09234e-11 73:1.29417e-10 74:0.0012292 75:6.1225e-10 76:0.000737519 77:0.423561 78:1.41954e-10 79:2.67794e-
472e-15 141:5.54281e-05 142:2.3328e-19 143:0.000232205 144:8.65155e-07 145:2.24411e-08 146:7.10125e-07 147:7.86
-1 1:0.845434 2:5.06478e-14 3:1.32383e-10 4:2.95362e-07 5:0.00275706 6:7.81491e-15 7:0.964099 8:3.31232e-09 9:0
75:4.07865e-11 76:0.000181499 77:0.196699 78:0.369821 79:0.000379074 80:0.00271158 81:0.00165076 82:1.14563e-12
44:2.93767e-09 145:7.36484e-07 146:4.68272e-05 147:0.211726 148:0.000618208 149:0.00200115 149:1.3861e-07 150:4.28546e-06 151:0
-1 1:0.944481 2:0.258223 3:1.89511e-11 4:0.000108014 5:0.811085 6:6.63497e-11 7:0.971421 8:0.957878 9:0.0007517
0222741 76:1.89119e-05 77:0.000492971 78:0.000427471 79:0.0295192 80:0.887906 81:1.27372e-06 82:1.55761e-16 83:
:0.045523 146:0.00313053 147:0.457306 148:0.000219394 149:0.000618208 150:8.8029e-13 151:0.591605 152:0.452833
-1 1:0.94245 2:0.806121 3:0.775894 4:7.72492e-10 5:2.99443e-10 6:1.57264e-13 7:0.000153056 8:2.35752e-11 9:1.92
6:8.8829e-05 77:0.0180584 78:0.00227126 79:0.0411931 80:8.26479e-06 81:1.59061e-07 82:8.67577e-26 83:0.00092343
e-06 146:0.00319812 147:0.401348 148:0.000275669 149:0.00075956 150:0.000279163 151:0.0109233 152:3.82285e-14 1
-1 1:0.994738 2:1.25493e-10 3:3.84578e-15 4:0.978258 5:0.00928155 6:0.89362 7:1.07299e-09 8:1.58704e-09 9:0.004
01792 74:0.310226 75:2.2678e-16 76:5.62269e-07 77:0.121236 78:1.89767e-15 79:4.02921e-24 80:0.0335896 81:3.5321
9003e-14 143:3.23258e-11 144:6.39681e-09 145:0.00154993 146:2.02388e-09 147:3.22373e-07 148:1.54942e-18 149:4.6
```

An example of the NMF-based matrix with the LibSVM format.

Obtaining the SVM parameters by using the grid search method with stratified ten-fold cross-validation.

- In order to obtain more suitable SVM parameters, based on stratified ten-fold cross-validation, the grid search method that derived by the LibSVM is utilized.

```
c:\>python C:\libsvm-3.22\tools\grid.py C:\NMF_Based_matrix\200.libsvm -v 10
```

EXPECTED RESULT

In this example, the suitable parameters values of C and g are 2 and 0.03125, respectively.

```
[local] 1 -15 46.4 (best c=2.0, g=0.03125, rate=76.5)
[local] 1 3 49.4 (best c=2.0, g=0.03125, rate=76.5)
[local] 1 -9 72.5 (best c=2.0, g=0.03125, rate=76.5)
[local] 5 -3 71.3 (best c=2.0, g=0.03125, rate=76.5)
[local] -1 -3 73.4 (best c=2.0, g=0.03125, rate=76.5)
[local] 11 -3 66.0 (best c=2.0, g=0.03125, rate=76.5)
[local] -3 -3 66.0 (best c=2.0, g=0.03125, rate=76.5)
[local] 9 -3 66.8 (best c=2.0, g=0.03125, rate=76.5)
[local] 3 -3 72.1 (best c=2.0, g=0.03125, rate=76.5)
[local] 15 -3 63.5 (best c=2.0, g=0.03125, rate=76.5)
[local] -5 -3 60.0 (best c=2.0, g=0.03125, rate=76.5)
[local] 7 -3 70.0 (best c=2.0, g=0.03125, rate=76.5)
[local] 1 -3 73.9 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 -7 67.6 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 -1 61.0 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 -13 73.7 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 1 51.9 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 -11 73.6 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 -5 65.7 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 -15 72.5 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 3 49.8 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 -9 71.2 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 -3 63.5 (best c=2.0, g=0.03125, rate=76.5)
2.0 0.03125 76.5
```

