

# On Deriving Actions for Improving Cost Overrun by Applying Association Rule Mining to Industrial Project Repository

Osaka University Japan

**Junya Debari**, Osamu Mizuno, Tohru Kikuno

IPA/SEC JAPAN

Nahomi Kikuchi, Masayuki Hirayama

# Overview

---

- ▶ Background
- ▶ Problem and Solution
- ▶ Case study
- ▶ Conclusion

# Overview

---

- ▶ **Background**

- ▶ What is association rule mining and why do we use this method?

- ▶ Problem and Solution

- ▶ Case study

- ▶ Conclusion

# Background

---

- ▶ It is important to avoid runaway projects.
  - ▶ Experts decide how to avoid runaway projects.
  - ▶ However, the decision is usually not based on data.
- ▶ In Japan, project data with software metrics have been collected, but are not utilized for controlling projects.
- ▶ We propose a method of deriving improvement actions of software projects by applying association rule mining.

# Association rule mining

- ▶ One of data mining methods
- ▶ Association rule
  - ▶ Described as  $[antecedent] \rightarrow [conclusion]$
  - ▶ Conclusion is restricted to either  $[project \text{ is } runaway]$  or  $[project \text{ is } success]$  in our study.
  - ▶ Example 1:
    - Rule A:  $[X \text{ is } a] \wedge [Y \text{ is } p] \rightarrow [project \text{ is } runaway]$
    - Rule B:  $[X \text{ is } b] \wedge [Y \text{ is } q] \rightarrow [project \text{ is } success]$
  - ▶ We call  $X, Y$ , and  $project$  as factors, and  $a, b, p, q, runaway$ , and  $success$  as values.

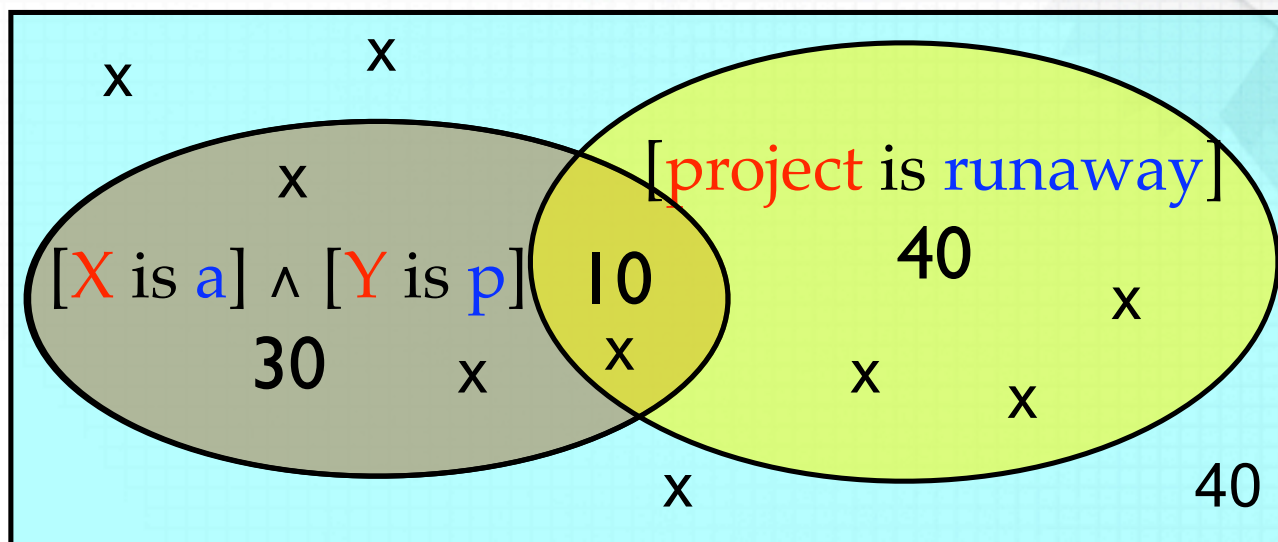
# Measures for association rules

---

- ▶ support
  - ▶ The probability that antecedent and conclusion hold simultaneously in the data set.
- ▶ confidence
  - ▶ The conditional probability that conclusion holds if antecedent is satisfied.

