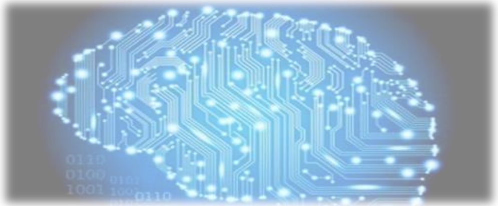
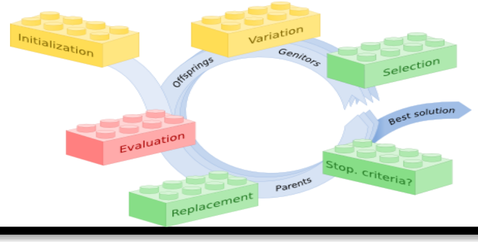




제 132회 석사 학위 졸업 논문 발표

A Hybrid Method for Multi-label Feature Selection considering Relative Effectiveness

상대적 효과성을 고려하는 다중 레이블 특징 선별을 위한 하이브리드 방법에 대한 연구

Jaegyun Park
jgp0566.cau@gmail.com
2019. 11. 22.

Introduction

- **Recently, multi-label data associated with more than one answers has explosively collected.**

Multi-label data

(Dog, Cat)

(Rabbit, Cat)

Single-label data

Cat

Dog

Multi-label Data and Feature Selection

- Through **Feature Selection** ➡ Select *relevant features* ➡ **High Dimensionality** ⬇

  = *Best feature subset* **Classification Accuracy** ⬆

Multi-label Classification (e.g. Text Sentiment Analysis)

I planned a trip to Japan, but I finally refunded my ticket, damn it. I should've gone to Jeju.

I hung up my Jeju ticket to travel with my friends. Can you recommend a good tourist spots?

Features

	go	*Japan*	Jeju	*damn*	*travel*	*good*
#1	1	1	1	1	0	0
#2	0	0	1	0	1	1

Labels

	worrying	angry	excited
#1	1	1	0
#2	0	0	1

Convention

Hybrid methods [1, 2, 3, 4, 5]

= Spend *Fitness Function Calls (FFCs)*
= Train the classification model

Feature subset s_i^t $\xrightarrow{\text{① Evaluate its fitness}}$ **Fitness**

② Modify it by *filter or evolutionary* operator

Filter operator [6, 7, 8]

- Rankings of features

damn > *go* *but* > *Japan* > *my*

Evolutionary operator [9, 10, 11]

