**Supplementary Materials for** The Role of Initial Feature in Information Theory-based Feature Selection Methods and an Improved Solution

**Supplementary Files**

**Content**

Figure

[Figure S1. Comparison of Acc achieved by different choice of bin number on datasets with total number of instances less than 1000. 2](#_Toc37834103)

[Figure S2. Comparison of Acc achieved by different choice of bin number on datasets with total number of instances more than 1000. 3](#_Toc37834104)

[Figure S3. The comparison between original methods and their corresponding improved methods with criteria of average Acc of Musk dataset with different feature subset size and different classifiers. 4](#_Toc37834105)

[Figure S4. The comparison between CMIM and improved CMIM on datasets with different classifiers. 5](#_Toc37834106)

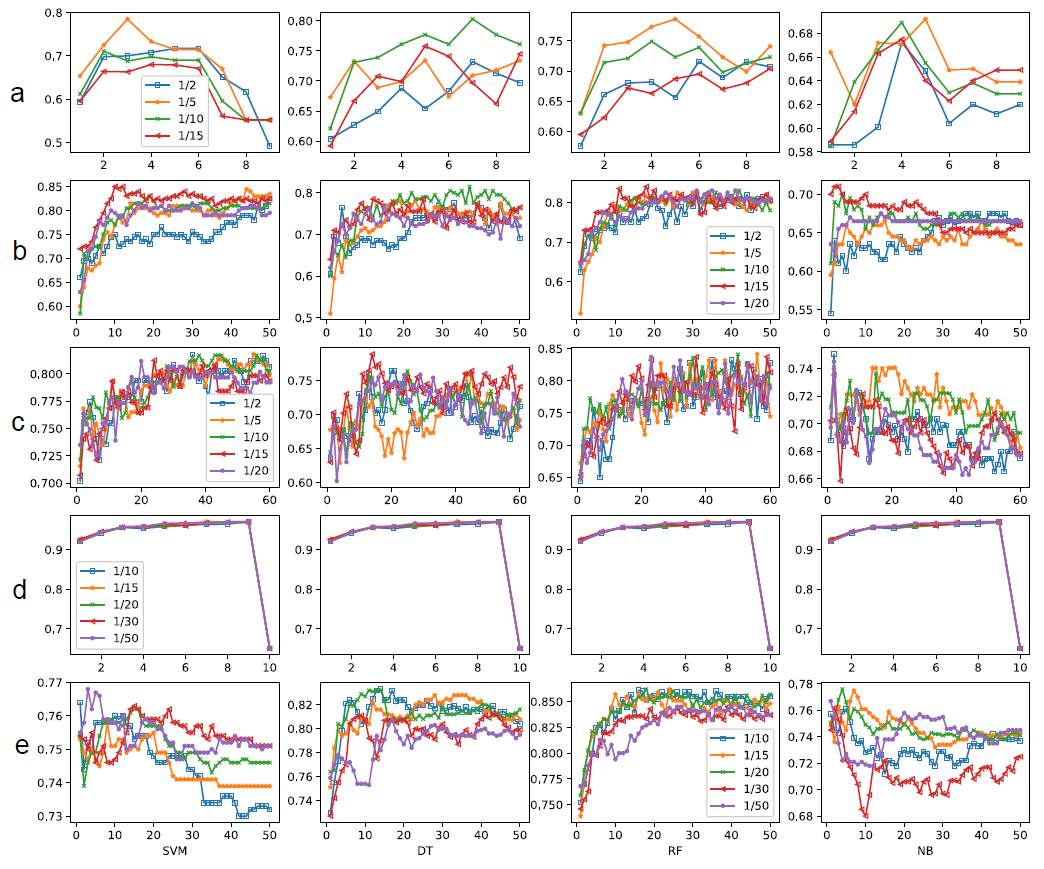
**Table**

[Table S1. Counts of times for best Acc achieved by different choice of bin number. 6](#_Toc37834107)

