Tian-bao DU, Guo-hua SHEN, Zhi-qiu HUANG, Yao-shen YU, De-xiang WU, 2020. Automatic traceability link recovery via active learning. *Frontiers of Information Technology & Electronic Engineering,* 21(8):1217-1225. https://doi.org/10.1631/FITEE.1900222

Automatic traceability link recovery via active learning

Key words: Automatic; Traceability link recovery; Manpower; Active learning

Corresponding author: Guo-hua SHEN
E-mail: ghshen@nuaa.edu.cn
ORCID: https://orcid.org/0000-0003-2182-0019

Motivation

1. Traceability link recovery (TLR) is an important and costly software task that requires humans establish relationships between source and target artifact sets within the same project. Previous research has proposed to establish traceability links by machine learning approaches.

2. Current machine learning approaches cannot be well applied to projects without traceability information (links), because training an effective predictive model requires humans label too many traceability links.

3. To save manpower, we propose a new TLR approach based on active learning (AL), which is called the AL-based approach.

Main idea

1. In recent years, some research teams have begun to apply machine learning to TLR. They used existing traceability information to train a classifier and then used this classifier to classify possible traceability links as valid or invalid (i.e., two artifacts are unrelated). Although the accuracy of these approaches is high, creating traceability information can require a lot of manpower, especially for projects without traceability information.

2. To save manpower, we propose a new TLR approach based on active learning. The main difference between TSL- and AL-based approaches is that the TSL-based approach randomly selects traceability links for labeling, while the AL-based approach selects traceability links for labeling based on a sample selection strategy.

Method

To save manpower, we use the AL-based approach to select a small number of representative samples for labeling. The AL-based approach includes mainly the following steps:

- It generates a training set.
- It establishes a set of features for traceability links.
- The training data is balanced by rebalancing.
- The learning engine trains a classifier, which is used to classify unlabeled links.

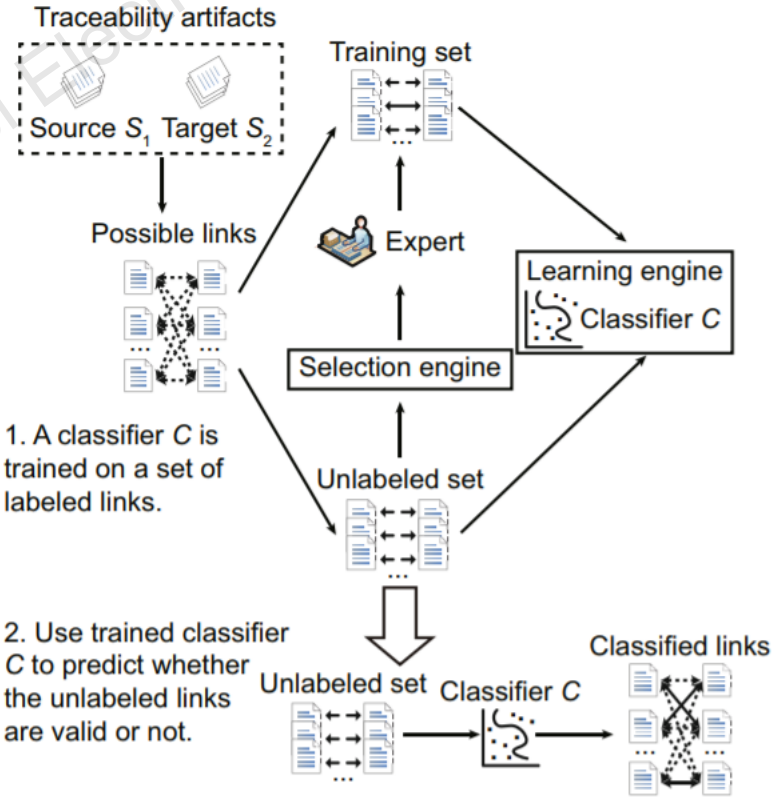

Fig. 1 Overview of the active learning based approach

Method

2. Datasets for the experiments

Table 2 Datasets used in the evaluation

Dataset	Number of invalid links	Number of valid links	Artifact type	
			Source	Target
eAnci	7091	554	UC	CC
SMOS	5656	1044	UC	CC
MODIS	890	41	HighR	LowR
EasyClinic (TC-UC)	1827	63	TC	UC
EasyClinic (TC-CC)	2757	204	TC	CC
EasyClinic (ID-TC)	1177	83	ID	TC
eTour	6363	365	UC	CC
Total	25 761	2354	–	–

HighR: high-level requirements; LowR: low-level requirements; UC: use cases; CC: code classes; ID: interaction diagrams; TC: test cases