Fitness	Feature subsets
0.23	s_1^t
0.42	s_2^t
0.48	s_3^t

recombination

Modified feature subsets

s_2^{t+1}

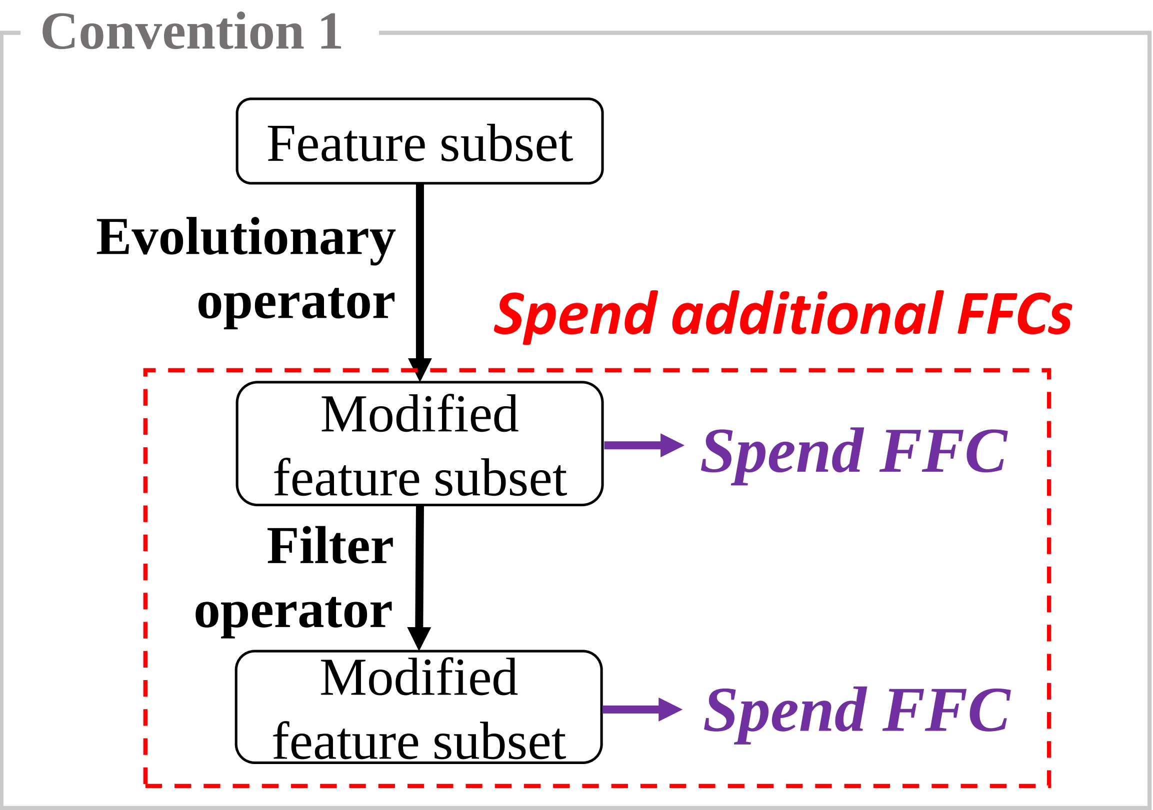
s_3^{t+1}

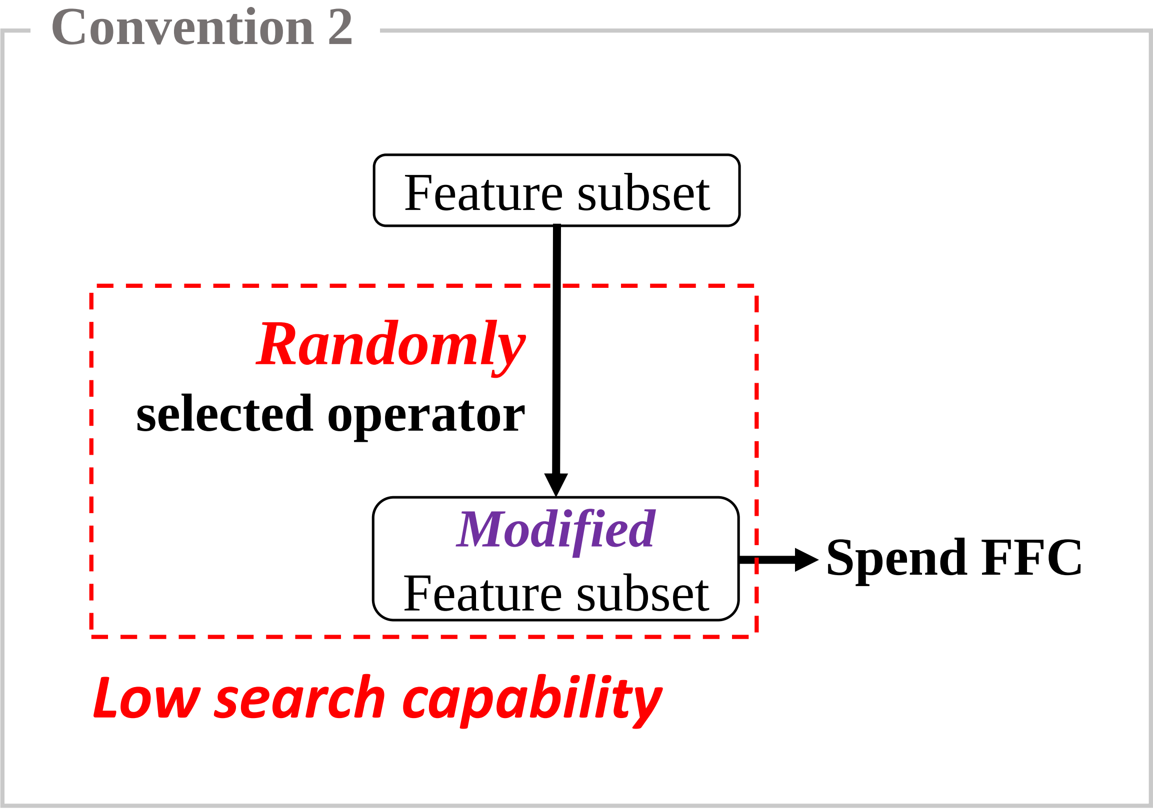
Example) s_i^t ={*go*, *my*} Filter s_i^{t+1} ={*damn*, *go*} Evolutionary s_i^{t+1} ={*damn*, *Japan*}

Convention : Problem

- **If something can check which operator will be more effective each time,**

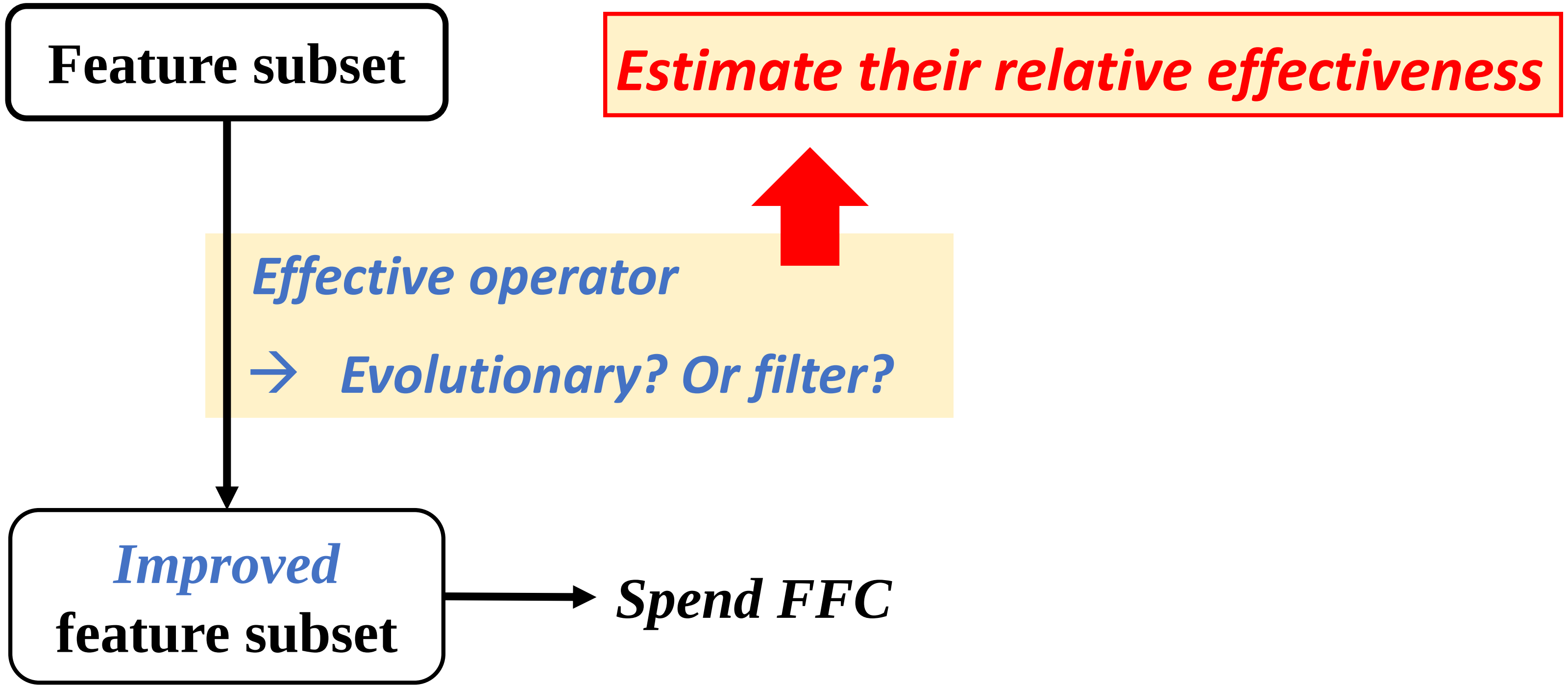
  then **Wasted FFCs ↓, Search Capability ↑**