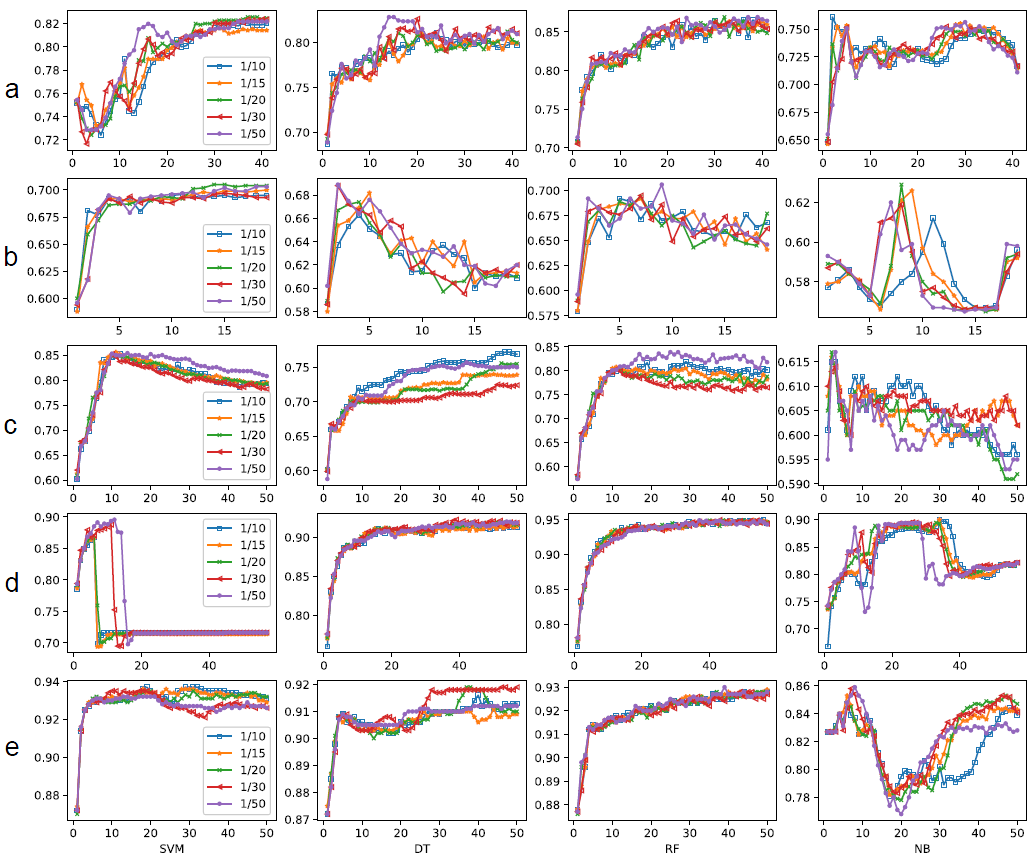
[Table S2. The best choice of bin number for 10 datasets. 6](#_Toc37834108)

[Table S3. Initial feature position in original feature set obtained with different choice of bin number. 7](#_Toc37834109)

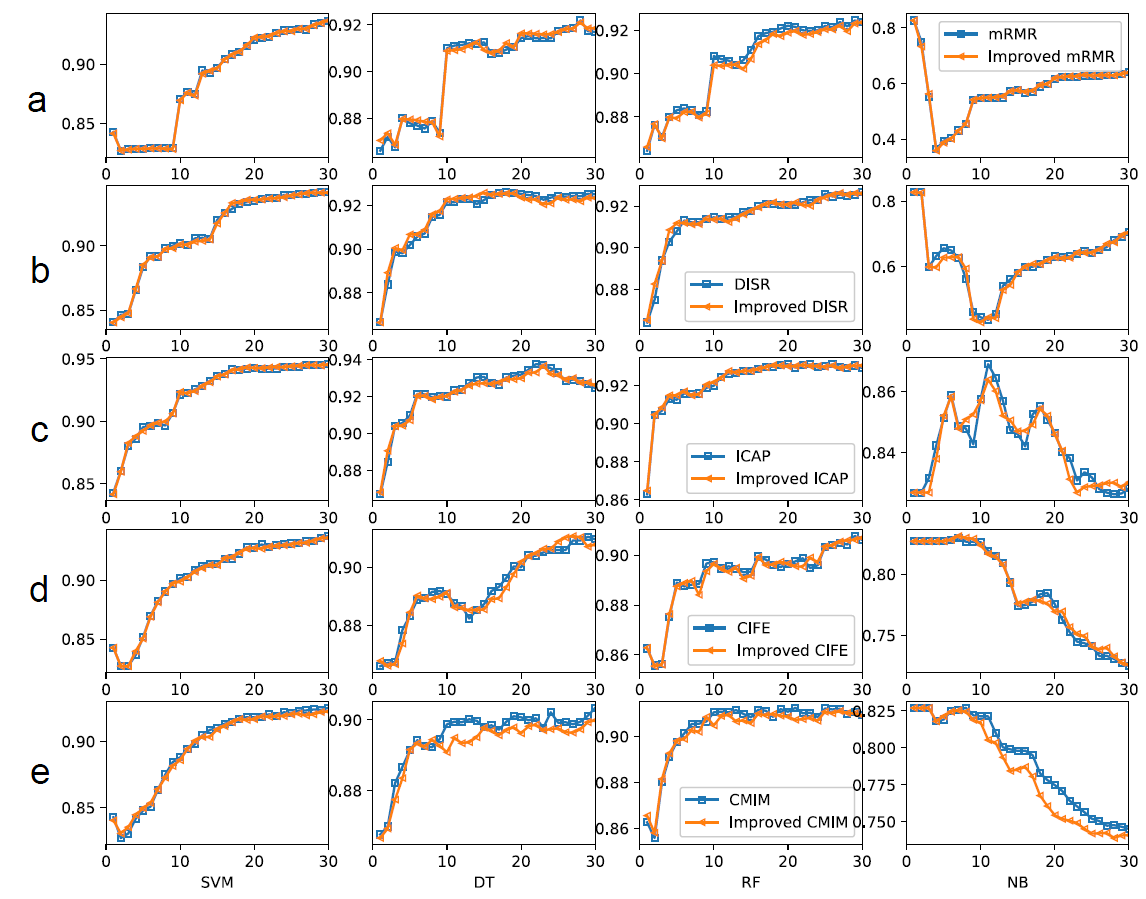
# **Figure S1.** **Comparison of Acc achieved by different choice of bin number on datasets with total number of instances less than 1000.** Dataset (a) Breast Cancer Coimbra, (b) Arcene, (c) Sonar, (d) Breast Cancer Wisconsin, and (e) Parkinson.



