**Information Theory-based Feature Selection: Minimum Distribution Similarity with Removed Redundancy**

Yu Zhang1,\* , Zhuoyi Lin1 and Chee Keong Kwoh1,\*

1School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore

\* To whom correspondence should be addressed. Tel: +65 6790 6057; Email: [asckkwoh@ntu.edu.sg](mailto:asckkwoh@ntu.edu.sg)

Correspondence may also be addressed to Email: [yu007@ntu.edu.sg](mailto:yu007@ntu.edu.sg)

**Supplementary Files**

**Content**

**Figures:**

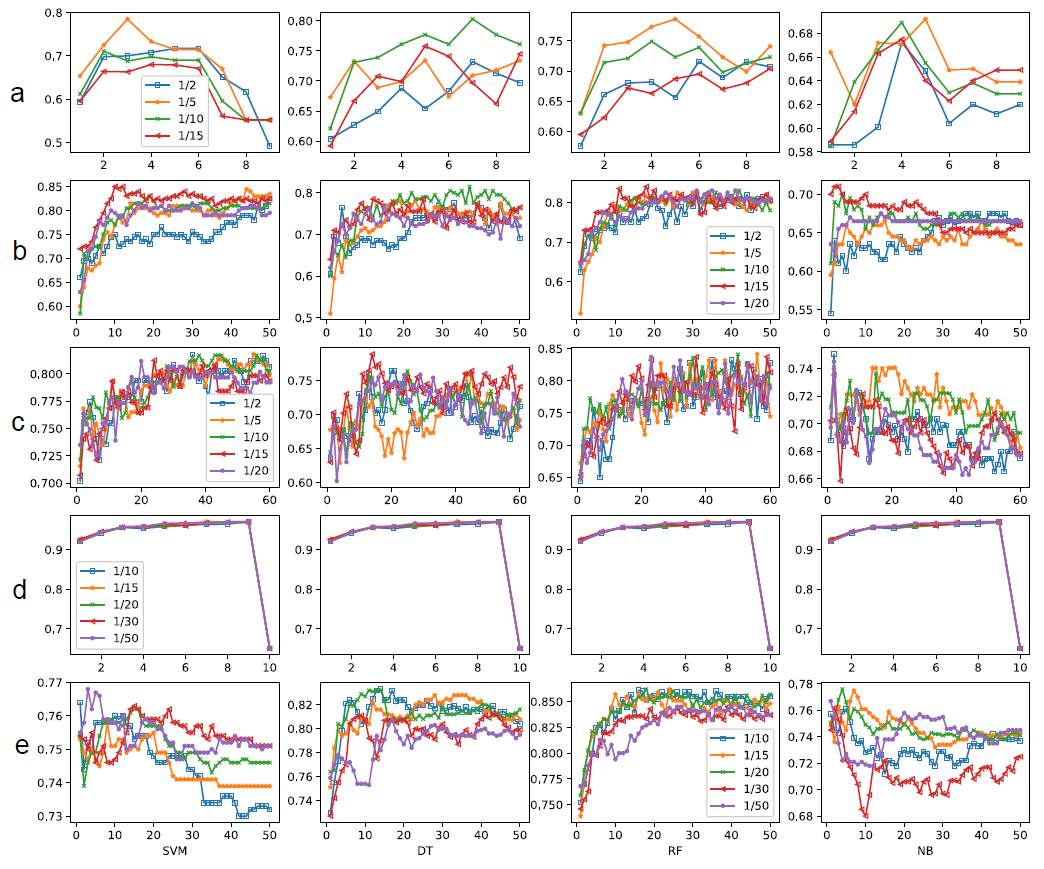
[Figure S1. Comparison of Acc achieved by different choice of bin number on datasets with total number of instances less than 1000. 2](#_Toc6219519)

[Figure S2. Comparison of Acc achieved by different choice of bin number on datasets with total number of instances more than 1000. 3](#_Toc6219520)

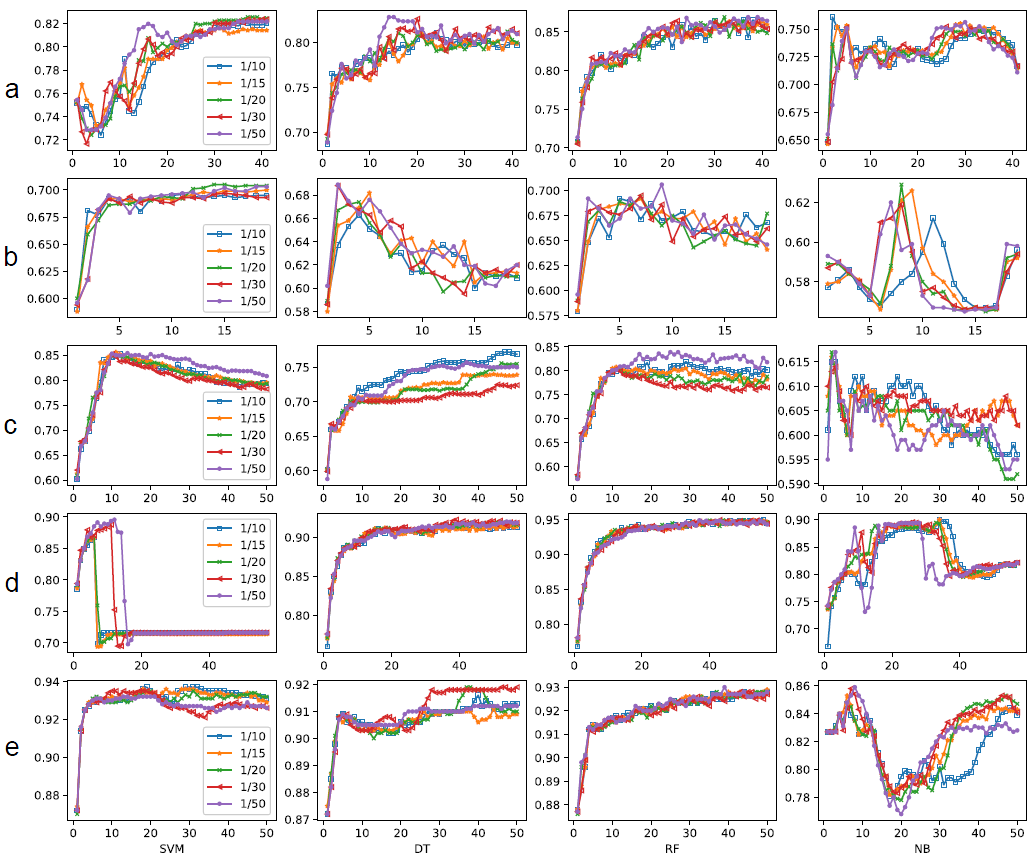
[Figure S3. The comparison between the original methods and their corresponding improved methods with criteria of average Acc on 10-fold cross validation of Musk dataset with different feature subset size and different classifiers. 4](#_Toc6219521)

