



清华大学
Tsinghua University



THUIR at the NTCIR-15 Micro-activity Retrieval Task

Jiayu Li, Ziyi Ye, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma

Tsinghua University

jy-li20@mails.tsinghua.edu.cn



Understanding Tasks: 2 in 1

Retrieval Task



Query i

Find activity i

Ranking Model



Instance 1

Instance 2

...

Instance n-1

Instance n

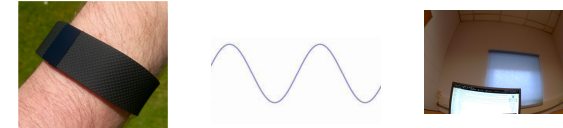
Instance
Ranking List

=

$$P(I_j | Q_i) = \frac{P(Q_i | I_j)P(I_j)}{\sum_{k=1}^n P(Q_i | I_k)P(I_k)} \propto P(Q_i | I_j)$$

Classification Task

Instance j

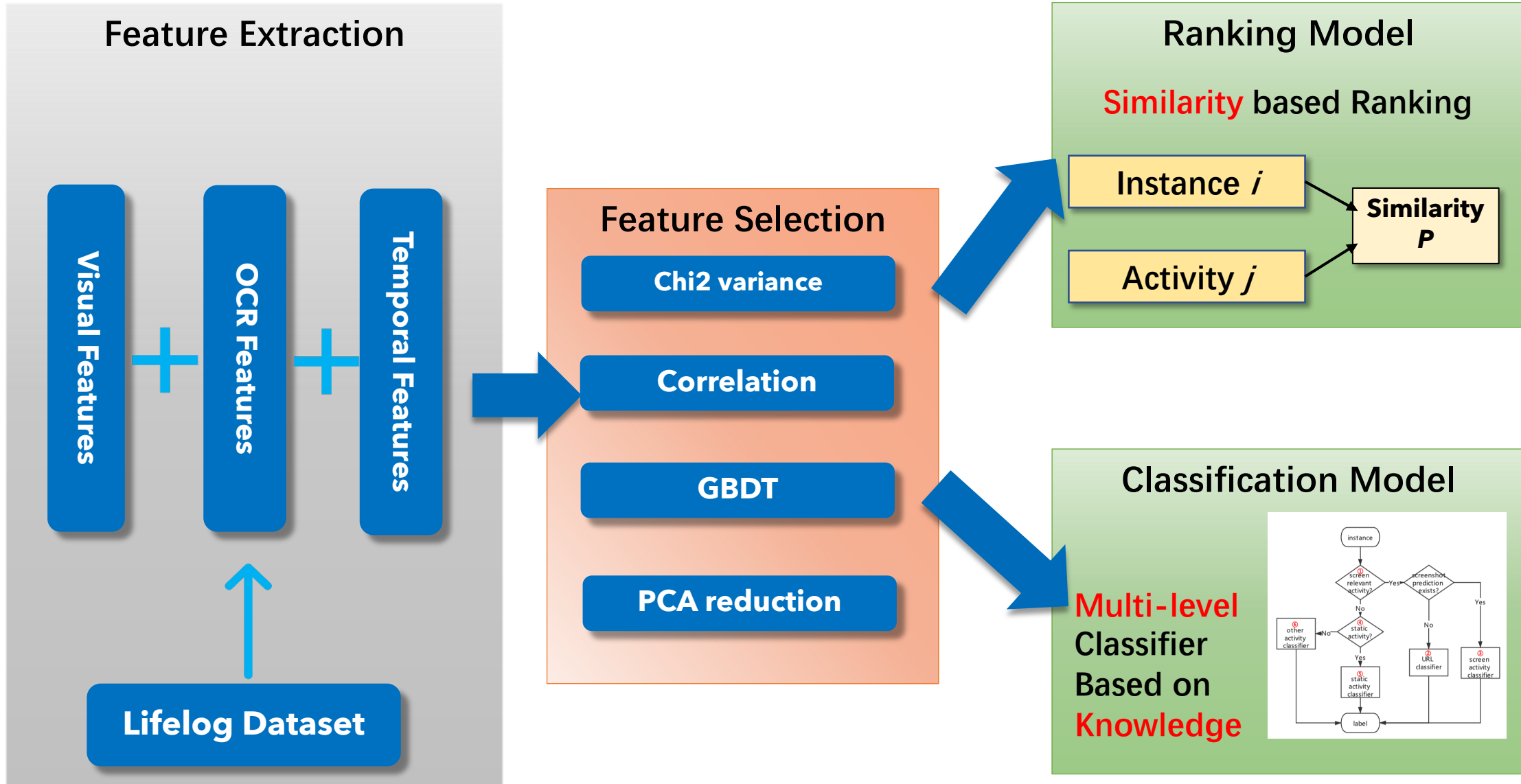


Classifier



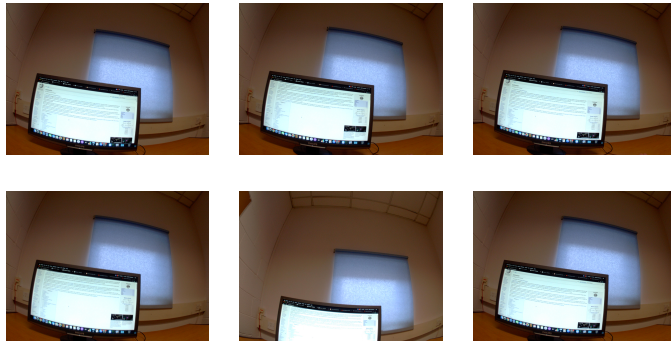
