Prevalence of Precision and Recall in ICSE19 Papers

Presenter: Yitong Li

Outline

- Motivation
- Dataset
- What to look for?
- Results
- Conclusion
- Limitations

Motivation

- Improper use of precision/recall would affect research results evaluation
- The tradeoff between precision and recall often makes it difficult to interpret the research results
- I personally find some papers confusing when they use precision and recall as granted without explaining why

Review of concept

- Precision = TP / (TP + FP)
  - the percentage of the tool-returned answers that are correct*
- Recall = TP / (TP + FN)
  - the percentage of the correct answers that the tool returns*

* https://cs.uwaterloo.ca/~dberry/ATRE/Slides/RvsPpanelTalk/ExpandedRvsPpanelSlides.pdf

Dataset

- ICSE
  - Top conference in SE field
  - Evaluation metrics used in the papers there would be used in future submissions
- ICSE 2019 main conference:
  - 315 papers*
    - Companion papers
    - SEIP, SEET, SEIS, NIER
    - Other technical papers

*unofficial number - got this number by counting the papers in conference proceedings

What to look for?

- RQ1: How many of these papers use “precision” or “recall” as evaluation metric?
- RQ2: What are some topics that use “precision” and “recall”?
- RQ3: Are the papers using “precision” and “recall” in sensible ways?

Results - RQ1

RQ1: How many of these papers use “precision” or “recall” as evaluation metric?

- Keyword filtering:
  - Take any papers that mention “precision” OR “recall”
  - 99/315 papers have the keywords
- Manually checked returned papers:
  - 39 True Positives (i.e. the paper actually used precision/recall as evaluation metric)
  - 60 False Positives (i.e. precision/recall are used in their other meanings)
  - Too many papers to look at to get True Negatives and False Negatives...

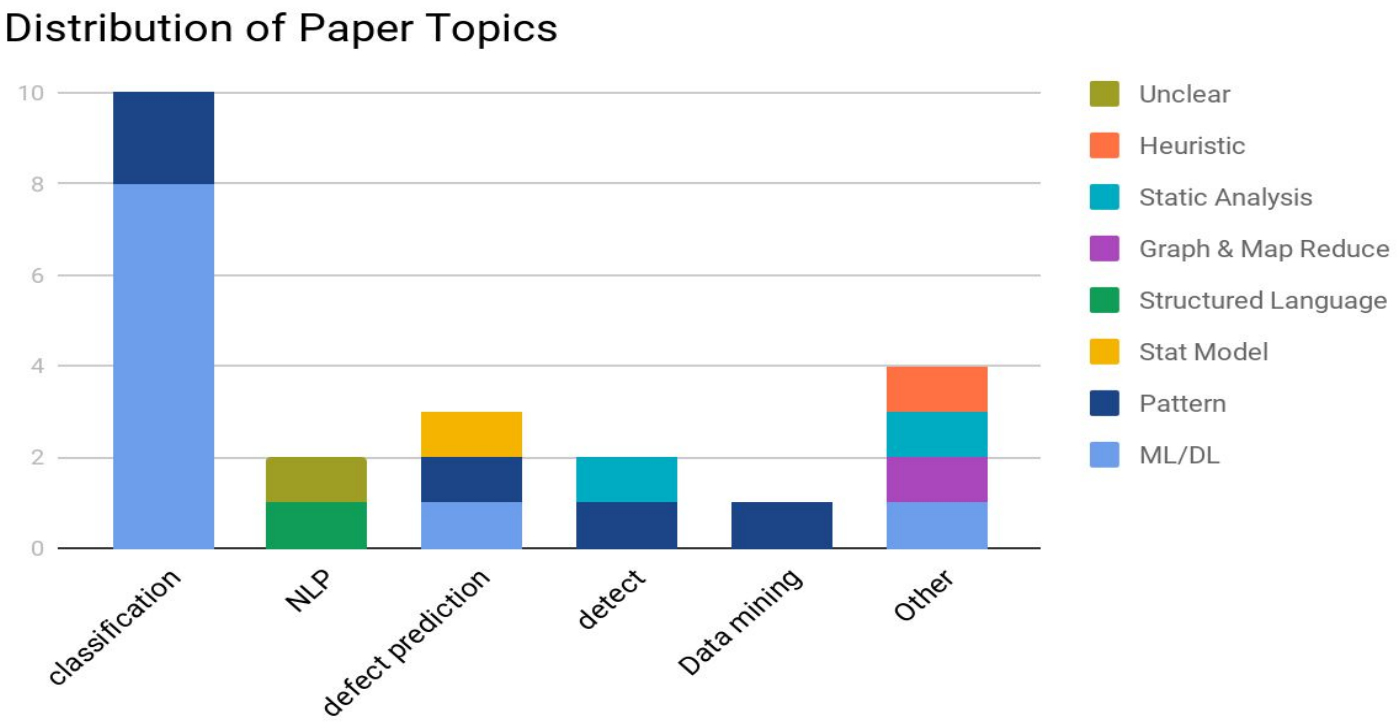
Results - RQ2*

RQ2: What are some topics that use "precision" and "recall"?