**Tables:**

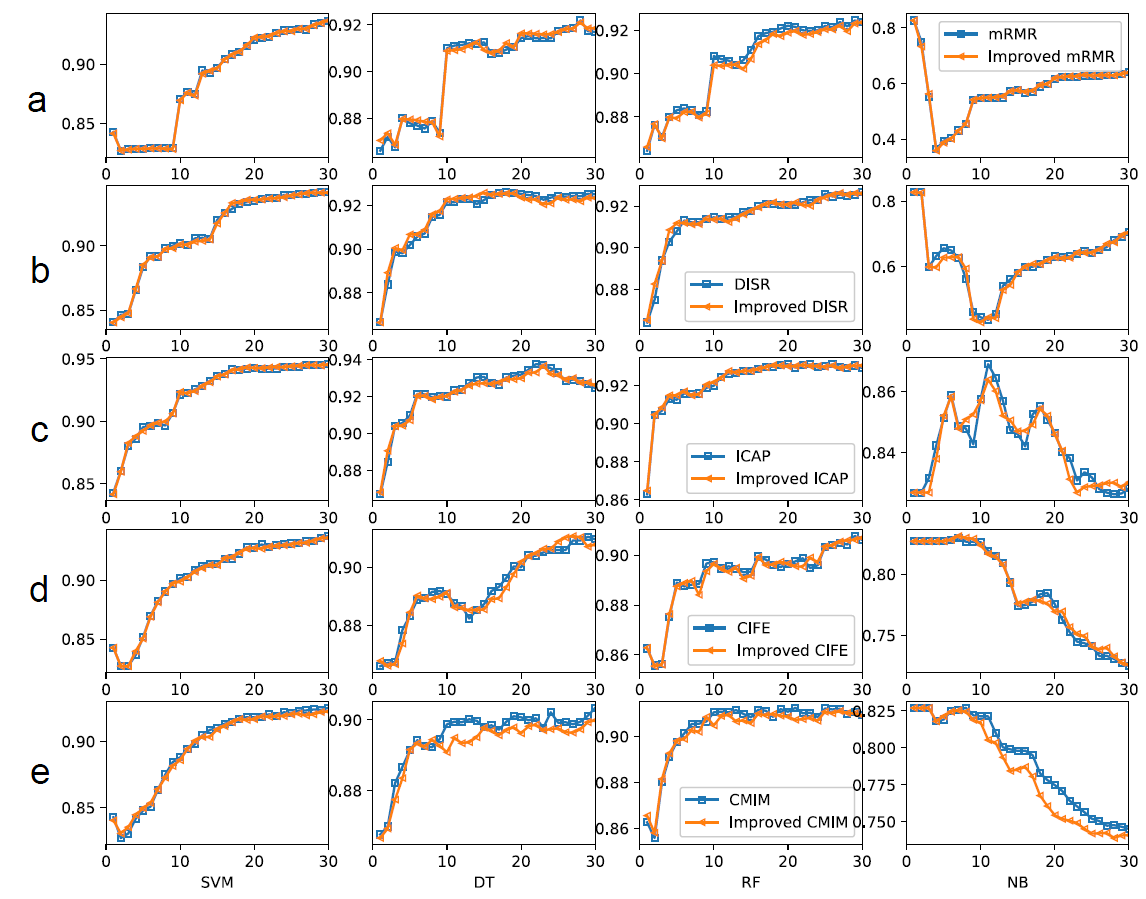
[Table S1. Counts of times for best Acc achieved by different choice of bin number. 7](#_Toc6219519)



**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.

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| datasets | 1/2 (percentage of total instance number) | 1/5 (percentage of total instance number) | 1/10 (percentage of total instance number) | 1/15 (percentage of total instance number) | 1/20 (percentage of total instance number) |
| Breast Cancer Coimbra |  | 3 | 1 |  |  |
| Arcene |  |  | 1 | 3 |  |
| Sonar | 1 | 1 | 1 | 1 |  |
| datasets | 1/10 (percentage of total instance number) | 1/15 (percentage of total instance number) | 1/20 (percentage of total instance number) | 1/30 (percentage of total instance number) | 1/50 (percentage 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.