Activity i

Overall Framework



I. Feature Extraction — (1) Visual & OCR Features

Visual



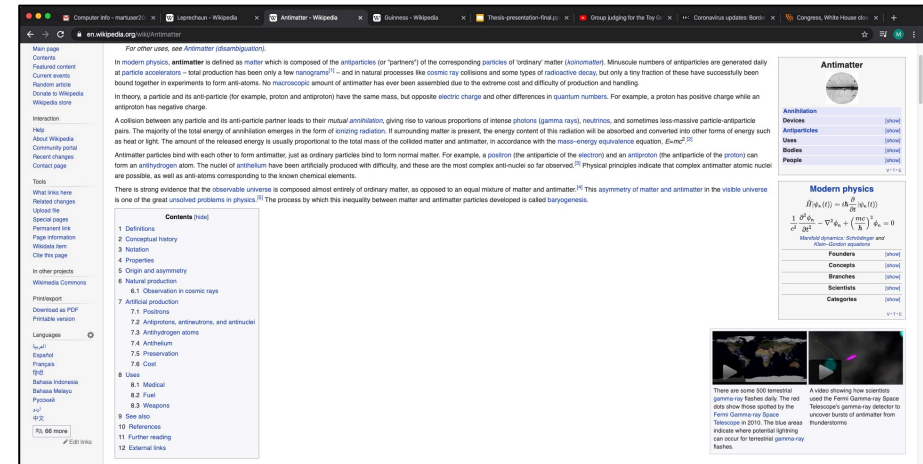
Histogram
Similarity

Pair-wise
Photo Similarity

Object Detection

Computer、Screen
Window、Table
Indoor
...

OCR



URL extraction

OCR detection

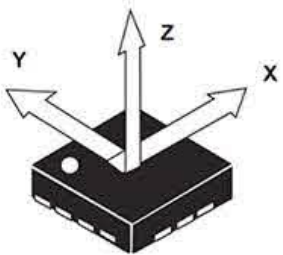
en.Wikipedia.org/wiki/Antimatter

I. Feature Extraction — (2) Temporal Features

Electro-oculography



Acceleration and heart rate



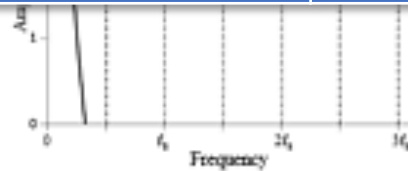
Mouse movements



Feature	Dimension
Photo	4180
Screenshot	1
EOG	62
Acceleration	523
Heart rate	26
Mouse Movement	14
User id	1
All	4806

