Hybrid Feature Selection Technique for Multi-label Text Mining

Injun Yu

Department of Computer Science & Engineering
The Graduate School, Chung-Ang University

INDEX

01 Introduction

02 Related Works

03 Proposed Method

04 Experimental Results

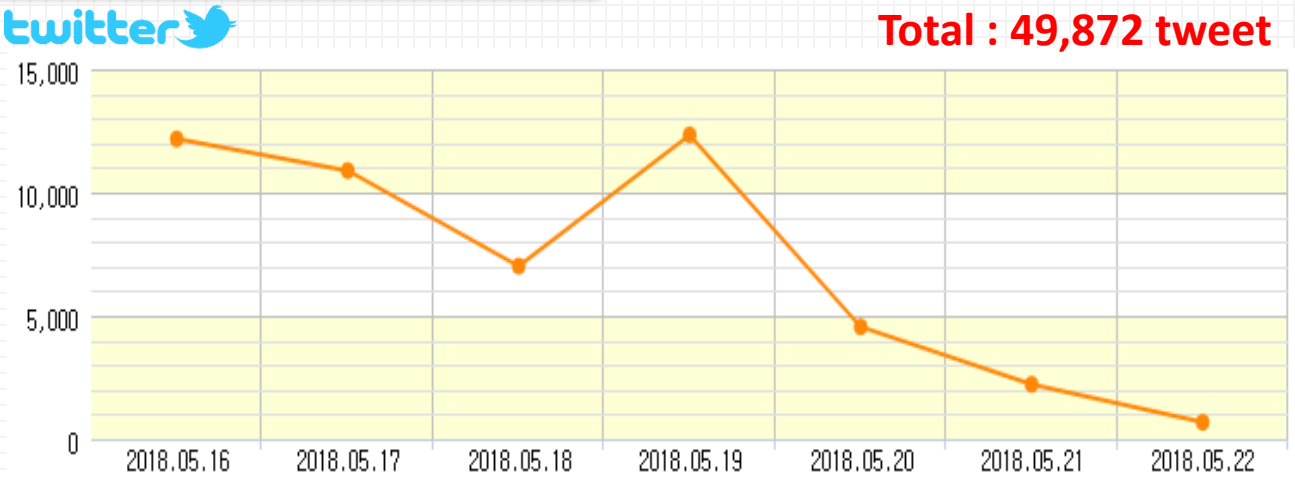
05 Conclusion

01 INTRODUCTION

- Text data is exploding on internet because of the appearance of SNS, such as Facebook, Twitter, Instagram, etc.
  (Ex. Number of tweets related to the keyword "데드풀 ")

- In text mining area, these text data is used for text applications such as sentiment analysis, spam detection, review recommender and so on.

Twitter : " 데드풀 " – 2018.05.16 ~ 2018.05.22



* http://tweetrend.com/

01 INTRODUCTION

- Text is generally represented as a "the bag of words", which is encoded in the text document whether words is presence or not, or the frequency of the word.
- Words = Terms = Features = Dimensions ⬆
- Many features correspond to rare words can mislead poor classification accuracy.
  (Noise)

Deadpool 2 is a wildly entertaining film that without a doubt audiences should not miss even if they aren't fans of the superhero genre.

When Deadpool 2 earned $53 million, making it the highest opening day ever for an R-rated movie.

Document

Docu.
D1
D2
D3
...

Features (= words)

wildly	entertain	miss	superhero	earn	R-rated	...
1	1	1	1	0	0	
0	0	0	0	1	1	
0	1	0	1	0	1	
...						

Noise

Labels

Review	Superhero	Info.	...
1	1	0	
0	0	1	
1	1	1	
...			

01 INTRODUCTION

- To check the important words associated with given categories(= Label) is necessary task.
- **Feature Selection** ➡ **High Dimensionality** ⬇
  Classification Accuracy ⬆

Deadpool 2 is a wildly entertaining film that without a doubt audiences should not miss even if they aren't fans of the superhero genre.

When Deadpool 2 earned $53 million, making it the highest opening day ever for an R-rated movie.