Method

3. Active learning process

Algorithm 1 Active learning process

Input: A sample set $D = \{x_1, x_2, \ldots, x_n\}$, a labeled sample set D_l, where D_l is initially empty, and an unlabeled sample set D_u, where $D_\mathrm{u} = D \backslash D_\mathrm{l}$

Output: D_l

```
D_l ← D_l0   // Randomly label a small number of
             // samples to initialize D_l
Train a classifier on D_l
while Termination condition is not met
    Select a sample x_i from D_u
    Experts label the sample x_i
    Add the labeled sample x_i to D_l
    Train the classifier on D_l
end while
Return D_l
```

Method

4. Two types of features for representing the links

- IR-based features
- Query quality (QQ) features

5. Data rebalancing

- To avoid sample imbalance, a synthetic minority oversampling technique (SMOTE, an oversampling method) is used to add minority class samples.

Table 3 Classification algorithms and rebalancing techniques

Category	Variable	Note
Classification algorithms	RF	Classifier that uses a multitude of RF
	Naive Bayes	Naive Bayes classifier using estimator classes
	Logistic	Regression model with a ridge estimator
	SVM	Support vector machine classifier
Rebalancing techniques	SMOTE	Adding minority class samples
	Undersampling	Reducing majority class samples
	None	No rebalancing technique is applied

Major results

1. Determining the best configuration of the AL-based approach

Table 4 Average F-score achieved by the implementation of the AL-based approach across all datasets

Rebalancing technique	Average F-score (%)			
	RF	Naive Bayes	Logistic	SVM
None	76.02	37.66*	46.00*	40.52*
SMOTE	**79.41**	35.02*	51.78*	49.25*
Undersampling	53.66*	32.46*	37.64*	36.63*

The bold font represents the best configuration. * performs statistically significantly worse than the best configuration (at the 0.05 significance level)

The combination of RF and SMOTE is the best one.

Major results

2. Determining a suitable training set size for the AL-based approach

Table 5 Average F-score achieved by the implementation of the AL-based approach using training sets of different sizes

Dataset	Average F-score (%)					
	AL					IR*
	2%	4%	6%	8%	10%	
eAnci (CC-UC)	24.24	50.07	55.05 (+30.50)	60.02	64.24	24.55
SMOS (CC-UC)	39.75	32.88	39.13 (+12.17)	40.47	42.62	26.96
MODIS (HighR-LowR)	26.77	29.13	38.18 (+11.31)	39.78	40.51	26.87
EasyClinic (TC-UC)	25.57	33.96	59.81 (+16.06)	62.89	75.65	43.75
EasyClinic (TC-CC)	51.97	66.01	76.56 (+29.82)	83.53	90.41	45.74
EasyClinic (ID-TC)	50.42	55.64	64.50 (+16.53)	74.39	81.45	47.97
eTour (CC-UC)	38.98	42.09	46.09 (−6.60)	47.48	47.83	52.69
Average	36.96	44.25	54.19 (+15.83)	58.37	63.24	38.36

* Baseline (IR: information retrieval). The number in the parentheses represents the difference of F-score between the AL- and IR-based approaches

The 6% dataset size is a suitable training set size for the AL-based approach.

Major results

3. Comparing the AL-based approach with the TSL-based approach

Table 6 Precision, recall, and F-score of the AL- and TSL-based approaches

Dataset	Precision (%)		Recall (%)		F-score (%)	
	AL	TSL	AL	TSL	AL	TSL
eAnci (CC-UC)	73.37	50.95	44.05	37.27	55.05 (+12.00)	43.05
SMOS (CC-UC)	57.78	52.39	29.58	24.92	39.13 (+5.35)	33.78
MODIS (HighR-LowR)	72.41	32.04	25.93	30.69	38.18 (+6.83)	31.35
EasyClinic (TC-UC)	95.05	75.44	43.64	28.76	59.81 (+18.16)	41.65
EasyClinic (TC-CC)	89.97	71.45	66.63	40.29	76.56 (+25.03)	51.53
EasyClinic (ID-TC)	87.56	67.79	51.06	25.83	64.50 (+27.09)	37.41
eTour (CC-UC)	68.98	59.40	34.60	18.56	46.09 (+17.81)	28.28
Average	77.87	58.50	41.59	29.48	54.19 (+16.04)	38.15

The number in the parentheses represents the difference of F-score between the AL- and TSL-based approaches

When we use 6% of the dataset as the training set, the AL-based approach outperforms the TSL-based approach in terms of precision, recall, and F-score for each of the seven datasets.

Conclusions

1. An AL-based approach has been proposed to save manpower.
2. We empirically derive the best configuration of the AL-based approach on seven datasets.
3. We choose a suitable training set size (6%) for the AL-based approach.
4. The AL-based approach outperforms the IR-based approach by more than 11% in terms of F-score.