c values, LPC
nt, number of
peaks

values, energy of
different frequency bands



II. Feature Selection

Methods

Chi2

Correlation

GBDT

Correlation + PCA

Hybrid



Experiment Result

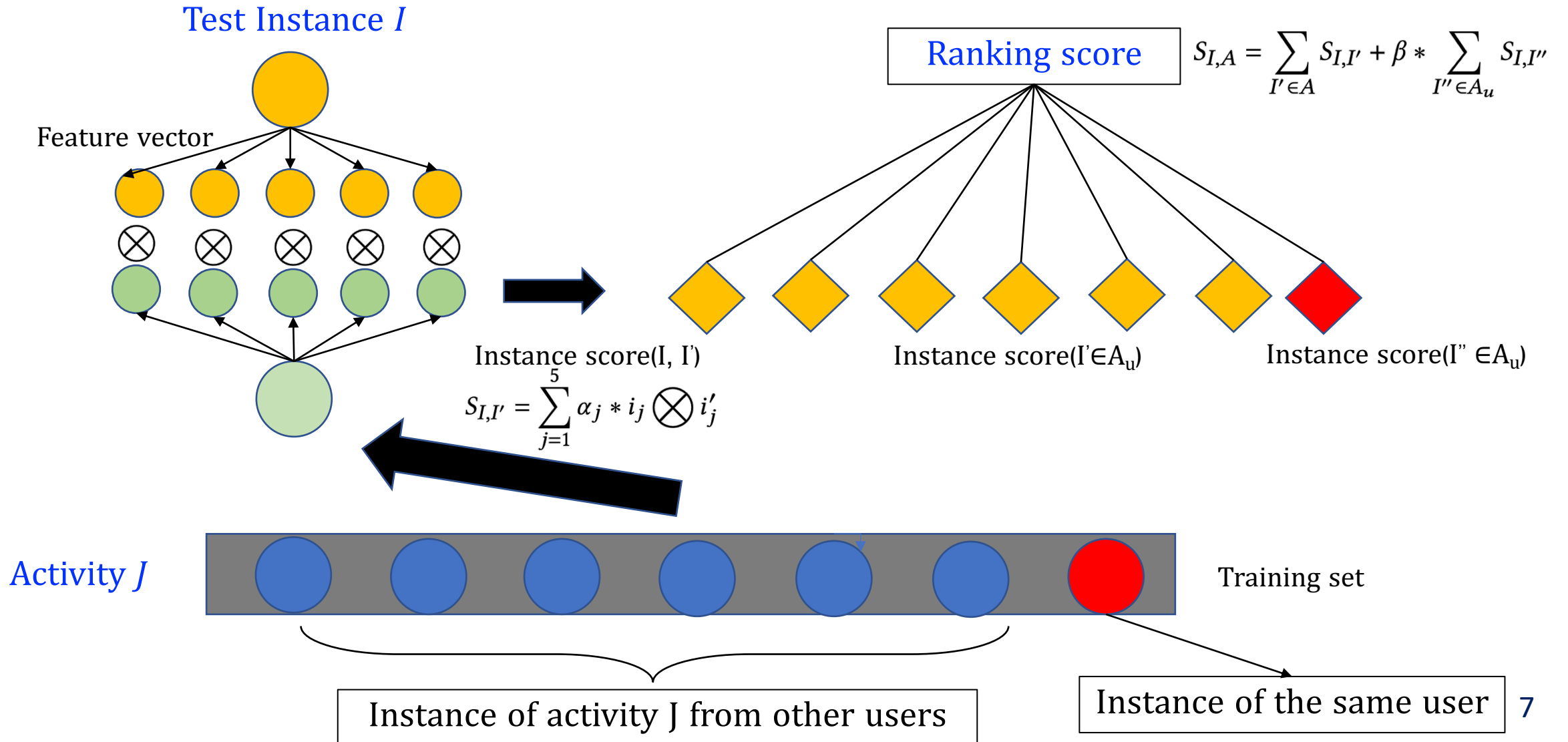
Method	Accuracy	mAP (classify)	mAP (ranking)
None	0.711	0.820	0.879
Chi2	0.879	0.928	0.971
Correlation	0.868	0.923	0.961
GBDT	0.897	0.940	0.973
Hybrid	0.900	0.937	0.972
Correlation + PCA(50)	0.872	0.920	0.960
Correlation + PCA(150)	0.889	0.933	0.974
Correlation + PCA(250)	0.878	0.926	0.973