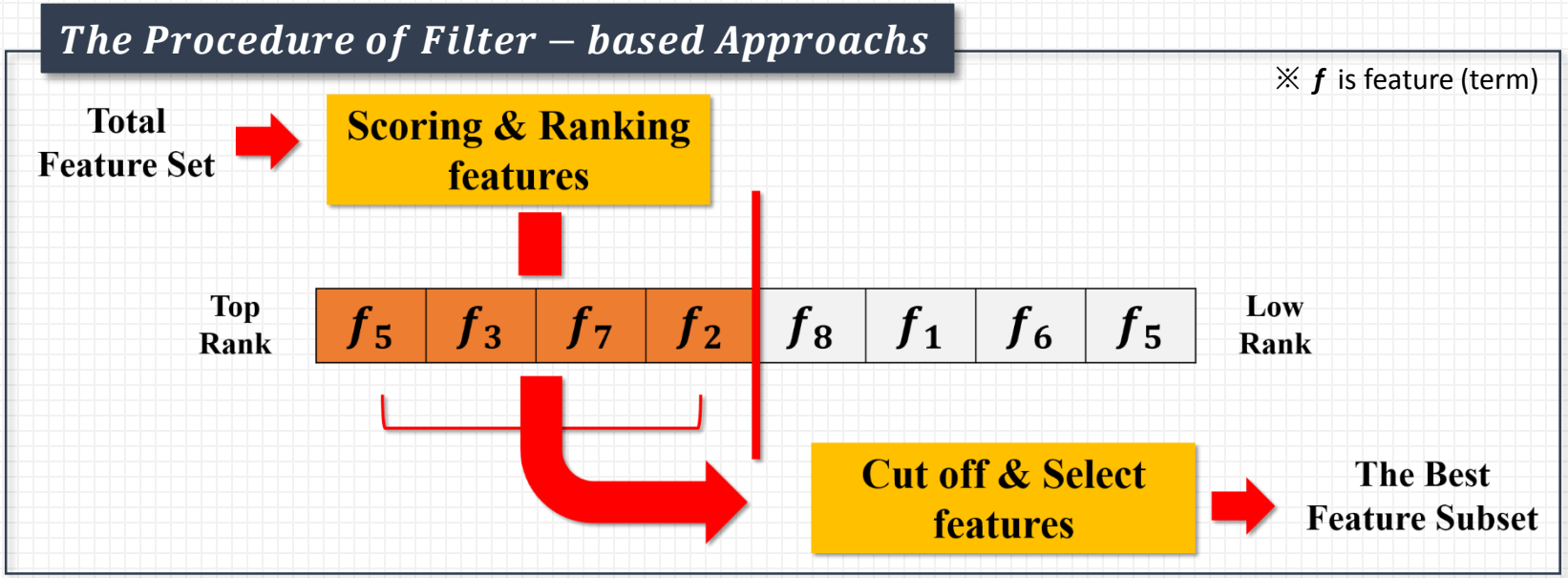
	Features						Labels		
	wildly	entertain	miss	superhero	earn	R-rated	Review	Superhero	Info.
D1	1	1	1	1	0	0	1	1	0
D2	0	0	0	0	1	1	0	0	1
D3	0	1	0	1	0	1	1	1	1

02 RELATED WORKS

Two Approaches of Single-Label Text Feature Selection

① Filter-based Approaches [1,2,3,4]

- **Score function** based on :
  Documents without words **vs** Documents with words
- NDM (2017) [4] : Consider the relative document frequencies.
- VGFSS (2017) [3] : new schema considering the variation of each class based on the distribution of words in classes

