Jinfu CHEN, Xiaoli WANG, Saihua CAI, Jiaping XU, Jingyi CHEN, Haibo CHEN, 2022. A software defect prediction method with metric compensation based on feature selection and transfer learning. *Frontiers of Information Technology & Electronic Engineering*, 23(5):715-731. https://doi.org/10.1631/FITEE.2100468

A software defect prediction method with metric compensation based on feature selection and transfer learning

Key words: Defect prediction; Feature selection; Transfer learning; Metric compensation

Corresponding author: Saihua CAI
E-mail: caisaih@ujs.edu.cn

ORCID: https://orcid.org/0000-0003-0743-1156

Motivation

- Traditional metric compensation methods consider only one-way adaptation of the data domain.
- Traditional metric compensation methods ignore the problem of redundancy of features used for model training.

Main idea

- The bi-directional metric compensation technique based on transfer learning largely reduces data discrepancies between different datasets during model training by adjusting the dataset in both directions.
- Dataset is optimized by filtering the data for irrelevant and redundant features through feature selection.

Method

- A novel metric compensation method is proposed based on feature selection and transfer learning.
- The Pearson feature selection method is used to filter the redundant features that are unrelated to the defect category; a combination of transfer component analysis and bidirectional metric compensation is used to reduce variability in the feature distribution of the source and target projects, to improve the similarity between projects.

Method (Cont'd)

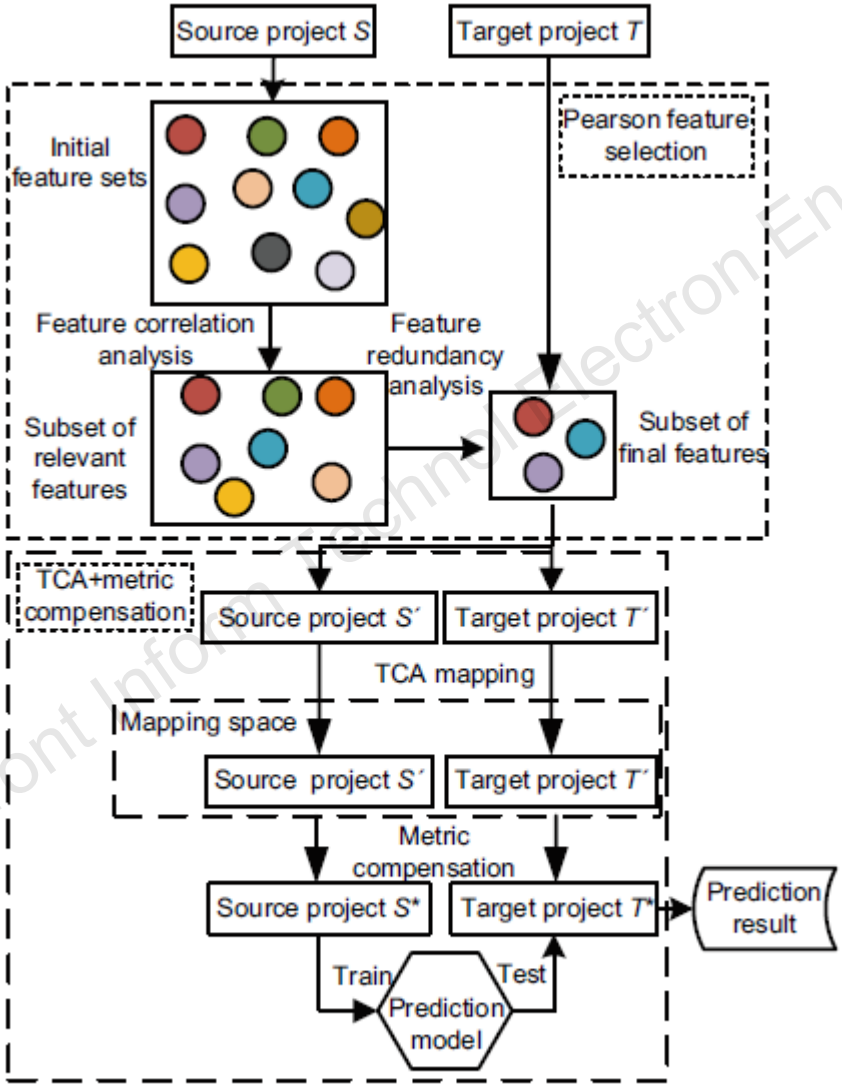

Fig. 1 Overall process of the proposed method for building a model

Method (Cont'd)