Comparison of
selected feature set

Methods	Correlation and Chi2	Correlation and GBDT	Chi2 and GBDT
Similarity	0.6219	0.5159	0.5318

III. Ranking Model



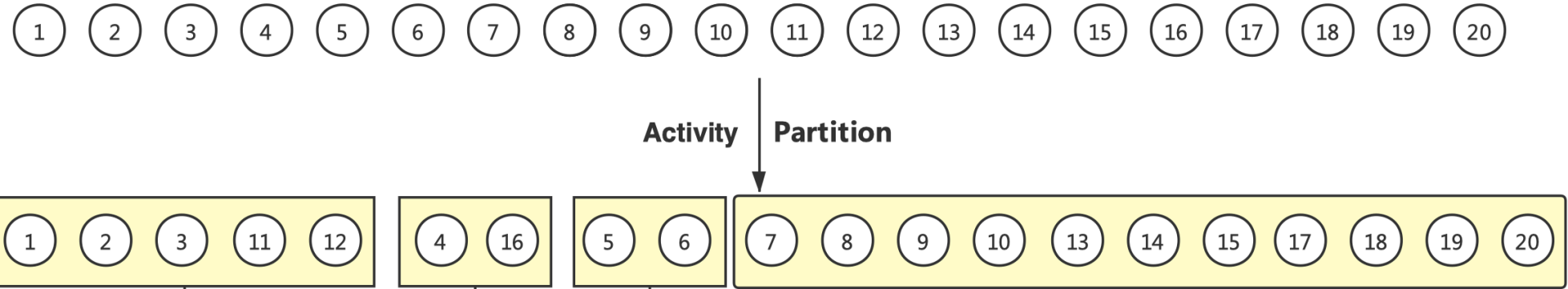
IV. Classification Models — (1)Basic Classifier

		Classifier	Accuracy	mAP(classify)	mAP(ranking)
Parameter based Classifier	{	LR	0.825	0.899	0.898
		SVM	0.821	0.875	0.848
		MLP	0.811	0.890	0.910
Tree + Boosting	{	Random Forest	0.779	0.869	0.927
		XGboost	0.826	0.882	0.921
		GBDT	0.836	0.901	0.947

Multi-level classifier of tree structure

Tree structure classifiers perform better at ranking task.

IV. Classification Models — (2)Rules Detection



Partition Method	Group 1	Group 2	Group 3	Group 4
Impurity	1,2,3,11,12	4, 16	others	-
Similarity	1,3	2, 11, 12	4, 16	others
Cluster	1,3,5,6,8,10	2,4,11,12,16	7,9,13,14,15	17,18,19,20