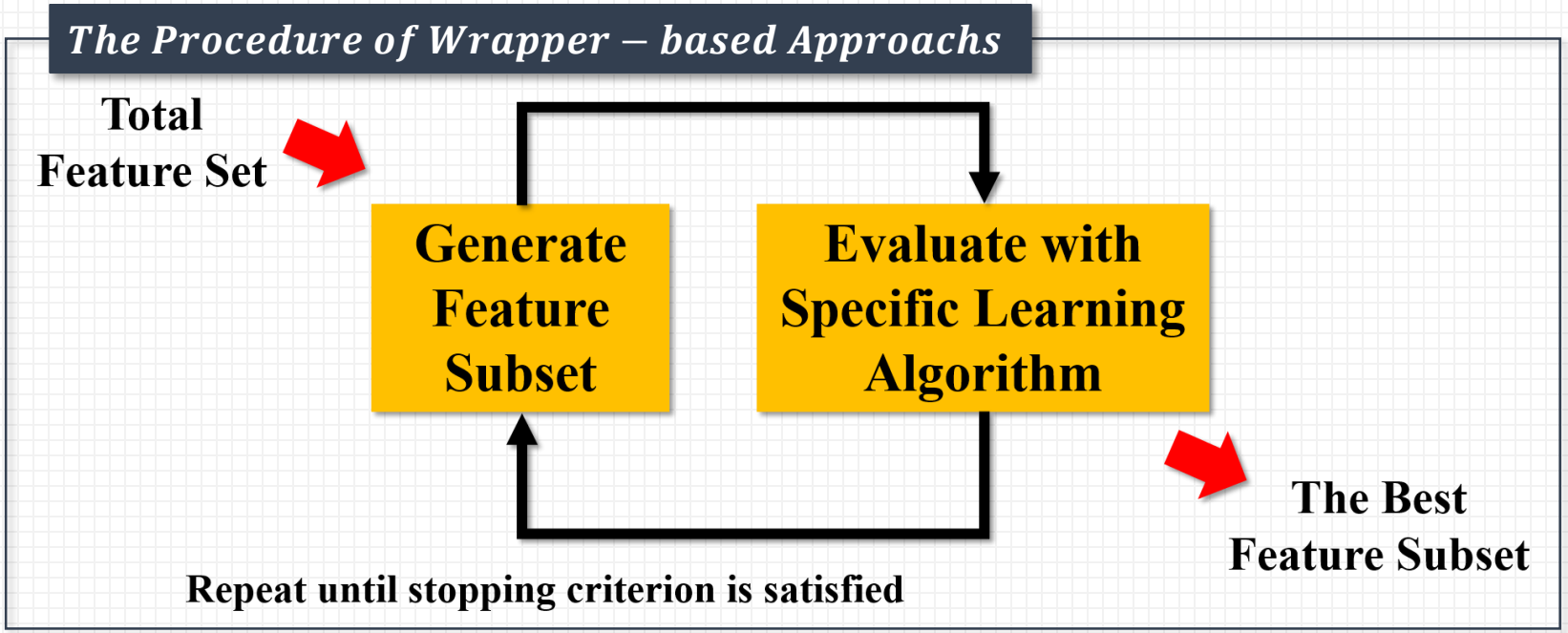


02 RELATED WORKS

Two Approaches of Single-Label Text Feature Selection

② Wrapper-based Approaches [5, 6]

- Considering a Specific Learning Algorithm
- Evolutionary Algorithms (EAs) are widely used such as GA, PSO, ACO.
- EGA (2016) [6] : genetic operators is devised or modified using word or document frequency.

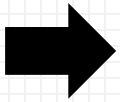


02 RELATED WORKS

Problem Transformation Approaches for Multi-Label Feature Selection

- Filter/Wrapper methods for single-label can't directly apply to multi-label dataset.
- Many existing studies applied Problem Transformation in Multi-Label environment .
- Problem Transformation approaches transform the multi-label problem into a single- label multi-class problem [7, 8, 9].

Instance	Labels
x_1	$\{L_1, L_4\}$
x_2	$\{L_3\}$
x_3	$\{L_1, L_3, L_4\}$
x_4	$\{L_2\}$



Instance	Label
x_1	C_1
x_2	C_3
x_3	C_3
x_4	C_4

OR

Instance	Label
x_1	C_1
x_1	C_4
x_2	C_2
x_3	C_1
x_3	C_3
x_3	C_4
x_4	C_2

The loss of information.
The generation of too many classes.

03 PROPOSED METHOD

The Motivation of The Proposed Method

Limitations of Existing Studies

In Single-Label

- ✓ Filter ➡ It is vulnerable to a specific learning algorithm.
- ✓ Wrapper ➡ Too computationally expensive to converge to the optimal solution.

In Multi-Label ➡ Problem Transformation

Classification Accuracy ⬇

Ideas of the proposed method

- ✓ Combining wrapper (EAs-based search) and filter (new score function).
- ✓ New score function directly deals with multi-label text data without transforming single-label.

