
# Machine-Learning-Guided Selectively Unsound Static Analysis

Kihong Heo
Seoul National University, Seoul, Korea

Hakjoo Oh
Korea University, Seoul, Korea

Kwangkeun Yi
Seoul National University, Seoul, Korea

# Meta Data & Stats

| | |
|---|---|
| **Conference:** | ICSE |
| **Track:** | Program Analysis II |
| **Year:** | 2017 |
| **Number of Authors:** | 3 |
| **Citations:** | 28 |
| **Pages (PDF):** | 11 |
| **Figures:** | 4 |
| **References:** | 24 |
| **Formals:** | 0 definitions |

# What is the Study About?

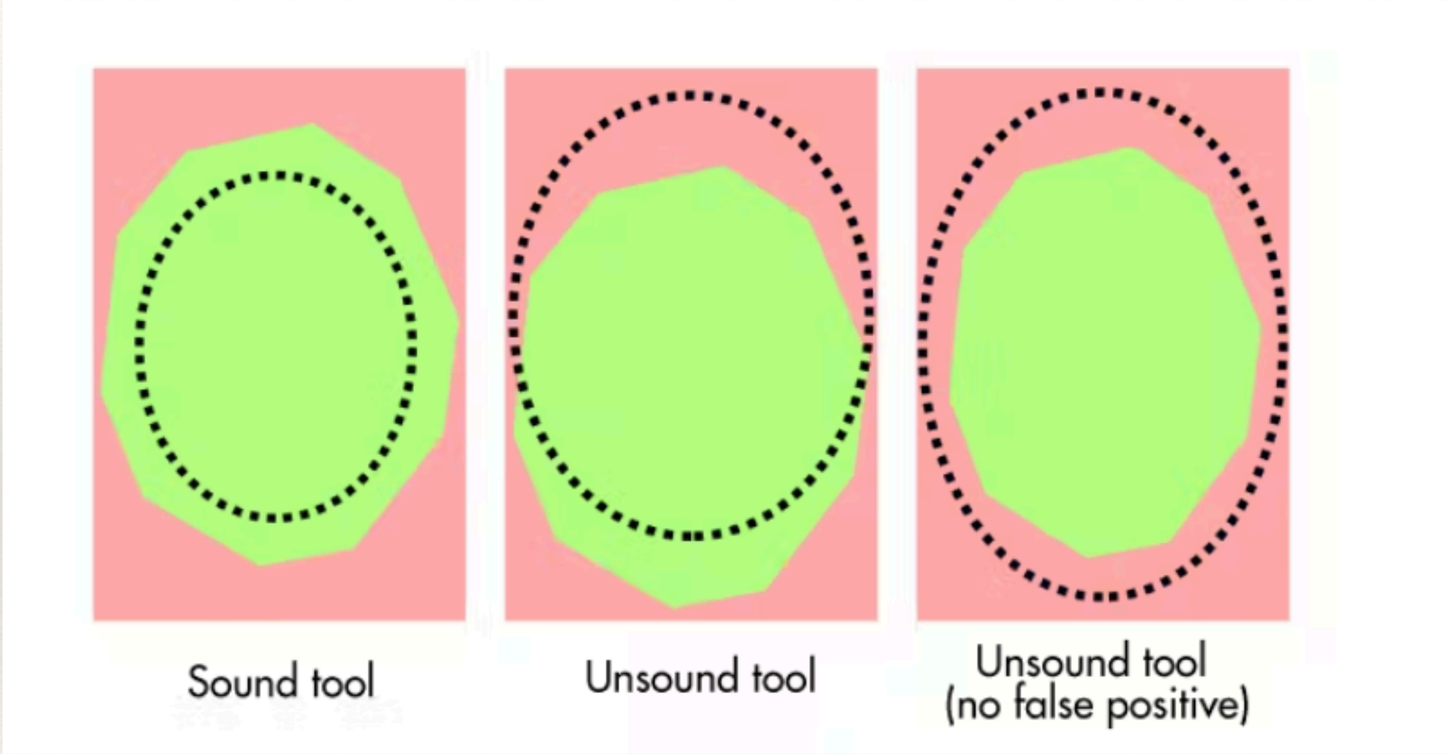
To present a machine-learning-based technique for selectively applying unsoundness in static analysis.

Experiments goals:

- Effectiveness of Approach: How much is the selectively unsound analysis better than the fully sound or fully unsound analyses?
- Efficacy of OC-SVM: Does the one-class classification algorithm outperform two-class classification algorithms?
- Time Cost: How does our technique affect cost of analysis?