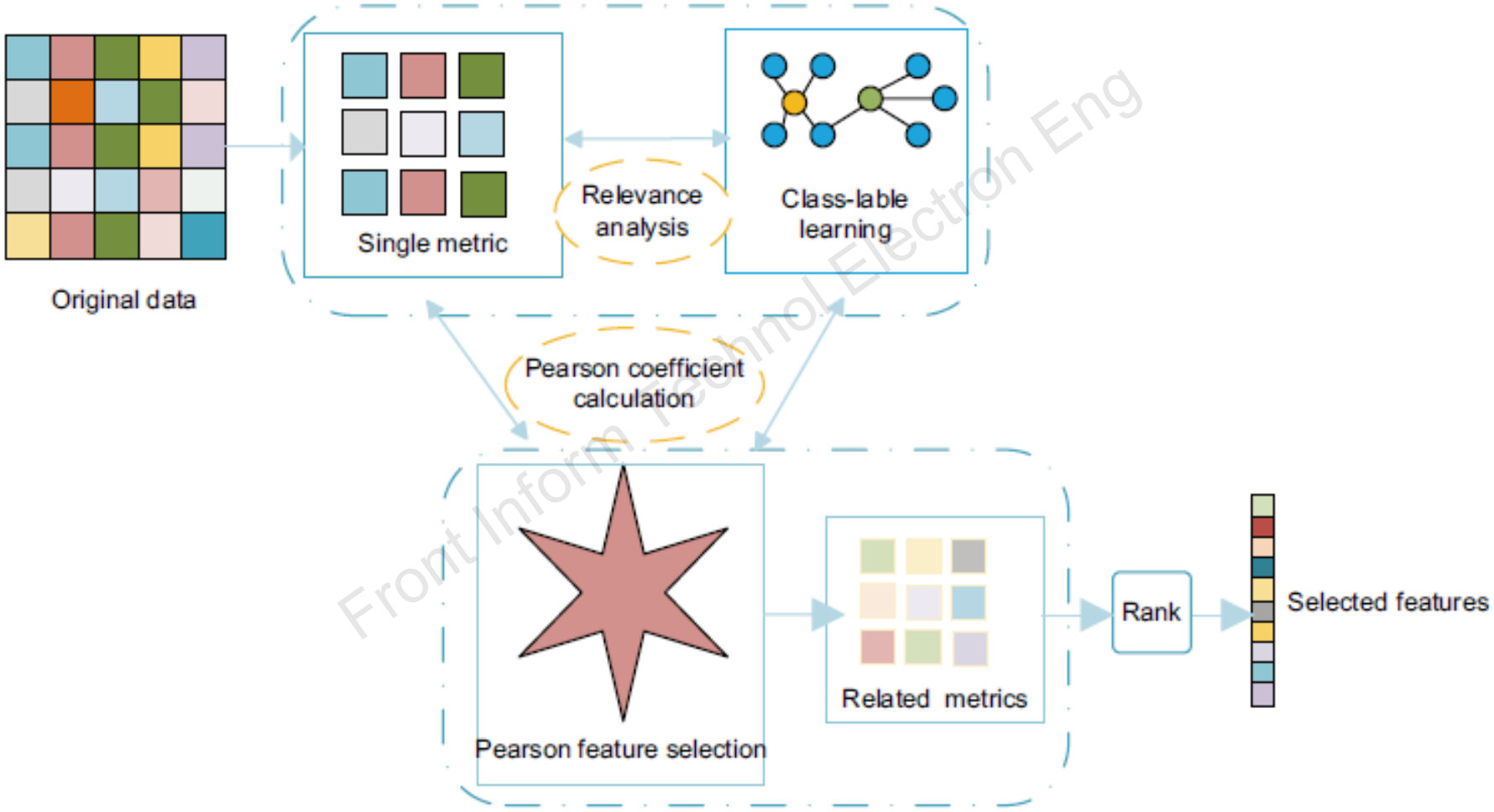

Fig. 2 Overall process of Pearson feature selection

Method (Cont'd)