Classification Accuracy ⬆

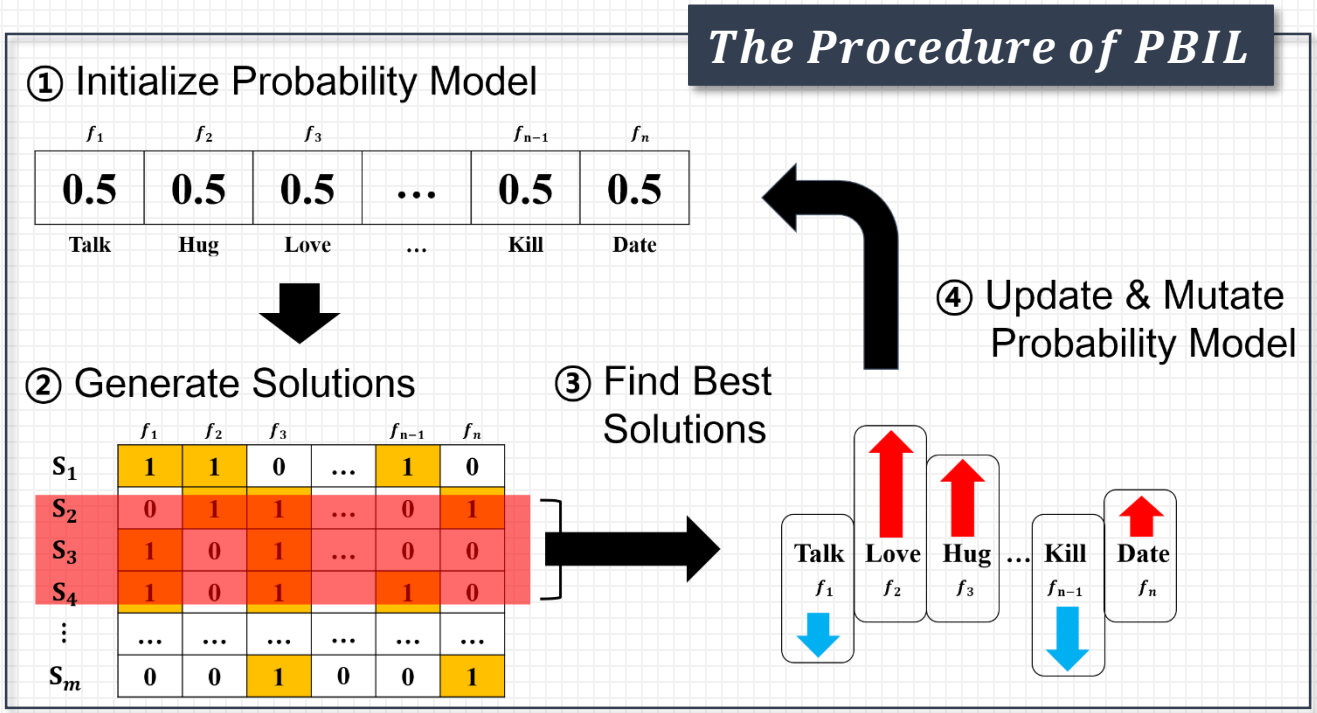
03 PROPOSED METHOD

Evolutionary Search – Population Based Incremental Learning (PBIL)

- PBIL is one of the simplest and most effective evolutionary algorithms using probability model [10, 11, 12].
- It takes too long to find a best feature subset because of the randomness of the initialization.

The initialization need to be guided to generate promising feature subset

New Score Function



03 PROPOSED METHOD

The Extension of Score Function Concept for Multi-Label Text Mining

Score Functions in Single-Label Text [1,2,3,4]

- A word that occurs frequently in a single class and does not occur in the other class is important feature.
- A word that occurs frequently in all classes is not important feature.

➡ **Documents with words VS Documents without words**

Idea of the important words in Multi-Label Text

The important word is a large difference between labels of documents with word w and labels of documents without word w.