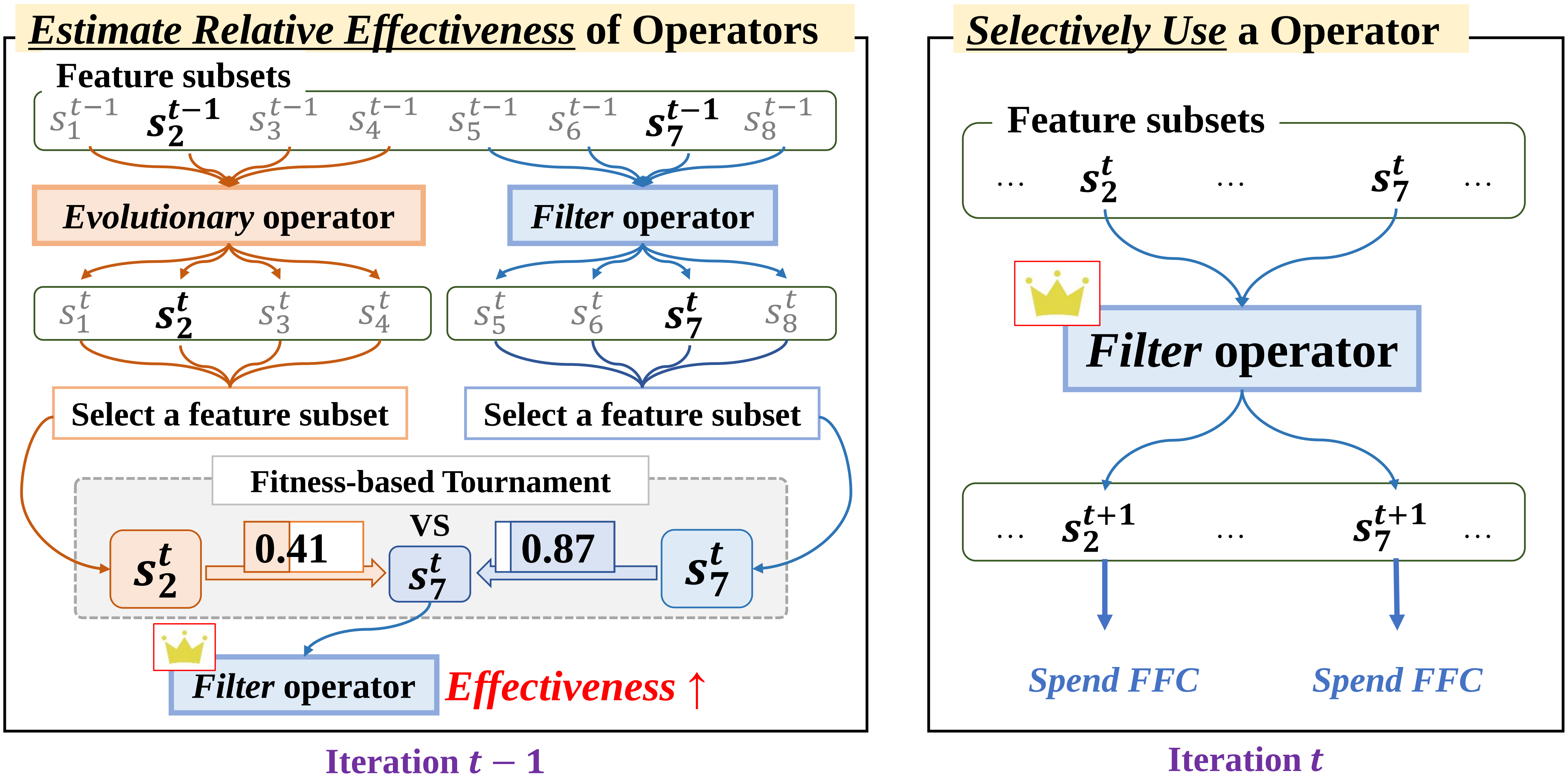


Motivation

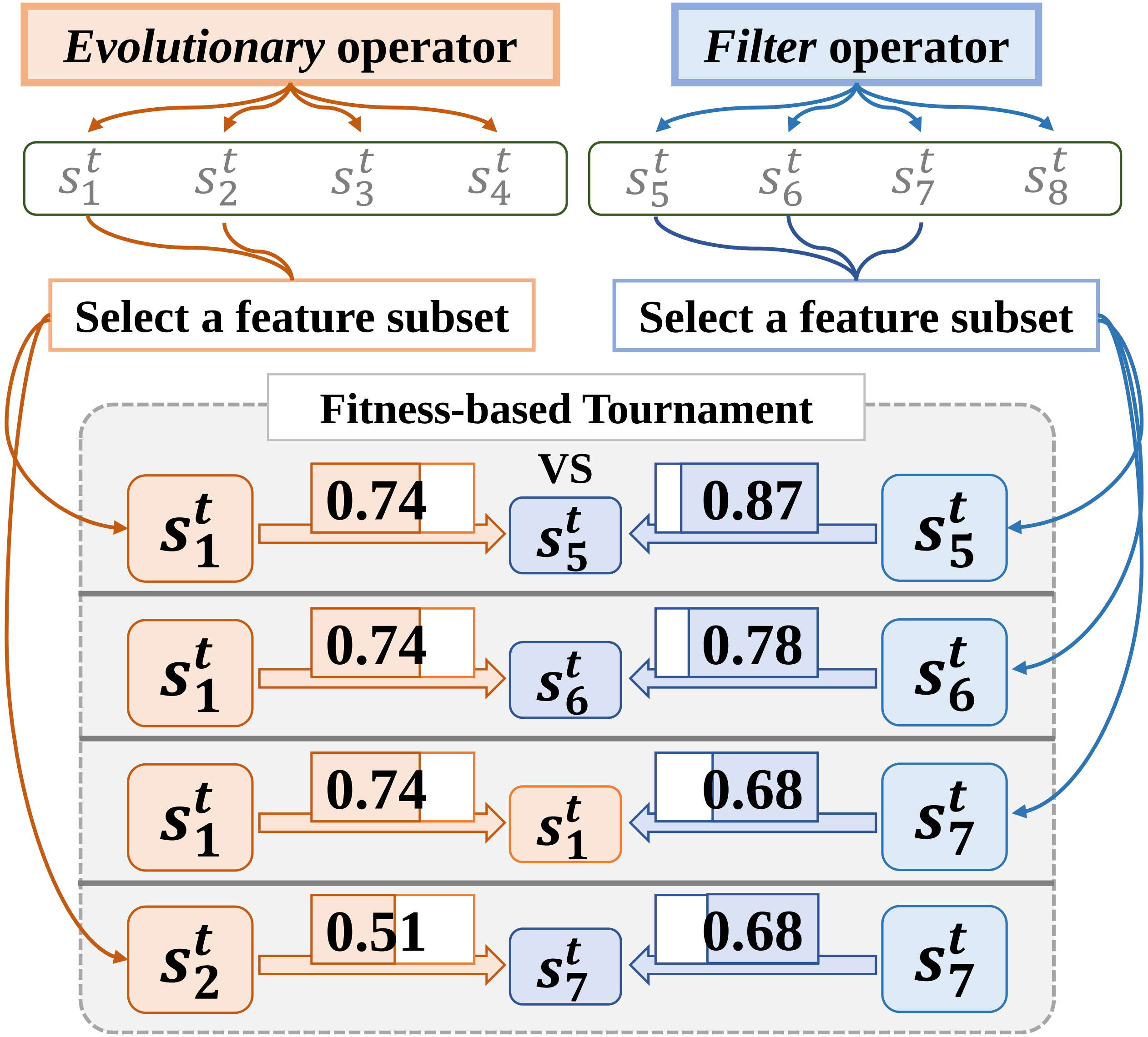
- **It is to selectively use an effective operator by estimating the relative effectiveness of operators.**



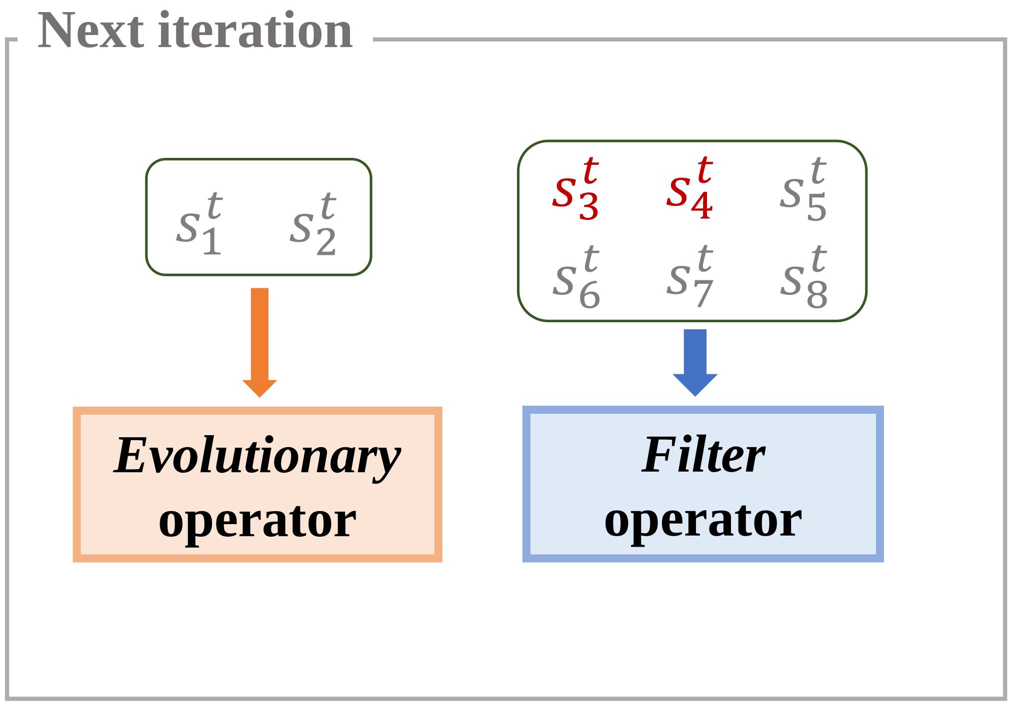
Proposed Method



Proposed Method : Estimation process



	Evolutionary	*Filter*
Win	1	3
Relative effectiveness	$1 - 3 = -2$	$3 - 1 = 2$



Experiment Setting

- **16 multi-label datasets**
- **Experiment :**
  - Hold-out cross validation (8:2)
  - Repeated 10 times
- **Evolutionary Operator**
  - Particle Swarm Optimization (PSO) [9].
- **Filter Operator**
  - Quadratic programming-based multi-label feature selection [12].

The standard statistics of employed multi-label datasets

Datasets	Patterns	Features	Labels	Card.	Den.	Distinct
RCV1 (S1)	6,000	945	101	2.880	0.029	1,028
RCV1 (S2)	6,000	945	101	2.634	0.026	954
RCV1 (S3)	6,000	945	101	2.614	0.026	939
RCV1 (S4)	6,000	945	101	2.484	0.025	816
RCV1 (S5)	6,000	945	101	2.642	0.026	946
Arts	7,484	1,157	26	1.654	0.064	599
Business	11,214	1,096	30	1.599	0.053	233
Computers	12,444	1,705	33	1.507	0.046	428
Education	12,030	1,377	33	1.463	0.044	511
Entertainment	12,730	1,600	21	1.414	0.067	337
Health	9,205	1,530	32	1.644	0.051	335
Recreation	12,828	1,516	22	1.429	0.065	530
Reference	8,027	1,984	33	1.174	0.036	275
Science	6,428	1,859	40	1.450	0.036	457
Social	12,111	2,618	29	1.279	0.033	361
Society	14,512	1,590	27	1.670	0.062	1,054

[9] Zhang, Y., Gong, D.W., Sun, X.Y., Guo, Y.N. (**2017**). A PSO-based multi-objective multi-label feature selection method in classification. *Sci. Rep.*, 7:376.
[12] Lim, H., Lee, J., Kim, D.W. (**2017**). Optimization approach for feature selection in multi-label classification. *Pattern Recognit. Lett.*, 89:25-30.