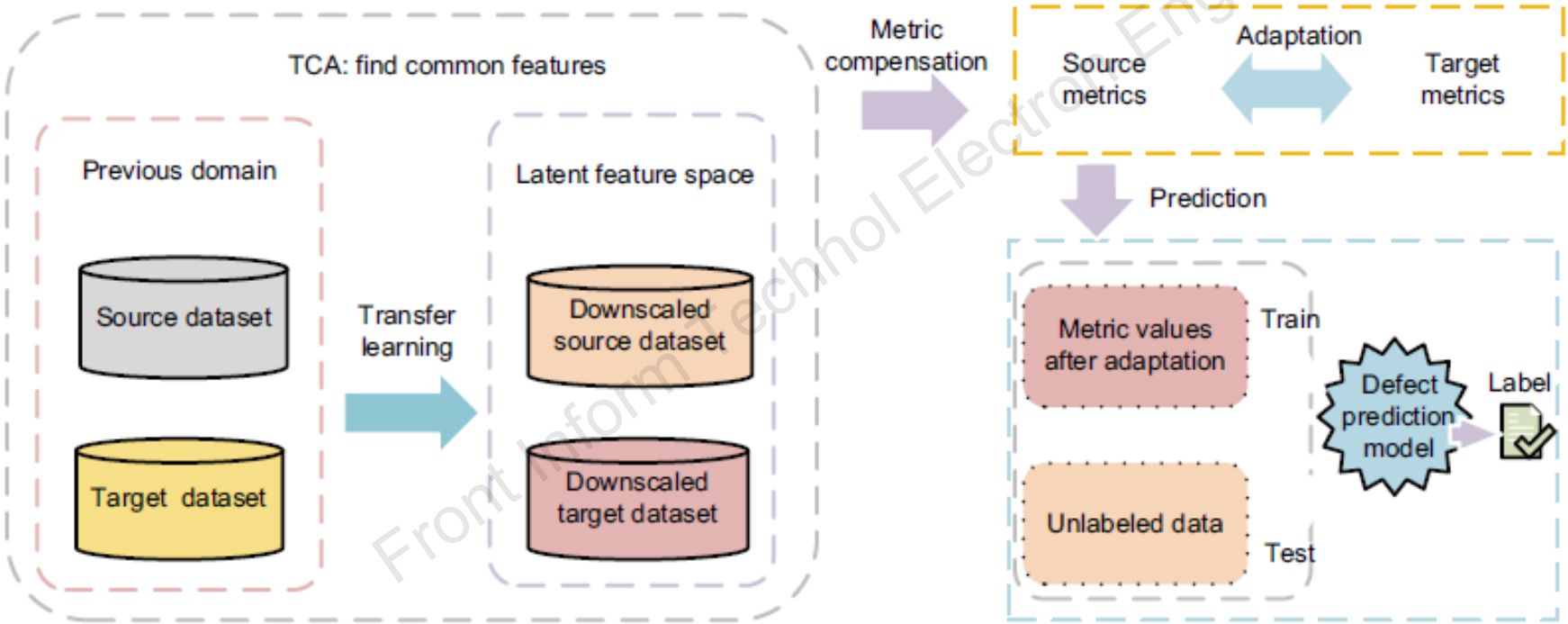

Fig. 3 Overall process of metric compensation based on transfer learning

Major results

Table 4 Comparison results among different methods based on the AEEEM dataset

Source=>target	AUC				F1-measure			
	meCom	meCom17	TCA	peUpMeCom	meCom	meCom17	TCA	peUpMeCom
EQ=>JDT	0.663	0.661	0.572	**0.686**	0.658	0.655	0.410	**0.684**
EQ=>LC	0.665	0.664	0.565	**0.688**	0.682	0.681	0.231	**0.703**
JDT=>PDE	0.647	0.647	0.548	**0.661**	0.616	0.618	0.278	**0.633**
JDT=>ML	0.626	0.626	0.530	**0.635**	0.525	0.523	0.240	**0.541**
PDE=>ML	0.606	0.605	0.534	**0.623**	0.612	0.613	0.226	**0.625**
EQ=>ML	0.594	0.594	0.522	**0.609**	0.560	0.561	0.241	**0.597**

Best results are in bold

Table 5 Comparison results among different methods based on the Relink dataset

Source=>target	AUC				F1-measure			
	meCom	meCom17	TCA	peUpMeCom	meCom	meCom17	TCA	peUpMeCom
Safe=>Apache	0.633	0.634	0.435	**0.656**	0.587	0.590	0.511	**0.628**
Safe=>Zxing	0.553	0.555	0.470	**0.581**	0.411	0.413	0.387	**0.539**
Apache=>Safe	0.649	0.655	0.389	**0.780**	0.591	0.595	0.413	**0.777**
Apache=>Zxing	0.566	0.563	0.528	**0.610**	0.427	0.424	0.362	**0.585**
Zxing=>Safe	0.612	0.613	0.384	**0.678**	0.482	0.486	0.372	**0.674**
Zxing=>Apache	0.588	0.584	0.551	**0.634**	0.487	0.481	0.480	**0.601**

Best results are in bold

Table 6 Comparison results among different methods based on the NASA MDP dataset

Source=>target	AUC				F1-measure			
	meCom	meCom17	TCA	peUpMeCom	meCom	meCom17	TCA	peUpMeCom
CM1=>JM1	0.607	0.606	0.522	**0.625**	0.513	0.516	0.309	**0.556**
KC1=>JM1	0.594	0.592	0.536	**0.613**	0.549	0.552	0.311	**0.575**
KC2=>JM1	0.608	0.609	0.567	**0.628**	0.548	0.544	0.349	**0.586**
PC1=>JM1	0.598	0.598	0.523	**0.614**	0.469	0.463	0.239	**0.518**
CM1=>KC1	0.645	0.647	0.525	**0.665**	0.582	0.599	0.275	**0.639**
KC2=>KC1	0.687	0.677	0.598	**0.697**	0.672	0.634	0.403	**0.678**
PC1=>KC1	0.653	0.654	0.501	**0.676**	0.589	0.596	0.202	**0.643**
CM1=>KC2	0.661	0.667	0.524	**0.702**	0.572	0.583	0.314	**0.648**
CM1=>PC1	0.616	0.621	0.549	**0.646**	0.542	0.560	0.255	**0.605**
KC2=>PC1	0.648	0.622	0.588	**0.653**	0.613	0.573	0.264	**0.621**

Best results are in bold

Major results (Cont'd)

Table 9 Results of the original data and Pearson feature selection based on the AEEEM dataset

Source=>target	AUC		F1-measure	
	Original	Pearson	Original	Pearson
EQ=>JDT	0.639	**0.686**	0.653	**0.684**
EQ=>LC	0.657	**0.688**	0.673	**0.703**
JDT=>PDE	0.646	**0.661**	0.615	**0.633**
JDT=>ML	0.605	**0.635**	**0.596**	0.541
PDE=>ML	0.596	**0.623**	0.587	**0.625**
EQ=>ML	0.582	**0.609**	0.594	**0.597**

Best results are in bold

Table 10 Results of the original data and Pearson feature selection based on the NASA MDP dataset

Source=>target	AUC		F1-measure	
	Original	Pearson	Original	Pearson
CM1=>JM1	0.570	**0.625**	0.453	**0.556**
KC1=>JM1	0.545	**0.613**	0.517	**0.575**
KC2=>JM1	0.585	**0.628**	0.501	**0.586**
PC1=>JM1	0.564	**0.614**	0.405	**0.518**
CM1=>KC1	0.637	**0.665**	0.557	**0.639**
KC2=>KC1	0.617	**0.697**	0.554	**0.678**
PC1=>KC1	0.636	**0.676**	0.547	**0.643**
CM1=>KC2	0.676	**0.702**	0.608	**0.648**
CM1=>PC1	0.609	**0.646**	0.517	**0.605**
KC2=>PC1	0.621	**0.653**	0.563	**0.621**

Best results are in bold

Major results (Cont'd)

Table 12 AUC obtained from the upMeCom and other metric compensation methods

Target project	AUC		
	meCom	meCom17	upMeCom
CM1	0.573	0.569	**0.612**
JM1	0.555	0.542	**0.561**
KC1	0.578	0.543	**0.628**
KC2	0.609	0.601	**0.671**
PC1	0.599	0.601	**0.615**
EQ	0.664	0.680	**0.695**
JDT	0.597	0.619	**0.644**
LC	0.631	0.642	**0.671**
ML	0.582	0.588	**0.608**
PDE	0.591	0.588	**0.618**
Safe	0.685	0.683	**0.730**
Zxing	0.588	0.572	**0.594**
Apache	0.619	0.612	**0.681**

Best results are in bold

Conclusions

- Metric compensation based on feature selection and transfer learning is an improvement of traditional metric compensation methods in handling the data distribution differences between the source project and target project.
- The introduction of the Pearson feature selection technique has effectively improved the prediction accuracy of the defect prediction model.
- A combination of transfer component analysis and bidirectional metric compensation is used to reduce variability in the feature distribution of the source project and target project, thus improving the similarity between projects.

Jinfu CHEN received his PhD degree in Computer Science and Technology from Huazhong University of Science and Technology, Wuhan, China, in 2009. He is currently a full professor with School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China. He is a member of IEEE and ACM, and also a member of the China Computer Federation. His major interests include software testing, software security, and trusted software.

Saihua CAI received his PhD degree from China Agricultural University, Beijing, China, in 2020. He is currently a lecture with School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China. His major interests include abnormal network traffic detection, outlier detection, and software testing.