1, 2, 3, 11, 12 : Screen relevant activities

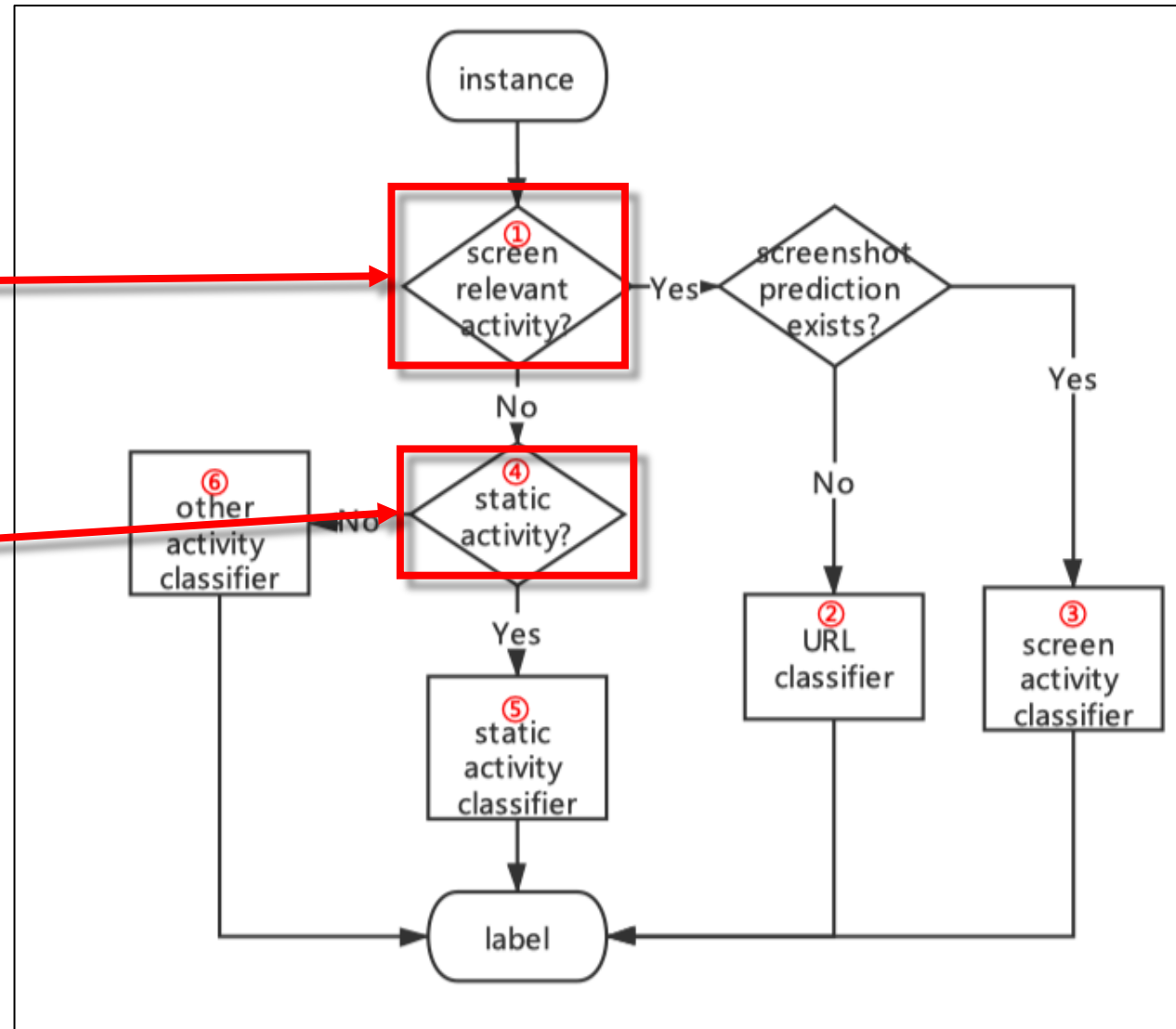
4, 16: Static activities (Zoning out & Close eyes)

IV. Classification Models — (3)Rule-based Classifier

Knowledge rules:

1. Screen-relevant
activities are different
from others.

2. Static activities
(Zoning out & closing
eyes) are special at EOG
and photo features.



1: Binary GBDT
2: 5-label GBDT
3: 5-label GBDT
4: Binary GBDT
5: Binary GBDT
6: 13-label GBDT

V. Results

Classifier	Accuracy	mAP (classify)	mAP (ranking)	Submission results
Basic GBDT Classifier	0.836	0.901	0.947	0.895
Similarity-based method	0.789	0.843	0.836	0.782
Two-level Classifier (Impurity partition)	0.875	0.921	0.971	0.901
Two-level Classifier (Similarity partition)	0.875	0.926	0.970	0.928
Two-level Classifier (Cluster partition)	0.796	0.880	0.931	0.886
Rule-based Classifier	0.889	0.933	0.974	0.950

Similarity based methods:
Repetition of micro activity is unstable.

Rule-based Classifier:
Knowledge about activities helps classification and ranking.

Takeaways

- ✦ Micro-activity is **detectable**
- ✦ Activity retrieval task is **equivalent** to classification task
- ✦ Performing **Feature selection** is better than nothing
 - ✦ Any methods for feature selection performs better than origin features.
- ✦ **Multi-level classifier is effective for MART**
 - ✦ **Rule detection** based on activity partition does help.
 - ✦ Decision tree and tree structure methods are efficient.



Team Name: **THUIR**

Thanks



清华大学
Tsinghua University



THANKS FOR YOUR ATTENTION !
Any Questions?

My email: jy-li20@mails.tsinghua.edu.cn