```
[local] 5 -7 74.2 (best c=32.0, g=0.0078125, rate=74.2)
[local] -1 -7 71.9 (best c=32.0, g=0.0078125, rate=74.2)
[local] 5 -1 62.9 (best c=32.0, g=0.0078125, rate=74.2)
[local] -1 -1 57.5 (best c=32.0, g=0.0078125, rate=74.2)
[local] 11 -7 69.2 (best c=32.0, g=0.0078125, rate=74.2)
[local] 11 -1 61.0 (best c=32.0, g=0.0078125, rate=74.2)
[local] 5 -13 72.4 (best c=32.0, g=0.0078125, rate=74.2)
[local] -1 -13 46.4 (best c=32.0, g=0.0078125, rate=74.2)
[local] 11 -13 72.7 (best c=32.0, g=0.0078125, rate=74.2)
[local] -3 -7 57.2 (best c=32.0, g=0.0078125, rate=74.2)
[local] -3 -1 54.9 (best c=32.0, g=0.0078125, rate=74.2)
[local] -3 -13 46.4 (best c=32.0, g=0.0078125, rate=74.2)
[local] 5 1 54.3 (best c=32.0, g=0.0078125, rate=74.2)
[local] -1 1 50.3 (best c=32.0, g=0.0078125, rate=74.2)
[local] 11 1 51.9 (best c=32.0, g=0.0078125, rate=74.2)
[local] -3 1 46.9 (best c=32.0, g=0.0078125, rate=74.2)
[local] 9 -7 71.6 (best c=32.0, g=0.0078125, rate=74.2)
[local] 9 -1 60.9 (best c=32.0, g=0.0078125, rate=74.2)
[local] 9 -13 75.7 (best c=512.0, g=0.0001220703125, rate=75.7)
[local] 9 1 51.9 (best c=512.0, g=0.0001220703125, rate=75.7)
[local] 5 -11 75.3 (best c=512.0, g=0.0001220703125, rate=75.7)
[local] -1 -11 46.4 (best c=512.0, g=0.0001220703125, rate=75.7)
[local] 11 -11 73.9 (best c=512.0, g=0.0001220703125, rate=75.7)
```

[local] -3 -11 46.4 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 9 -11 73.3 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 3 -7 75.1 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 3 -1 63.7 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 3 -13 60.0 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 3 1 55.2 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 3 -11 72.7 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 5 -5 72.9 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] -1 -5 75.2 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 11 -5 68.2 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] -3 -5 70.8 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 9 -5 68.7 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 3 -5 74.2 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 15 -7 66.1 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 15 -1 61.0 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 15 -13 72.9 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 15 1 51.9 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 15 -11 72.2 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 15 -5 64.1 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 5 -15 60.0 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] -1 -15 46.4 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 11 -15 75.6 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] -3 -15 46.4 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 9 -15 75.3 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 3 -15 46.4 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 15 -15 73.7 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] -5 -7 46.4 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] -5 -1 46.4 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] -5 -13 46.4 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] -5 1 46.4 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] -5 -11 46.4 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] -5 -5 47.9 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] -5 -15 46.4 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 5 3 49.6 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] -1 3 48.0 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 11 3 49.8 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] -3 3 46.4 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 9 3 49.8 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 3 3 49.4 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 15 3 49.8 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] -5 3 46.4 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 7 -7 73.5 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 7 -1 61.4 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 7 -13 75.2 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 7 1 52.1 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 7 -11 75.7 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 7 -5 71.4 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 7 -15 72.3 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 7 3 49.8 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 5 -9 75.7 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] -1 -9 59.5 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 11 -9 73.0 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] -3 -9 46.4 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 9 -9 73.6 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 3 -9 75.3 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 15 -9 68.6 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] -5 -9 46.4 (best c=512.0, g=0.0001220703125, rate=75.7)
 [local] 7 -9 73.7 (best c=512.0, g=0.0001220703125, rate=75.7)

```
[local] 1 -7 75.3 (best c=512.0, g=0.0001220703125, rate=75.7)
[local] 1 -1 63.8 (best c=512.0, g=0.0001220703125, rate=75.7)
[local] 1 -13 46.4 (best c=512.0, g=0.0001220703125, rate=75.7)
[local] 1 1 55.2 (best c=512.0, g=0.0001220703125, rate=75.7)
[local] 1 -11 60.0 (best c=512.0, g=0.0001220703125, rate=75.7)
[local] 1 -5 76.5 (best c=2.0, g=0.03125, rate=76.5)
[local] 1 -15 46.4 (best c=2.0, g=0.03125, rate=76.5)
[local] 1 3 49.4 (best c=2.0, g=0.03125, rate=76.5)
[local] 1 -9 72.5 (best c=2.0, g=0.03125, rate=76.5)
[local] 5 -3 71.3 (best c=2.0, g=0.03125, rate=76.5)
[local] -1 -3 73.4 (best c=2.0, g=0.03125, rate=76.5)
[local] 11 -3 66.0 (best c=2.0, g=0.03125, rate=76.5)
[local] -3 -3 66.0 (best c=2.0, g=0.03125, rate=76.5)
[local] 9 -3 66.8 (best c=2.0, g=0.03125, rate=76.5)
[local] 3 -3 72.1 (best c=2.0, g=0.03125, rate=76.5)
[local] 15 -3 63.5 (best c=2.0, g=0.03125, rate=76.5)
[local] -5 -3 60.0 (best c=2.0, g=0.03125, rate=76.5)
[local] 7 -3 70.0 (best c=2.0, g=0.03125, rate=76.5)
[local] 1 -3 73.9 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 -7 67.6 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 -1 61.0 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 -13 73.7 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 1 51.9 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 -11 73.6 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 -5 65.7 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 -15 72.5 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 3 49.8 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 -9 71.2 (best c=2.0, g=0.03125, rate=76.5)
[local] 13 -3 63.5 (best c=2.0, g=0.03125, rate=76.5)
2.0 0.03125 76.5
```

Effectiveness Evaluation on the SVM-based disease risk assessment model

- For Effectiveness Evaluation on the SVM-based disease risk assessment model, the stratified ten-fold cross-validation is conducted by the LibSVM.

```
C:\NMF_Based_matrix>c:\libsvm-3.22\windows\svm-train.exe -v 10 -c 2 -g 0.03125 testing-1.libsvm
```

EXPECTED RESULT

The execution result will be shown finally.

```
AUC = 0.82188
Accuracy = 74.3% (743/1000)
AP = 0.834817
BAC = 0.743
Sensitivity = 0.74 (370/(370+130))
Specificity = 0.746 (373/(373+127))
Cross Validation = 74.3%
C:\NMF_Based_matrix>
```



This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited