Qiao Yu, Shu-juan Jiang, Rong-cun Wang, Hong-yang Wang, 2017. A feature selection approach based on a similarity measure for software defect prediction. *Frontiers of Information Technology & Electronic Engineering*, **18**(11):1744-1753. https://doi.org/10.1631/FITEE.1601322

A feature selection approach based on a similarity measure for software defect prediction

Key words: Software defect prediction; Feature selection; Similarity measure; Feature weights; Feature ranking list

Corresponding author: Shu-juan Jiang
E-mail: shjjiang@cumt.edu.cn

ORCID: http://orcid.org/0000-0003-0643-0565

Motivation

- Software defect prediction aims to find the potential defects based on historical data and software features.
- Software features could reflect the characteristics of software modules. However, some of these features may be more relevant to the class (defective or non-defective), but others may be redundant or irrelevant.
- Feature selection can select the high correlation features from high-dimensional features. Therefore, introducing feature selection into software defect prediction could not only improve its efficiency, but also improve its accuracy (Miao et al., 2012, Khoshgoftaar et al., 2014, Liu et al., 2014).

Main idea

- In order to fully measure the correlation between different features and the class, this paper presents a feature selection approach based on similarity measure (SM) for software defect prediction.

Method

1. The feature weights are updated according to the similarity of samples in different classes.
2. A feature ranking list is generated by sorting the feature weights in descending order, and all feature subsets are selected from the feature ranking list in sequence.
3. Finally, all feature subsets are evaluated on KNN model and measured by AUC metric for their classification performance.

Major results

- As shown in the line charts in Fig. 2, we can find that SM performs better than RF with different number of features.

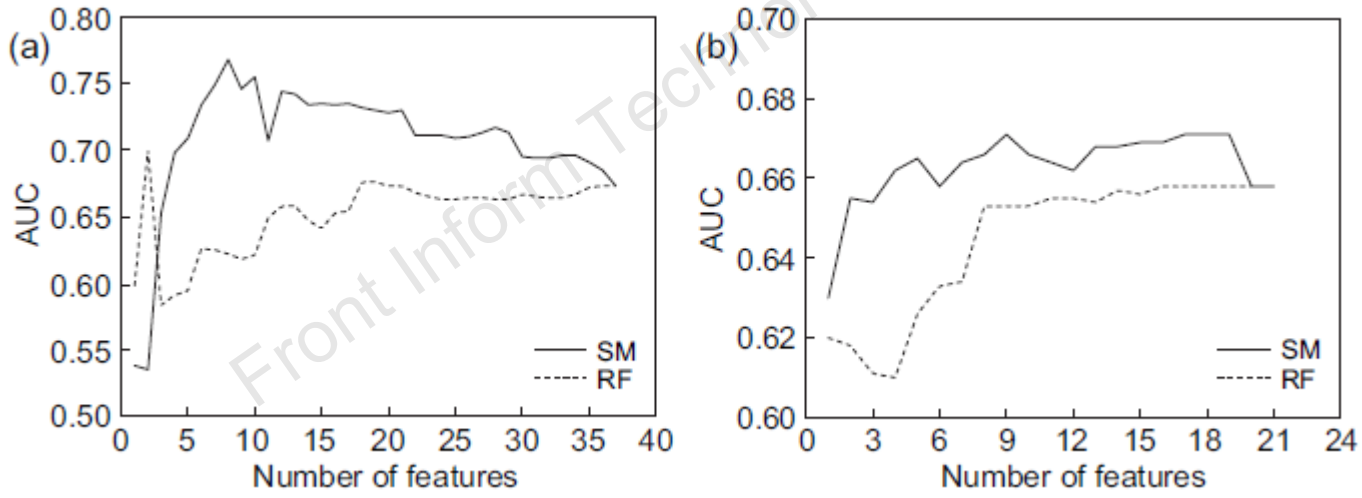

Fig. 2 Comparisons of similarity measure (SM) and ReliefF (RF) with different numbers of features on NASA datasets: (a) CM1 and (b) JM1 (examples)

Major results

- As can be seen from Table 4, we can conclude that SM outperforms OR and RF significantly, and SM is comparable to CL and GR.

Table 4 Win/Tie/Loss results

	SM-OR	SM-RF	SM-CL	SM-GR
Win	10	8	6	3
Tie	1	2	0	4
Loss	0	1	5	4

Conclusions

- This paper presents a feature selection approach based on similarity measure for software defect prediction
- we conduct experiments on 11 NASA datasets and make comparisons with four feature selection approaches.
- The experimental results show that our approach performs better than or is comparable to the compared approaches.