# Table of Content

# Sound and Unsound Analysis



# Example

```
str = "hello world";
for(i=0; !str[i]; i++)// buffer access 1
  skip;

size = positive_input();
for(i=0; i<size; i++)
  skip;

... = str[i];           // buffer access 2
```

# Uniformly Sound Analysis

## Example

```
str = "hello world";
for(i=0; !str[i]; i++)// buffer access 1
  skip;

size = positive_input();
for(i=0; i<size; i++)
  skip;

... = str[i];           // buffer access 2
```

F

T

# Uniformly Unsound Analysis

## Example

```
str = "hello world";
for(i=0; !str[i]; i++)// buffer access 1
  skip;

size = positive_input();
for(i=0; i<size; i++)
  skip;

... = str[i];           // buffer access 2
```

## UUA

```
str = "hello world";
i = 0;
if (!str[i])           // buffer access 1   T
  skip;

size = positive_input();
i = 0;
if (i < size)
  skip;

... = str[i];           // buffer access 2  F
```

# Selectively Unsound Analysis

## Example

```
str = "hello world";
for(i=0; !str[i]; i++)// buffer access 1
  skip;

size = positive_input();
for(i=0; i<size; i++)
  skip;

... = str[i];           // buffer access 2
```

## SUA

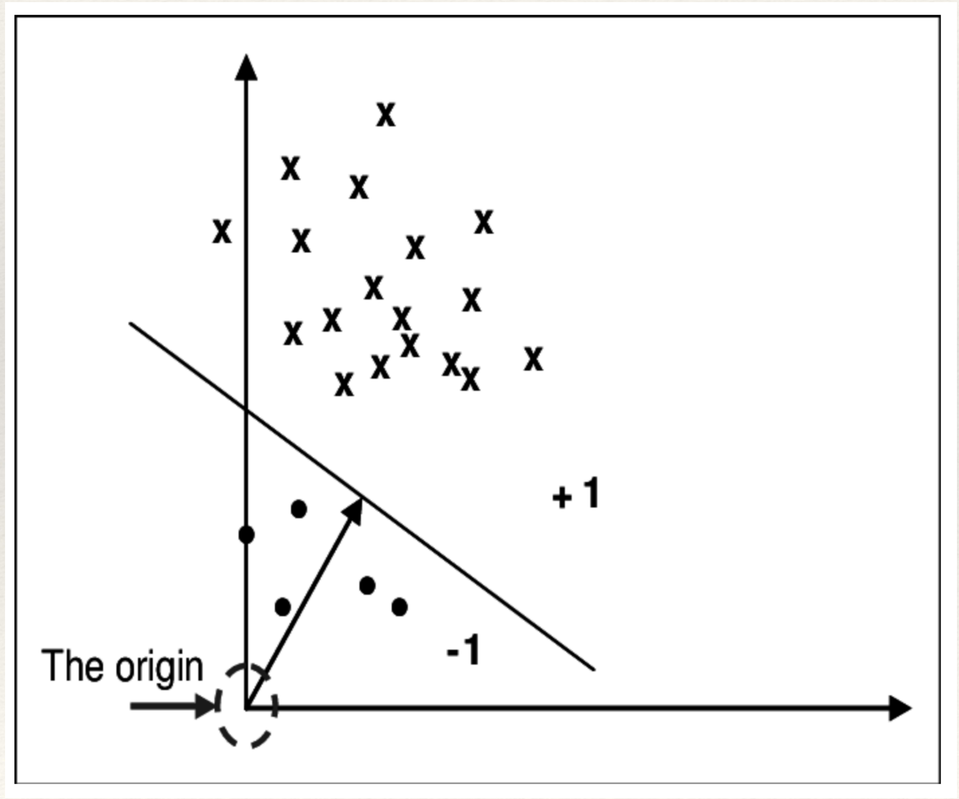
```
str = "hello world";
i = 0;
if(!str[i])          // buffer access 1   T
  skip;

size = positive_input();
for(i = 0; i < size; i++)
  skip;

... = str[i];         // buffer access 2   T
```