- *Defect/vulnerability predictions:*
  - "LEOPARD: Identifying Vulnerable Code for Vulnerability Assessment through Program Metrics"
  - "A System Identification based Oracle for Control-CPS Software Fault Localization"
  - "Class Imbalance Evolution and Verification Latency in Just-in-Time Software Defect Prediction"
- *Classification*
  - "PIVOT: Learning API-Device Correlations to Facilitate Android Compatibility Issue Detection"
  - "Supporting Analysts by Dynamic Extraction and Classification of Requirements-Related Knowledge"
  - "Pattern-Based Mining of Opinions in Q&A Websites"
  - "DLFinder: Characterizing and Detecting Duplicate Logging Code Smells"
  - "NL2Type: Inferring JavaScript Function Types from Natural Language Information"
  - ...
- *NLP*
  - "Automatically Generating Precise Oracles from Structured Natural Language Specifications"

* results from 22 papers I've looked at so far (the rest are in progress)

Results - RQ2


Distribution of Paper Topics
10
8
6
4
2
0
classification
NLP
defect prediction
detect
Data mining
Other
Unclear
Heuristic
Static Analysis
Graph & Map Reduce
Structured Language
Stat Model
Pattern
ML/DL

Results - RQ3

RQ3: Are the papers using “precision” and “recall” in sensible ways?

- Report both precision and recall
  - often compare proposed technique to a baseline and report better results in both “precision” and “recall”
- Emphasize on either “precision” or “recall”
  - one is more important than the other in the application domain

Results - RQ3

- Examples: Report both precision and recall
  - "NL2Type: Inferring JavaScript Function Types from Natural Language Information"
    - Using ML on Natural Language comments to predict javascript function return type
    - Reported better precision and recall compared to previous work
    - In my opinion: both precision are recall are important; any false positives or false negatives would cause incorrect type prediction and the type mismatch would cause issues
  - "DLFinder: Characterizing and Detecting Duplicate Logging Code Smells"
    - Manually crafted patterns for duplicate logging code smells and detect potential problematic logging statements using these patterns
    - Reported both precision and recall for the proposed technique (no comparison)
    - 100% precision and recall for 2 patterns but low precision and high recall for the other
    - In my opinion: low FNs => less code smells being missed; however, with very low precision, high recall indicates the majority of possible results were returned

Results - RQ3

- Examples: Report only precision/recall
  - "LEOPARD: Identifying Vulnerable Code for Vulnerability Assessment through Program Metrics"
    - Use program complexity metric, function metric, etc. to rank potential vulnerable code
    - Use recall as one of the evaluation metric; precision was not used; compared recall with previous work
    - In my opinion: for security domain, it's important to have (or close to) 100% recall. It makes sense for the paper to only use recall for evaluation.
  - "PIVOT: Learning API-Device Correlations to Facilitate Android Compatibility Issue Detection"
    - Automatically extract API-Device correlations to detect API-device incompatibility for android
    - Use only precision as evaluation metric (no explanation on why)
    - In my opinion: high precision => low FPs => less work for developers to look at reported incompatibilities

Conclusion

- Precision and recall are less often used than I expected in ICSE 19
  - 39/315 = 12%
- With the increasing popularity of ML/DL and NLP techniques, more papers would likely use precision and recall to evaluate their results
- Provide justification on why using precision/recall would make the evaluation results more clear/convincing (IMO)

Limitations

- Only checked the papers returned from keyword filtering
  - However, TPs / total # papers would suggest a lower bound on how prevalent “precision” and “recall” are used in ICSE19 papers
- Only looked at ICSE19 papers
  - Results may not generalize to other conference/journal papers
  - However, ICSE is the top conference for SE community. The evaluation metrics used in accepted papers would attract future submissions to use the same
- Manual check error
  - Only one person looked at the papers => human errors are possible

Thank you!