➡ **New Score Function : Label Frequency Difference (LFD)**

03 PROPOSED METHOD

Label Frequency Difference (LFD)

- We propose new score function to guide initialization of PBIL considering the influence of multiple labels on a word.

$$LFD(t) = \|lf(t) - lf(\bar{t})\|$$

※ ***lf(t)*** *is the average label frequency of each label given presence of term t.*
※ ***lf*($\bar{t}$)** *is the average label frequency of each label given absence of term t.*
※ $\|\cdot\|$ *is Euclidean distance*

Example

Docu.	Features		Labels			
	Kill	*Talk*	*Action*	*Thriller*	*Drama*	*Romance*
D1	1	1	1	0	0	1
D2	1	0	1	1	1	0
D3	1	1	1	1	0	0
D4	0	1	0	1	1	1
D5	0	0	0	0	1	1
D6	1	0	1	1	0	0
D7	0	1	1	0	1	1

For term t_1(= "***Kill***"),

$lf(t_1) = [\,1.00,\ 0.75,\ 0.25,\ 0.25\,]$,

$lf(\bar{t_1}) = [0.33, 0.00, 1.00, 1.00]$,

$\mathrm{LFD}(t_1) = \|lf(t_1) - lf(\bar{t_1})\| =$ **1.3202**

And for term t_2(= "***Talk***"),

$lf(t_2) = [\,0.75,\ 0.50,\ 0.50,\ 0.75\,]$,

$lf(\bar{t_2}) = [\,0.67,\ 0.67,\ 0.67,\ 0.33\,]$,

$\mathrm{LFD}(t_2) = \|lf(t_2) - lf(\bar{t_2})\| =$ **0.4859**

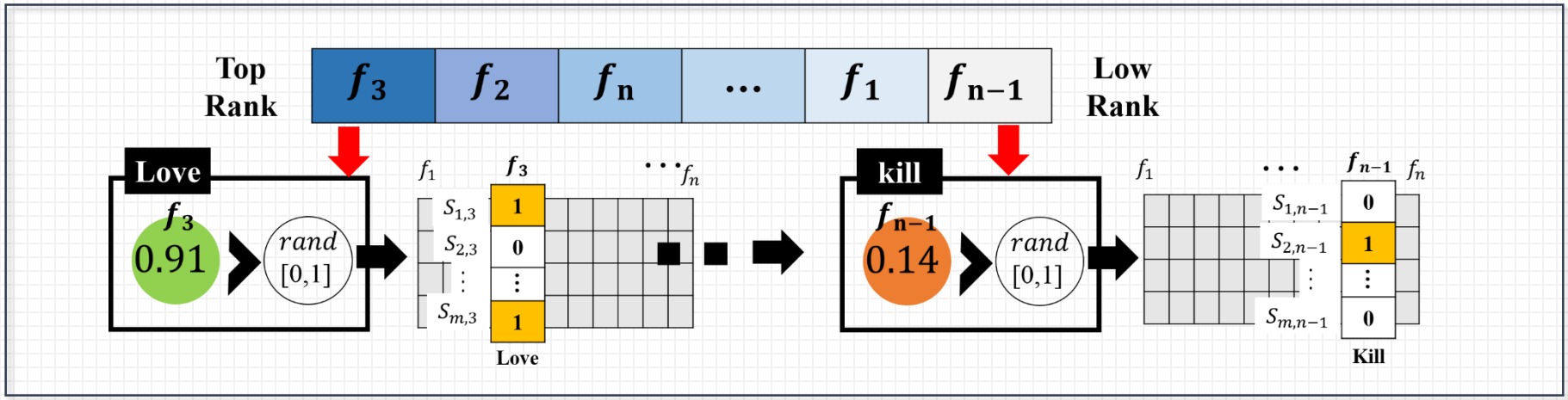
Importance of terms : "***Kill***" > "***Talk***"

03 PROPOSED METHOD

Hybrid Approach using LFD and PBIL

① Initialize the Probability Model using a Normalized LFD between 0 and 1

② Rank the Probability Model and then Generate *m* Solutions in order of high feature ranking



③ Finding Top 50% Best Samples after Evaluating Samples

④ Updating & Mutating Probability Model

➡ Repeat ② - ④until stopping criterion is satisfied

04 EXPERIMENTAL RESULT

Experimental Settings

- **11 Text Datasets with Multi-Label : Yahoo Datasets** [13]
- **Experiment :** Hold-out cross validation (8:2), Repeated 10 times

Dataset	Patterns	Features	Labels	Card.	Den.
Arts	7,484	1,157	26	1.654	0.064
Business	11,214	1,096	30	1.599	0.053
Computers	12,444	1,705	33	1.507	0.046
Education	12,030	1,377	33	1.463	0.044
Entertainment	12,730	1,600	21	1.414	0.067
Health	9,205	1,530	32	1.644	0.051
Recreation	12,828	1,516	22	1.429	0.065
Reference	8,027	1,984	33	1.174	0.036
Science	6,428	1,859	40	1.450	0.036
Social	12,111	2,618	29	1.279	0.033
Society	14,512	1,590	27	1.670	0.062

04 EXPERIMENTAL RESULT

Experimental Settings

- **Comparison Methods :**

  Filter : Apply problem transformation - label powerset

  MD [1], PR [2], VGFSS [3], NDM [4]

  Wrapper : Population size : 50, FFC : 100

  AIPSO [5], EGA [6], CDM+EGA [6]

- **Classifier : Multi-Label Naïve Bayes (MLNB)** [14]

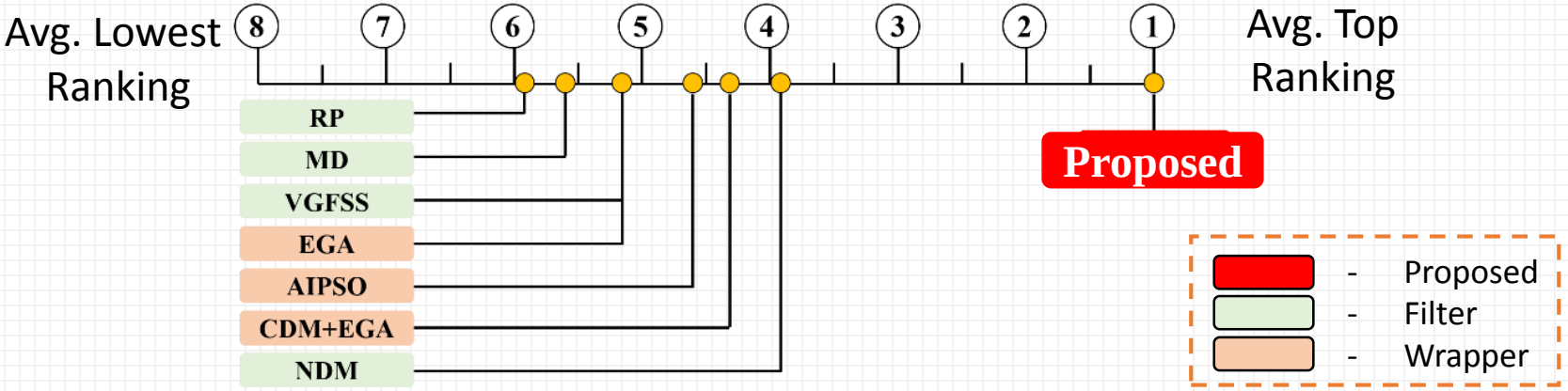
- **Evaluation Measure :**

- **Hamming Loss** : evaluates the fraction of misclassified instance-label pairs.
- **Ranking Loss** : evaluates the fraction of reversely ordered label pairs.
- **Multi-label Accuracy** : evaluates the overall effectiveness of a classifier.
- **Subset accuracy** : evaluates the fraction of correctly classified examples.