Experimental Setting

- **Comparison Methods :**
  Hybrid-based methods : EGA + CDM [1], bALO-QR [13]
  PSO-based method : CSO [14]

- **Classifier : Multi-Label Naïve Bayes (MLNB) [15]**

- **Termination Criterion : After the number of evaluation (FFCs) of feature subset is over 300**

- **Evaluation Measure :**

◆ **Hamming Loss : evaluates the fraction of misclassified instance-label pairs.**
◆ **One-error : evaluates the fraction of misclassified instance-label pairs for the top-ranked label.**
◆ **Multi-label Accuracy : evaluates the overall effectiveness of a classifier.**
◆ **Subset Accuracy : evaluates the fraction of correctly classified examples.**

[1] Ghareb, A.S., Bakar, A.A., Hamdan, A.R. (**2016**). Hybrid feature selection based on enhanced genetic algorithm for text categorization. *Expert Syst. Appl.*, 49:31-47.
[13] Mafarja, M.M., Mirjalili, S. (**2018**). Hybrid binary ant lion optimizer with rough set and approximate entropy reducts for feature selection. *Soft Comput.*, 1-17.
[14] Gu, S., Cheng, R., Jin, Y. (**2018**). Feature selection for high-dimensional classification using a competitive swarm optimizer. *Soft Comput.*, 22:811-822.
[15] Zhang, M.L., Peña, J.M., Robles, V. (**2009**). Feature selection for multi-label naive Bayes classification. *Inf. Sci.*, 179:3218-3229.

Experimental Result

- **Comparison results in terms of each evaluation measure**

Datasets	Hamming loss				One-error			
	Proposed	EGA+CDM	bALO-QR	CSO	Proposed	EGA+CDM	bALO-QR	CSO
RCV1 (S1)	0.029 ± 0.001	0.030 ± 0.001	0.030 ± 0.000	**0.029 ± 0.001**	**0.578 ± 0.149**	0.637 ± 0.129	0.621 ± 0.134	0.648 ± 0.125
RCV1 (S2)	**0.027 ± 0.001**	0.028 ± 0.003	0.027 ± 0.001	0.027 ± 0.001	**0.524 ± 0.014**	0.654 ± 0.023	0.580 ± 0.015	0.599 ± 0.019
RCV1 (S3)	**0.026 ± 0.000**	0.027 ± 0.001	0.027 ± 0.001	0.026 ± 0.001	**0.607 ± 0.198**	0.718 ± 0.150	0.671 ± 0.174	0.683 ± 0.168
RCV1 (S4)	**0.024 ± 0.001**	0.025 ± 0.001	0.025 ± 0.001	0.024 ± 0.001	**0.583 ± 0.199**	0.696 ± 0.160	0.671 ± 0.175	0.672 ± 0.174
RCV1 (S5)	**0.026 ± 0.001**	0.028 ± 0.003	0.028 ± 0.001	0.026 ± 0.001	**0.612 ± 0.201**	0.695 ± 0.161	0.656 ± 0.182	0.652 ± 0.185
Arts	**0.060 ± 0.001**	0.067 ± 0.002	0.069 ± 0.002	0.066 ± 0.002	**0.669 ± 0.169**	0.712 ± 0.149	0.710 ± 0.149	0.712 ± 0.149
Business	**0.031 ± 0.001**	0.036 ± 0.004	0.034 ± 0.001	0.034 ± 0.002	**0.388 ± 0.411**	0.399 ± 0.409	0.398 ± 0.400	0.396 ± 0.406
Computers	**0.042 ± 0.001**	0.051 ± 0.004	0.046 ± 0.001	0.047 ± 0.001	**0.410 ± 0.010**	0.469 ± 0.006	0.445 ± 0.009	0.448 ± 0.007
Education	**0.042 ± 0.001**	0.048 ± 0.002	0.048 ± 0.002	0.047 ± 0.001	**0.578 ± 0.015**	0.661 ± 0.008	0.616 ± 0.020	0.639 ± 0.016
Entertainment	**0.057 ± 0.001**	0.069 ± 0.003	0.065 ± 0.001	0.065 ± 0.001	**0.512 ± 0.007**	0.605 ± 0.019	0.563 ± 0.015	0.586 ± 0.015
Health	**0.039 ± 0.001**	0.050 ± 0.003	0.047 ± 0.001	0.047 ± 0.002	**0.739 ± 0.324**	0.774 ± 0.282	0.764 ± 0.300	0.778 ± 0.238
Recreation	**0.057 ± 0.001**	0.070 ± 0.003	0.067 ± 0.002	0.065 ± 0.001	**0.553 ± 0.011**	0.739 ± 0.013	0.675 ± 0.013	0.675 ± 0.011
Reference	**0.031 ± 0.001**	0.040 ± 0.003	0.037 ± 0.002	0.037 ± 0.001	**0.702 ± 0.253**	0.715 ± 0.243	0.718 ± 0.241	0.715 ± 0.243
Science	**0.036 ± 0.001**	0.043 ± 0.003	0.042 ± 0.001	0.042 ± 0.001	**0.630 ± 0.020**	0.707 ± 0.018	0.696 ± 0.027	0.696 ± 0.023
Social	**0.026 ± 0.001**	0.042 ± 0.004	0.032 ± 0.002	0.032 ± 0.001	**0.437 ± 0.198**	0.571 ± 0.152	0.472 ± 0.186	0.490 ± 0.179
Society	**0.057 ± 0.001**	0.065 ± 0.004	0.064 ± 0.001	0.063 ± 0.001	**0.448 ± 0.013**	0.510 ± 0.017	0.489 ± 0.019	0.479 ± 0.016
Avg. Rank	**1.06**	3.88	2.94	2.13	**1.00**	3.75	2.31	2.94