# **Figure S2.** Comparison of Acc achieved by different choice of bin number on datasets with total number of instances more than 1000. Dataset (a) Biodegradation, (b) Diabetic Retinopathy Debrecen, (c) Madelon, (d) Spambase and (e) Musk.



# **Figure S3.** The comparison between original methods and their corresponding improved methods with criteria of average Acc of Musk dataset with different feature subset size and different classifiers, (a) mRMR and improved mRMR, (b) DISR and improved DISR, (c) ICAP and improved ICAP, (d) CIFE and improved CIFE, and (e) CMIM and improved CMIM.



# **Figure S4.** The comparison between CMIM and improved CMIM on dataset (a) Arcene, (b) Biodegradation, (c) Breast Cancer Wisconsin, (d) Breast Cancer Coimbra, (e) Diabetic Retinopathy Debrecen, (f) Madelon, (g) Parkinson, (h) Sonar and (i) Spambase with different classifiers.

# **Table S1**. Counts of times for best Acc achieved by different choice of bin number.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| datasets | 1/2 (portion of total instance number) | 1/5 (portion of total instance number) | 1/10 (portion of total instance number) | 1/15 (portion of total instance number) | 1/20 (portion of total instance number) |
| Breast Cancer Coimbra |  | 3 | 1 |  |  |
| Arcene |  |  | 1 | 3 |  |
| Sonar | 1 | 1 | 1 | 1 |  |
| datasets | 1/10 (portion of total instance number) | 1/15 (portion of total instance number) | 1/20 (portion of total instance number) | 1/30 (portion of total instance number) | 1/50 (portion of total instance number) |
| Breast Cancer Wisconsin\* | - | - | - | - | - |
| Parkinson | 1 |  | 2 |  | 1 |
| Biodegradation | 1 |  | 1 |  | 2 |
| Diabetic Retinopathy Debrecen |  |  | 2 |  | 2 |
| Madelon | 1 | 1 |  |  | 2 |
| Spambase \*\* |  | 1 |  |  | 1 |
| Musk | 1 |  |  | 1 | 2 |

\*the performances of different bin number choices for breast cancer Wisconsin dataset are not distinguishable.

\*\* the performances of different bin number choices with DT and RF classifiers are not distinguishable.

# **Table S2**. The best choice of bin number for 10 datasets.

|  |  |  |
| --- | --- | --- |
| Datasets | Number of instances | Best bin number in portion\* |
| Breast Cancer Coimbra | 116 | 1/5 |
| Arcene | 200 | 1/15 |
| Sonar | 208 | -\*\* |
| Breast Cancer Wisconsin | 683 | - |
| Parkinson | 756 | 1/20 |
| Biodegradation | 1055 | 1/50 |
| Diabetic Retinopathy Debrecen | 1151 | 1/50 |
| Madelon | 2600 | 1/50 |
| Spambase | 4601 | 1/50 |
| Musk | 7074 | 1/50 |

# **Table S3**. Initial feature position in original feature set obtained with different choice of bin number.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| datasets | 1/2 (portion of total instance number) | 1/5 (portion of total instance number) | 1/10 (portion of total instance number) | 1/15 (portion of total instance number) | 1/20 (portion of total instance number) |
| Breast Cancer Coimbra | 1 | 1 | 1 | 1 | 1 |
| Arcene | 7868 | 4381 | 6336 | 1516 | 5083 |
| Sonar | 11 | 12 | 12 | 12 | 11 |
| datasets | 1/10 (portion of total instance number) | 1/15 (portion of total instance number) | 1/20 (portion of total instance number) | 1/30 (portion of total instance number) | 1/50 (portion of total instance number) |
| Breast Cancer Wisconsin\* | 4 | 4 | 4 | 4 | 4 |
| Parkinson | 427 | 427 | 427 | 58 | 404 |
| Biodegradation | 36 | 36 | 36 | 36 | 36 |
| Diabetic Retinopathy Debrecen | 3 | 5 | 3 | 3 | 3 |
| Madelon | 242 | 242 | 242 | 339 | 339 |
| Spambase \*\* | 52 | 53 | 53 | 53 | 53 |
| Musk | 151 | 151 | 151 | 151 | 151 |