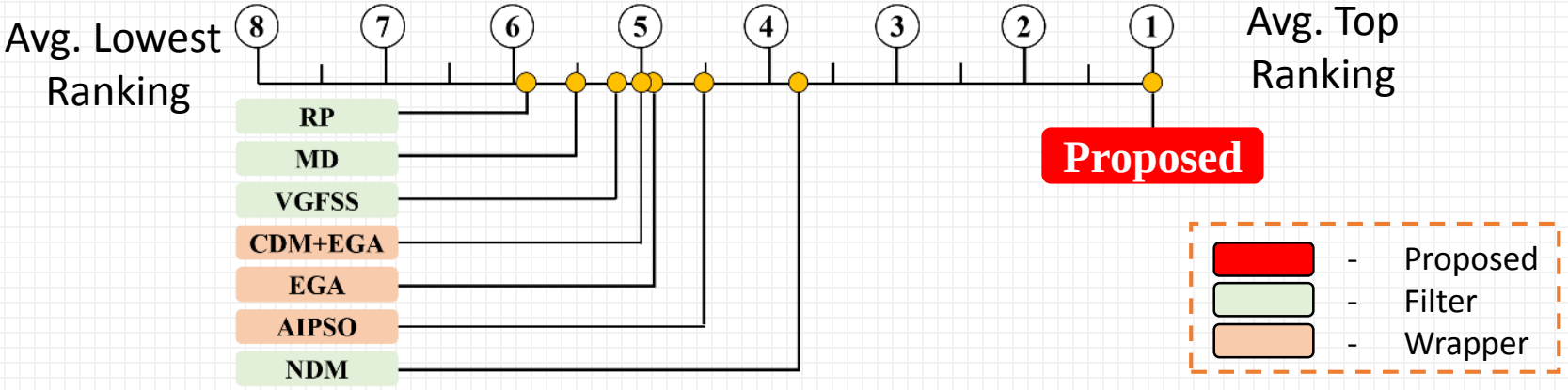
04 EXPERIMENTAL RESULT

Experimental Results in terms of Bonferroni-Dunn test

- **Multi-label accuracy (⬆) : Proposed > Wrapper & Filter**



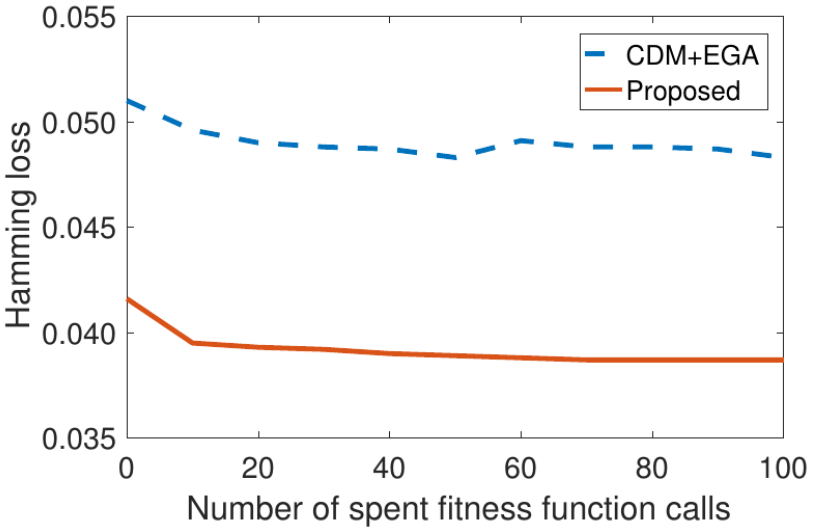
- **Subset accuracy (⬆) : Proposed > Wrapper & Filter**