Experimental Result

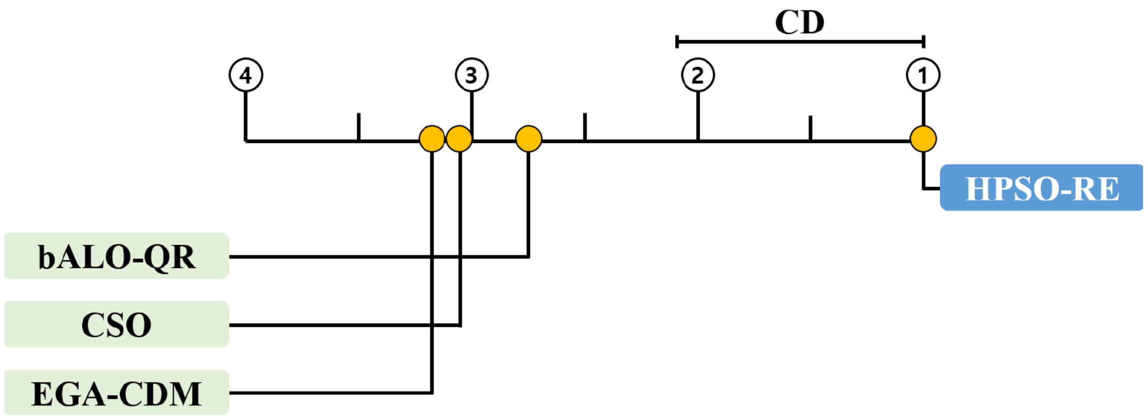
- **Comparison results in terms of each evaluation measure**

Datasets	**Multi-label accuracy**				**Subset accuracy**			
	Proposed	EGA+CDM	bALO-QR	CSO	Proposed	EGA+CDM	bALO-QR	CSO
RCV1 (S1)	**0.199 ± 0.010**	0.176 ± 0.011	0.168 ± 0.013	0.124 ± 0.013	**0.019 ± 0.005**	0.009 ± 0.002	0.012 ± 0.007	0.012 ± 0.005
RCV1 (S2)	**0.243 ± 0.011**	0.177 ± 0.011	0.179 ± 0.014	0.157 ± 0.018	**0.102 ± 0.016**	0.011 ± 0.005	0.087 ± 0.010	0.087 ± 0.004
RCV1 (S3)	**0.220 ± 0.016**	0.161 ± 0.004	0.178 ± 0.019	0.168 ± 0.014	**0.113 ± 0.014**	0.025 ± 0.005	0.093 ± 0.009	0.102 ± 0.005
RCV1 (S4)	**0.268 ± 0.014**	0.170 ± 0.007	0.192 ± 0.014	0.183 ± 0.019	**0.153 ± 0.011**	0.033 ± 0.014	0.120 ± 0.016	0.126 ± 0.016
RCV1 (S5)	**0.232 ± 0.011**	0.187 ± 0.009	0.191 ± 0.016	0.165 ± 0.012	**0.092 ± 0.008**	0.013 ± 0.003	0.082 ± 0.012	0.091 ± 0.011
Arts	**0.180 ± 0.013**	0.094 ± 0.008	0.099 ± 0.008	0.106 ± 0.012	**0.151 ± 0.007**	0.058 ± 0.007	0.071 ± 0.007	0.075 ± 0.006
Business	**0.673 ± 0.007**	0.662 ± 0.009	0.654 ± 0.008	0.656 ± 0.011	**0.531 ± 0.012**	0.514 ± 0.016	0.507 ± 0.011	0.512 ± 0.011
Computers	**0.423 ± 0.007**	0.369 ± 0.010	0.388 ± 0.006	0.391 ± 0.008	**0.352 ± 0.008**	0.299 ± 0.011	0.316 ± 0.010	0.319 ± 0.009
Education	**0.149 ± 0.007**	0.075 ± 0.008	0.109 ± 0.012	0.085 ± 0.018	**0.092 ± 0.009**	0.047 ± 0.009	0.074 ± 0.007	0.064 ± 0.013
Entertainment	**0.229 ± 0.013**	0.173 ± 0.007	0.220 ± 0.011	0.188 ± 0.011	**0.232 ± 0.014**	0.130 ± 0.009	0.188 ± 0.010	0.176 ± 0.022
Health	**0.507 ± 0.010**	0.410 ± 0.017	0.397 ± 0.019	0.423 ± 0.020	**0.386 ± 0.012**	0.307 ± 0.016	0.314 ± 0.009	0.308 ± 0.054
Recreation	**0.236 ± 0.009**	0.045 ± 0.004	0.111 ± 0.011	0.119 ± 0.008	**0.194 ± 0.009**	0.020 ± 0.003	0.093 ± 0.013	0.106 ± 0.016
Reference	**0.387 ± 0.008**	0.360 ± 0.010	0.352 ± 0.009	0.350 ± 0.013	**0.350 ± 0.016**	0.321 ± 0.006	0.316 ± 0.011	0.294 ± 0.074
Science	**0.133 ± 0.005**	0.075 ± 0.007	0.064 ± 0.010	0.070 ± 0.017	**0.106 ± 0.010**	0.053 ± 0.008	0.048 ± 0.005	0.055 ± 0.011
Social	**0.535 ± 0.012**	0.340 ± 0.021	0.471 ± 0.014	0.449 ± 0.025	**0.488 ± 0.014**	0.287 ± 0.022	0.432 ± 0.016	0.412 ± 0.012
Society	**0.329 ± 0.013**	0.290 ± 0.019	0.254 ± 0.012	0.211 ± 0.041	**0.295 ± 0.017**	0.215 ± 0.012	0.179 ± 0.028	0.157 ± 0.021
Avg. Rank	**1.00**	3.19	2.75	3.06	**1.00**	3.56	2.81	2.63