# OC-SVM

## One Class Support Vector Machine



# Experiments

| Program | LOC | Bug | Baseline T | Baseline F | Selective T | Selective F | Uniform T | Uniform F |
|---|---|---|---|---|---|---|---|---|
| SM-1 | 0.5K | 28 | 28 | 18 | 28 | 15 | 13 | 5 |
| SM-2 | 0.8K | 2 | 2 | 16 | 1 | 4 | 0 | 0 |
| SM-3 | 0.7K | 3 | 3 | 3 | 3 | 3 | 0 | 0 |
| SM-4 | 0.7K | 10 | 10 | 6 | 10 | 6 | 6 | 0 |
| SM-5 | 1.7K | 3 | 3 | 6 | 3 | 6 | 0 | 0 |
| SM-6 | 0.4K | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| SM-7 | 1.1K | 2 | 2 | 32 | 0 | 2 | 0 | 0 |
| BIND-1 | 1.2K | 1 | 1 | 35 | 1 | 33 | 0 | 0 |
| BIND-2 | 1.7K | 1 | 1 | 45 | 0 | 41 | 0 | 0 |
| BIND-3 | 0.5K | 1 | 1 | 4 | 0 | 1 | 0 | 0 |
| BIND-4 | 1.1K | 2 | 2 | 0 | 0 | 0 | 0 | 0 |
| FTP-1 | 0.8K | 4 | 4 | 13 | 4 | 3 | 0 | 0 |
| FTP-2 | 1.5K | 1 | 1 | 7 | 1 | 6 | 0 | 3 |
| FTP-3 | 1.5K | 24 | 24 | 25 | 23 | 17 | 7 | 12 |
| polymorph-0.4.0 | 0.7K | 10 | 10 | 6 | 3 | 6 | 0 | 6 |
| ncompress-4.2.4 | 1.9K | 12 | 0 | 10 | 4 | 0 | 0 | 0 |
| 129.compress | 2.0K | 7 | 7 | 34 | 7 | 14 | 4 | 7 |
| spell-1.0 | 2.2K | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| man-1.5h1 | 4.7K | 6 | 5 | 60 | 1 | 28 | 0 | 13 |
| 256.bzip2 | 4.7K | 3 | 3 | 149 | 3 | 21 | 3 | 21 |
| gzip-1.2.4a | 8.2K | 13 | 11 | 87 | 8 | 34 | 0 | 24 |
| bc-1.06 | 17.0K | 2 | 0 | 57 | 0 | 10 | 0 | 9 |
| sed-4.0.8 | 25.9K | 1 | 0 | 64 | 0 | 14 | 0 | 4 |
| **Total** | | 138 | 118 | 677 | 100 | 264 | 33 | 104 |

TABLE I
THE NUMBER OF ALARMS IN INTERVAL ANALYSIS

| Program | LOC | Bug | Baseline T | Baseline F | Selective T | Selective F | Uniform T | Uniform F |
|---|---|---|---|---|---|---|---|---|
| mp3rename-0.6 | 0.6K | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| ghostscript-8.71 | 1.5K | 2 | 2 | 0 | 2 | 0 | 2 | 0 |
| uni2ascii-4.14 | 5.7K | 7 | 7 | 0 | 7 | 0 | 7 | 0 |
| pal-0.4.3 | 7.4K | 3 | 3 | 0 | 0 | 0 | 0 | 0 |
| shntool-3.0.1 | 16.3K | 1 | 1 | 10 | 1 | 1 | 1 | 0 |
| sdop-0.61 | 23.9K | 65 | 65 | 78 | 65 | 0 | 0 | 0 |
| latex2rtf-2.3.8 | 28.7K | 2 | 2 | 9 | 2 | 8 | 0 | 1 |
| rrdtool-1.4.8 | 34.8K | 1 | 1 | 12 | 1 | 1 | 1 | 0 |
| daemon-0.6.4 | 58.4K | 1 | 1 | 7 | 1 | 1 | 1 | 0 |
| rplay-3.3.2 | 61.0K | 3 | 3 | 7 | 2 | 4 | 1 | 2 |
| urjtag-0.10 | 64.2K | 12 | 12 | 78 | 6 | 0 | 0 | 0 |
| a2ps-4.14 | 64.6K | 6 | 6 | 26 | 3 | 12 | 1 | 0 |
| dico-2.0 | 84.3K | 2 | 2 | 46 | 1 | 1 | 1 | 2 |
| **Total** | | 106 | 106 | 273 | 92 | 28 | 16 | 5 |

TABLE II
THE NUMBER OF ALARMS IN TAINT ANALYSIS

# Feedback

- *Problem statement*
- *Innovation*
- *Contribution*
- *Logical correcteness*
- *Proof of statements*
- *Readablity*

# What is good/interesing about the paper

- *Structured*
- *Detailed example*
- *Novel approach*

# What could be better

- *These is no code base*
- *Examples are hard to read*
- *Did not explain their choice in Experiments part*
- *Not enough references*
- *Hard to read for non-ML person*

# 8. Conclusion