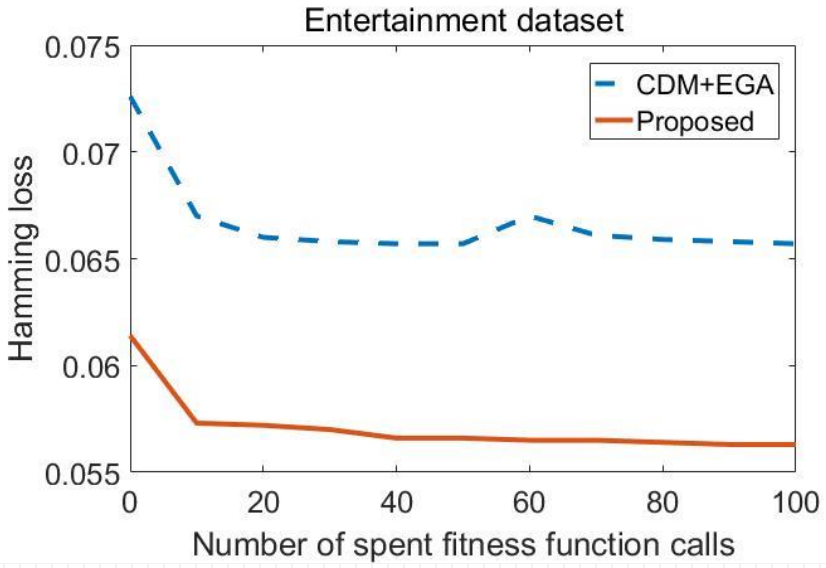
04 EXPERIMENTAL RESULT

Comparison results of convergence according to FFCs

- Comparison results of convergence between CDM+EGA and the proposed method in terms of Hamming loss (⬇) to analyze the performance of local search.
- CDM + EGA is a wrapper-based hybrid method that combines the local search method CDM (filter) with the EGA (wrapper) to guide the initialization steps. [6]
- As a result, LFD showed better initialization than CDM and accelerated convergence.



Health dataset



Entertainment dataset

05 CONCLUSION

- We proposed effective hybrid text feature selection technique in multi-label environment.
- To reinforce local search, we measure new score considering the multi-label itself of important word.
- For the best use of LFD, we revise the initialization & generation steps of PBIL
- The experimental results prove a superior performance of the proposed method compared the state-of-the-art methods.

✓ Classification Accuracy ⬆
✓ LFD leads to the better promising feature subset than other local search.
✓ Avg. Ranking of Multi-Label Accuracy & Subset Accuracy:

Proposed > Wrapper & Filter


Thank
You

Reference

[1] Tang, B., Kay, S., and He, H. (2016). Toward optimal feature selection in naive bayes for text categorization. *IEEE transactions on knowledge and data engineering*, 28(9):2508–2521.
[2] Feng, G., An, B., Yang, F., Wang, H., and Zhang, L. (2017). Relevance popularity: A term event model based feature selection scheme for text classification. *PloS one*, 12(4):1–15.
[3] Agnihotri, D., Verma, K., and Tripathi, P. (2017). Variable global feature selection scheme for automatic classification of text documents. *Expert Systems with Applications*, 81:268–281.
[4] Rehman, A., Javed, K., and Babri, H. A. (2017). Feature selection based on a normalized difference measure for text classification. *Information Processing & Management*, 53(2):473–489.
[5] Lu, Y., Liang, M., Ye, Z., and Cao, L. (2015). Improved particle swarm optimization algorithm and its application in text feature selection. *Applied Soft Computing*, 35:629–636.
[6] Ghareb, A. S., Bakar, A. A., and Hamdan, A. R. (2016). Hybrid feature selection based on enhanced genetic algorithm for text categorization. *Expert Systems with Applications*, 49:31–47.
[7] Pereira, R. B., Plastino, A., Zadrozny, B., and Merschmann, L. H. (2018a). Categorizing feature selection methods for multilabel classification. *Artificial Intelligence Review*, 49(1):57–78.

Reference

[8] Chen, W., Yan, J., Zhang, B., Chen, Z., and Yang, Q. (2007). Document transformation for multi-label feature selection in text categorization. In *7th IEEE International Conference on Data Mining (ICDM 2007)*, pages 451–456.
[9] Spolaˆ oR, N., Cherman, E. A., Monard, M. C., and Lee, H. D. (2013). A comparison of multi-label feature selection methods using the problem transformation approach. *Electronic Notes in Theoretical Computer Science*, 292:135–151.
[10] Baluja, S. (1994). Population-based incremental learning. A method for integrating genetic search based function optimization and competitive learning. Technical report, Carnegie-Mellon Univ Pittsburgh Pa Dept Of Computer Science.
[11] Hauschild, M. and Pelikan, M. (2011). An introduction and survey of estimation of distribution algorithms. *Swarm and evolutionary computation*, 1(3):111–128.
[12] Zangari, M., Santana, R., Mendiburu, A., and Pozo, A. T. R. (2017). Not all PBILs are the same: Unveiling the different learning mechanisms of pbil variants. *Applied Soft Computing*, 53:88–96.
[13] Ueda, N. and Saito, K. (2003). Parametric mixture models for multi-labeled text. In *Advances in neural information processing systems*, pages 737–744.
[14] Zhang, M.-L., Pe˜na, J. M., and Robles, V. (2009). Feature selection for multi-label naive bayes classification. *Information Sciences*, 179(19):3218–3229.