## Example of measures

- ▶ Rule A:  $[X \text{ is } a] \wedge [Y \text{ is } p] \rightarrow [\text{project is runaway}]$



The number of projects : 100

- ▶ support :  $10 / 100 = 10\%$
- ▶ confidence :  $10 / 30 = 33\%$

## Outline of proposed method

---

- (1) Construct a number of association rules using data mining.
- (2) Select a subset of association rules, which hold several property, from association rules constructed in (1). We define them as improvement rules.
- (3) Extract action plans which make projects success from improvement rules.

# Overview

---

- ▶ Background
- ▶ **Problem and Solution**
  - ▶ What is the problem with association rule mining and how do we solve the problem?
- ▶ Case study
- ▶ Conclusion

# Problem with association rule mining

---

1. A great number of association rules are extracted, if the threshold of support is too low.
  - ▶ However, if the threshold is too high, no rule is extracted.
2. No criteria or method exists for selecting improvement rules from association rules.
  - ▶ We propose the following method.
    - Step1 Construct a set of analogous rules from association rules.
    - Step2 Select a set of improvement rules from the set of analogous rules.

# Analogous rules

---

## ▶ Example 2

Rule A:  $[X \text{ is } a] \wedge [Y \text{ is } p] \rightarrow [\text{project is runaway}]$

Rule C:  $[X \text{ is } b] \wedge [Y \text{ is } q] \wedge [Z \text{ is } t] \rightarrow [\text{project is success}]$

## ▶ Conditions for analogous rules

1. Factors in antecedent of Rule C include factors in antecedent of Rule A. ( $\{X, Y, Z\} \supseteq \{X, Y\}$ )
2. Factors in both conclusions are the same. ( $\{\text{project}\} = \{\text{project}\}$ )

## ▶ Rule C is an analogous rule of rule A.

---

# Improvement rules

---

▶ Example 3

Rule A:  $[X \text{ is } a] \wedge [Y \text{ is } p] \rightarrow [\text{project is runaway}]$

Rule C:  $[X \text{ is } b] \wedge [Y \text{ is } q] \wedge [Z \text{ is } t] \rightarrow [\text{project is success}]$

- ▶ In this case, the following conditions hold:
1. Conclusion of Rule A is [project is runaway].
  2. Rule C is an analogous rule of Rule A.
  3. Conclusion of Rule C is [project is success].
- ▶ Then we call Rule C an improvement rule of Rule A.
-

# Procedure for extracting improvement rules

---

- ▶ From association rules, we can extract a set of improvement rules.
- ▶ Step 1
  - ▶ Construct a set P of such association rules that the conclusion is [project is runaway]
- ▶ Step 2
  - ▶ Construct a set R<sub>j</sub> of improvement rules of Rule  $j \in P$ .  
R<sub>j</sub> = {Rule i | Rule i is an analogous rule of Rule j, and Conclusion of Rule i is [project is success] }

# Overview

---

- ▶ Background
- ▶ Problem and Solution
- ▶ **Case Study**
  - ▶ Experimental data  
and the result of the case study.
- ▶ Conclusion

# Outline of Experiment

---

- ▶ Phase 1
  - ▶ Collect and prepare the data for analysis.
- ▶ Phase 2
  - ▶ Apply association rule mining to the data.
- ▶ Phase 3
  - ▶ Extract improvement rules from the obtained rules.
- ▶ Phase 4
  - ▶ Inspect software project improvement actions based on improvement rules.

## Experimental data

---

- ▶ We apply the proposed method to the year 2006 version of IPA/SEC repository.
  - ▶ The number of project data is 1,419.
  - ▶ Most of the projects are the developments of business application software.
  - ▶ The quantitative data as well as qualitative data are collected using software metrics.
  
- ▶ IPA/SEC : Information-technology Promotion Agency, Japan/Software Engineering Center

# Metrics

---

- ▶ The metrics are classified into 3 groups.
- ▶ Controllable metrics
  - ▶ They can be easily controlled to improve a running project.
- ▶ Environmental metrics
  - ▶ They are useful to analyze the condition of projects.
  - ▶ But they are hard to control
- ▶ Resultant metrics
  - ▶ They indicate the result of projects.