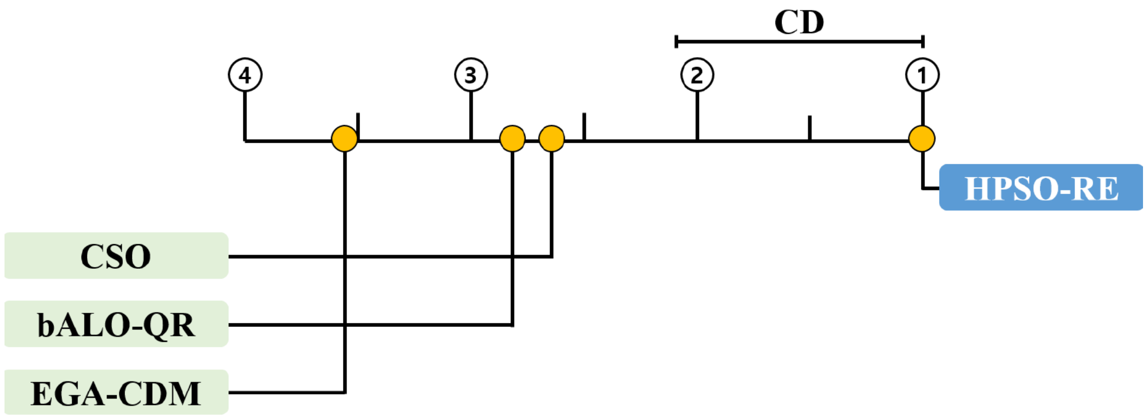
Experimental Result

- **Friedman test**
  - From the experimental results of the four algorithms, it showed a significant statistical difference.

- **Bonferroni-Dunn test**

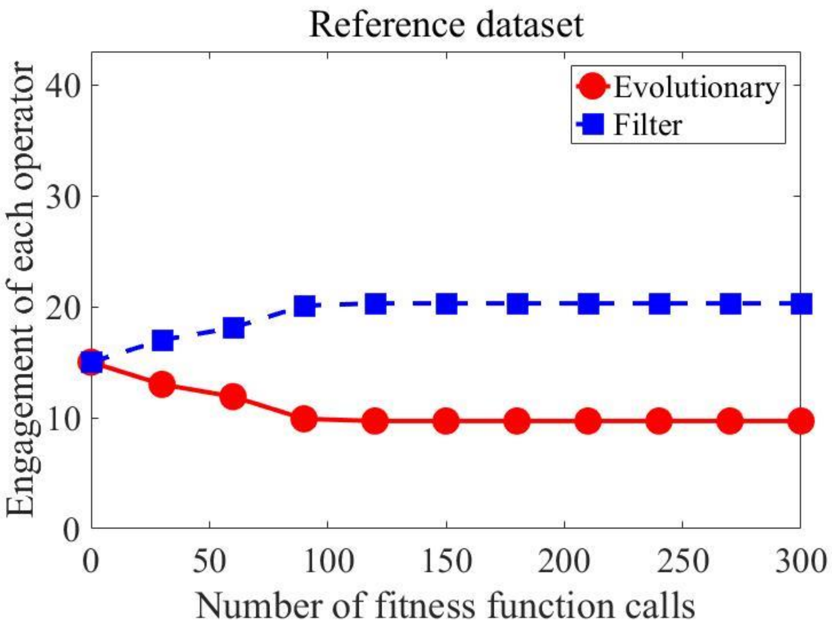


Multi-label accuracy



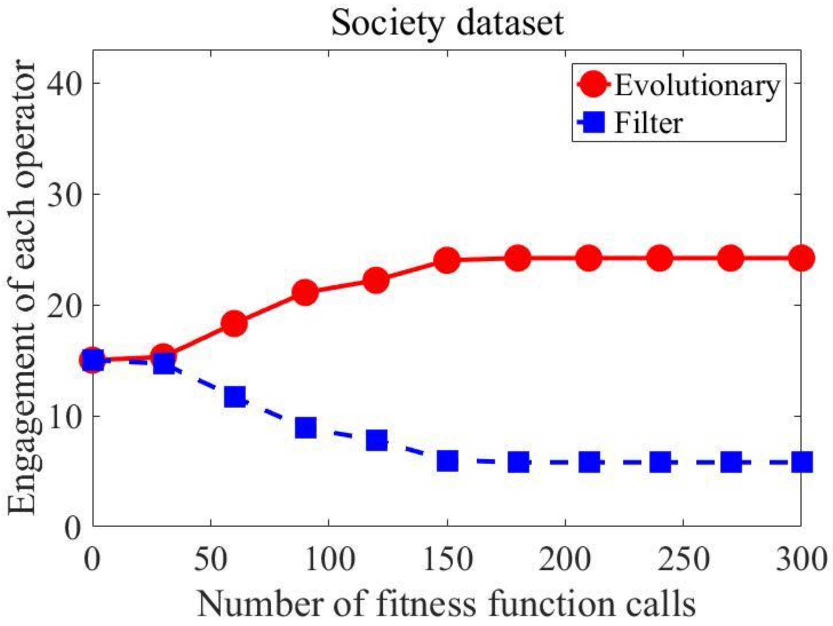
Subset accuracy

Analysis



Ranking by filter

$f_2 > f_5 > \cdots > f_1$



Ranking by filter

$f_1 > \cdots > f_2 > f_5$

→ They show that the proposed method is robust against datasets.

Contribution

The contribution of this study

- It is the proposal of a new process for advance estimation of the relative effectiveness of the operators and their selective engagement in each iteration.
- The experimental results verified that the proposed method significantly outperformed three state-of-the-art feature selection methods, on 16 multi-label datasets.

International Publication (SCI(E))

- Competitive Particle Swarm Optimization for Multi-Category Text Feature Selection, **Entropy**, 2019, **Co-author**.
- Memetic Feature Selection for Multi-label Text Categorization using Label Frequency Difference, **Information Sciences**, 2019, **Co-author**.

Thank You

Reference

[1] Ghareb, A.S., Bakar, A.A., Hamdan, A.R. (**2016**). Hybrid feature selection based on enhanced genetic algorithm for text categorization. *Expert Syst. Appl.*, 49:31-47.
[2] Lee, J., Seo, W., Kim, D.W. (**2018**). Effective evolutionary multilabel feature selection under a budget constraint. *Complexity*, 2018.
[3] Moradi, P., Gholampour, M. (**2016**). A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. *Appl. Soft Comput.***,** 43:117-130.
[4] Dong, H., Li, T., Ding, R., Sun, J. (**2018**). A novel hybrid genetic algorithm with granular information for feature selection and optimization. *Appl. Soft Comput.*, 65:33-46.
[5] Mafarja, M.M., Mirjalili, S. (**2017**). Hybrid whale optimization algorithm with simulated annealing for feature selection. *Neurocomputing*, 260:302-312.
[6] Lee, J., Kim, D.W. (**2015**). Mutual information-based multi-label feature selection using interaction information. *Expert Syst. Appl.*, 42:2013-2025.
[7] Lee, J., Kim, D.W. (**2015**). Fast multi-label feature selection based on information-theoretic feature ranking. *Pattern Recognit.*, 48:2761-2771.
[8] Rehman, A., Javed, K., Babri, H.A. (**2017**). Feature selection based on a normalized difference measure for text classification. *Inf. Process. Manag.*, 53:473-489.
[9] Zhang, Y., Gong, D.W., Sun, X.Y., Guo, Y.N. (**2017**). A PSO-based multi-objective multi-label feature selection method in classification. *Sci. Rep.*, 7:376.

Reference

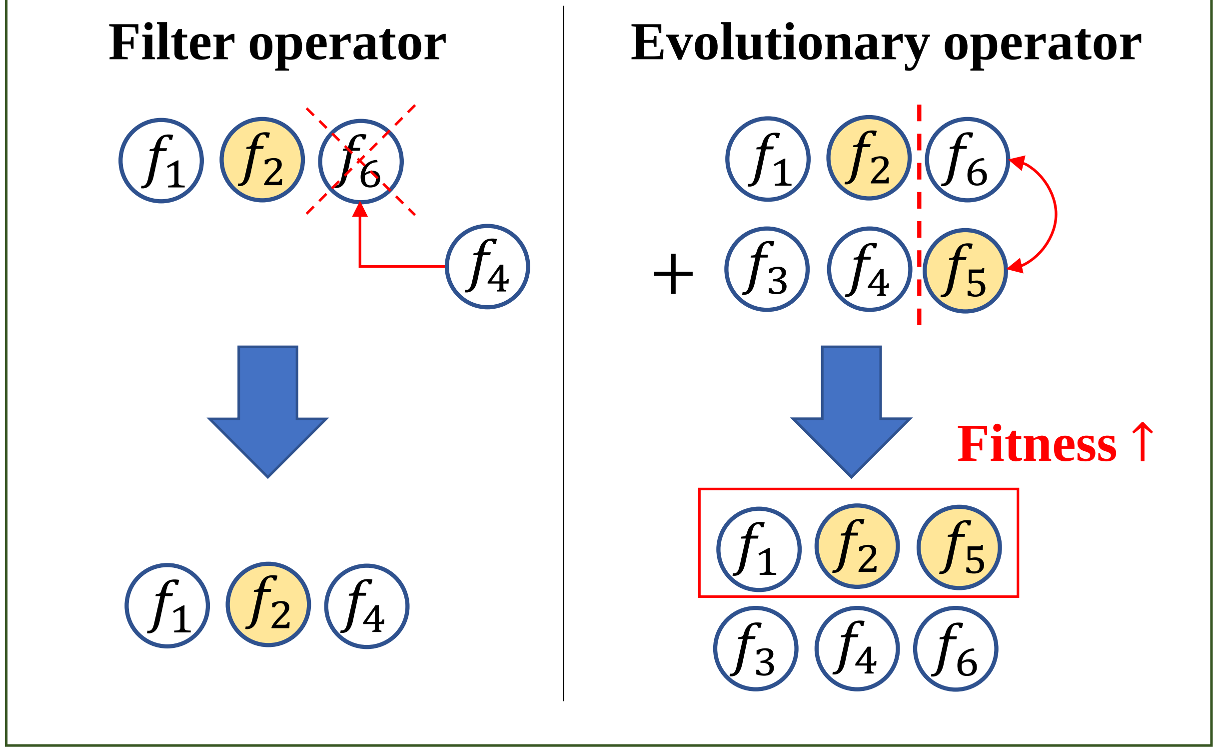
[10] Lu, Y., Liang, M., Ye, Z., Cao, L. (**2015**). Improved particle swarm optimization algorithm and its application in text feature selection. *Appl. Soft Comput.*, 35:629-636.
[11] Lin, K.C., Zhang, K.Y., Huang, Y.H., Hung, J.C., Yen, N. (**2016**). Feature selection based on an improved cat swarm optimization algorithm for big data classification. *J. Supercomput.*, 72:3210-3221.
[12] Lim, H., Lee, J., Kim, D.W. (**2017**). Optimization approach for feature selection in multi-label classification. *Pattern Recognit. Lett.*, 89:25-30.
[13] Mafarja, M.M., Mirjalili, S. (**2018**). Hybrid binary ant lion optimizer with rough set and approximate entropy reducts for feature selection. *Soft Comput.*, 1-17.
[14] Gu, S., Cheng, R., Jin, Y. (**2018**). Feature selection for high-dimensional classification using a competitive swarm optimizer. *Soft Comput.*, 22:811-822.
[15] Zhang, M.L., Peña, J.M., Robles, V. (**2009**). Feature selection for multi-label naive Bayes classification. *Inf. Sci.*, 179:3218-3229.

Appendix : example

- **Assumption : Rankings of features by filter**

$f_2 > f_1 > f_4 > f_3 > f_5 > f_6$ ← ***damn*** > ***go but*** > ***Japan*** > ***my***

Example 1) Feature subset f_1 f_2 f_6



Example 2) Feature subset f_3 f_4 f_6