# Controllable metrics

---

1	New technology
2	Involvement of user for requirement specification
3	Skill of project manager
...	
15	Evaluation of plan (cost)
...	
38	Experience for the system of user

# Environmental metrics

---

1	Purpose of system
2	Use of DBMS
3	Requirement for performance
...	...
47	Actual FP Per Effort

# Resultant Metrics

---

1	Cost overrun
2	Quality evaluation
3	Duration overrun
4	Number of failures
5	Success of the project (self evaluation)
6	Customer satisfaction

# Focus in experiment

---

- ▶ Three concerns of software projects.
  - ▶ Quality
  - ▶ **Cost**
  - ▶ Duration
- ▶ In this study, we focus on the **cost** problem.
  - ▶ We restricted the factor of conclusion to **Cost overrun**.
    - ▶ [Evaluation of plan(cost) is partially clear and feasible]  
 $\wedge$ [Effort is High]  $\rightarrow$  [Cost overrun is bad]
    - ▶ [Cost overrun is good] means “Cost does not run over”  
i.e. [project is success].
    - ▶ [Cost overrun is bad] means “Cost runs over”  
i.e. [project is runaway].

## Result of association rule mining (Phase 2)

---

- ▶ Threshold of association rule mining
  - ▶ minimum support : 1%
  - ▶ minimum confidence : 90%
- ▶ The number of extracted association rules :
  - ▶ rules with conclusion [**Cost overrun** is **bad**] is 180.
  - ▶ rules with conclusion [**Cost overrun** is **good**] is 545,926.

## Result of extracting improvement rules (Phase 3)

---

- ▶ The number of the following rules is 180.  
[antecedent]→[**cost overrun** is **bad**].
- ▶ Among 180 rules, 35 rules had improvement rules.
- ▶ The total number of improvement rules is 4,116.
  
- ▶ Execution time of Phase 2 and Phase 3 was 320 minutes.
  - ▶ @Core2Duo 2.33GHz, Mem 3.0GB

## Example 1 of extracted improvement rules

[Evaluation of plan (cost) is partially clear and feasible]  
^ [Involvement of user for requirements specification is relatively]  
^ [Requirement for performance is medium]  
=> [Cost overrun is bad]      the rule concerned with cost overrun

[Evaluation of plan (cost) is very clear and feasible]  
^ [Involvement of user for requirements specification is relatively]  
^ [Requirement for performance is high]  
=> [Cost overrun is good]      the improvement rule

Based on observation of two rules,  
the following factors seem to be important:

- ▶ Evaluation of plan (cost)
- ▶ Requirement for performance

# Example 1 of improvement actions

[Evaluation of plan (cost) is partially clear and feasible]  
^ [Involvement of user for requirements specification is relatively]  
^ [Requirement for performance is medium]  
=> [Cost overrun is bad]      the rule concerned with cost overrun

[Evaluation of plan (cost) is very clear and feasible]  
^ [Involvement of user for requirements specification is relatively]  
^ [Requirement for performance is high]  
=> [Cost overrun is good]      the improvement rule

Based on observation of two rules,  
the following action should be remarked:

- ▶ Make new cost plan

## Example 2 of extracted improvement rules

[Evaluation of plan is(cost) is very clear and feasible]

^ [New technology is used]

^ [Category of business is legacy]

^ [Actual FP per effort is low]

=> [Cost overrun is bad]      the rule concerned with cost overrun

[Evaluation of plan (cost) is very clear and feasible]

^ [New technology is not used]

^ [Category of business is legacy]

^ [Actual FP per effort is high]

=> [Cost overrun is good]      the improvement rule

Based on observation of two rules,  
the following factors seem to be important:

- ▶ New technology
- ▶ Actual FP per effort

## Example 2 of extracted improvement rules

[Evaluation of plan (cost) is very clear and feasible]

^ [New technology is used]

^ [Category of business is legacy]

^ [Actual FP per effort is low]

=> [Cost overrun is bad]      the rule concerned with cost overrun

[Evaluation of plan (cost) is very clear and feasible]

^ [New technology is not used]

^ [Category of business is legacy]

^ [Actual FP per effort is high]

=> [Cost overrun is good]      the improvement rule

Based on observation of two rules,  
the following action should be remarked:

- ▶ Do not use new technology

## Evaluation by experts

---

- ▶ Finally, 35 improvement actions are extracted based on 4,116 improvement rules.
- ▶ The experts in industries evaluated the extracted improvement actions.
  - ▶ They said that all of improvement actions are basically valid and expectable ones.

# Overview

---

- ▶ Background
- ▶ Problem and Solution
- ▶ Case Study
- ▶ **Conclusion**

## Conclusion

---

- ▶ We proposed a method of deriving improvement actions of software projects by applying association rule mining.
- ▶ We applied proposed method to industrial project repository in Japan.
  - ▶ From 546,106 rules, we successfully extracted 35 improvement actions.

# Q & A

---

- ▶ Thank you.
  - ▶ If any questions, please ask in **EASY and SLOW** English.