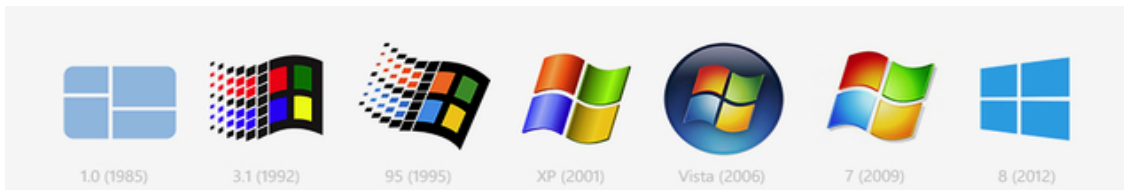
Learning From Evolution History to Predict Future Requirement Changes

Lin Shi, Qing Wang, MingShu Li
Institute of Software Chinese Academy of Sciences,
Beijing, China

Motivation

- Longtime evolution
  - Requirement changes
  - A large number of versions



- Managing the costs and risks of evolution is a challenging problem in the RE community.

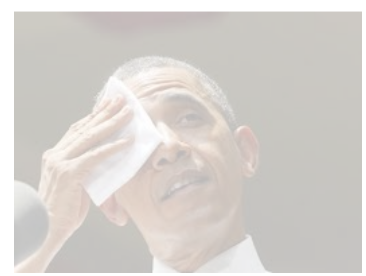
Difficult to analyze and assess the proneness to **requirement changes** across multiple versions

Motivation

- Longtime evolution
  – Requirement changes
  – A large number of versions

Find requirements that may change in the FUTURE!

- Managing the costs and risks of evolution is a challenging problem in the RE community.

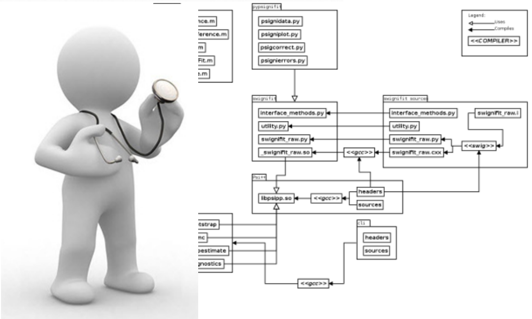


! Difficult to analyze and assess the proneness to <u>requirement changes</u> across multiple versions

Why Conduct Such a Study ?

- Being aware of volatile requirements
  - Help to reduce the workload of requirements volatility analysis

  - Important to make robust designs that could adapt to changes



  - Help to identify potential changes as early as possible.

Challenge

- Existing literatures
  - Analyze possible future scenarios (Bush, D. et.al, 2003)
  - Assess the volatility of the overall requirements (Loconsole, A. et.al, 2005)

- Studies that can predict which requirements are prone to change are rarely exploited.
  - Predicting factors are difficult to comprehend and construct.

Our Contribution

- Provide a set of metrics to measure the evolution history of a requirement.
- A methodology to predict requirements that are prone to change in the future.
- Conduct a case study in industrial settings, based on our methodology.

Agenda

1 Motivation

2 Methodology

3 Case Study

4 Conclusion

Methodology Overview - idea

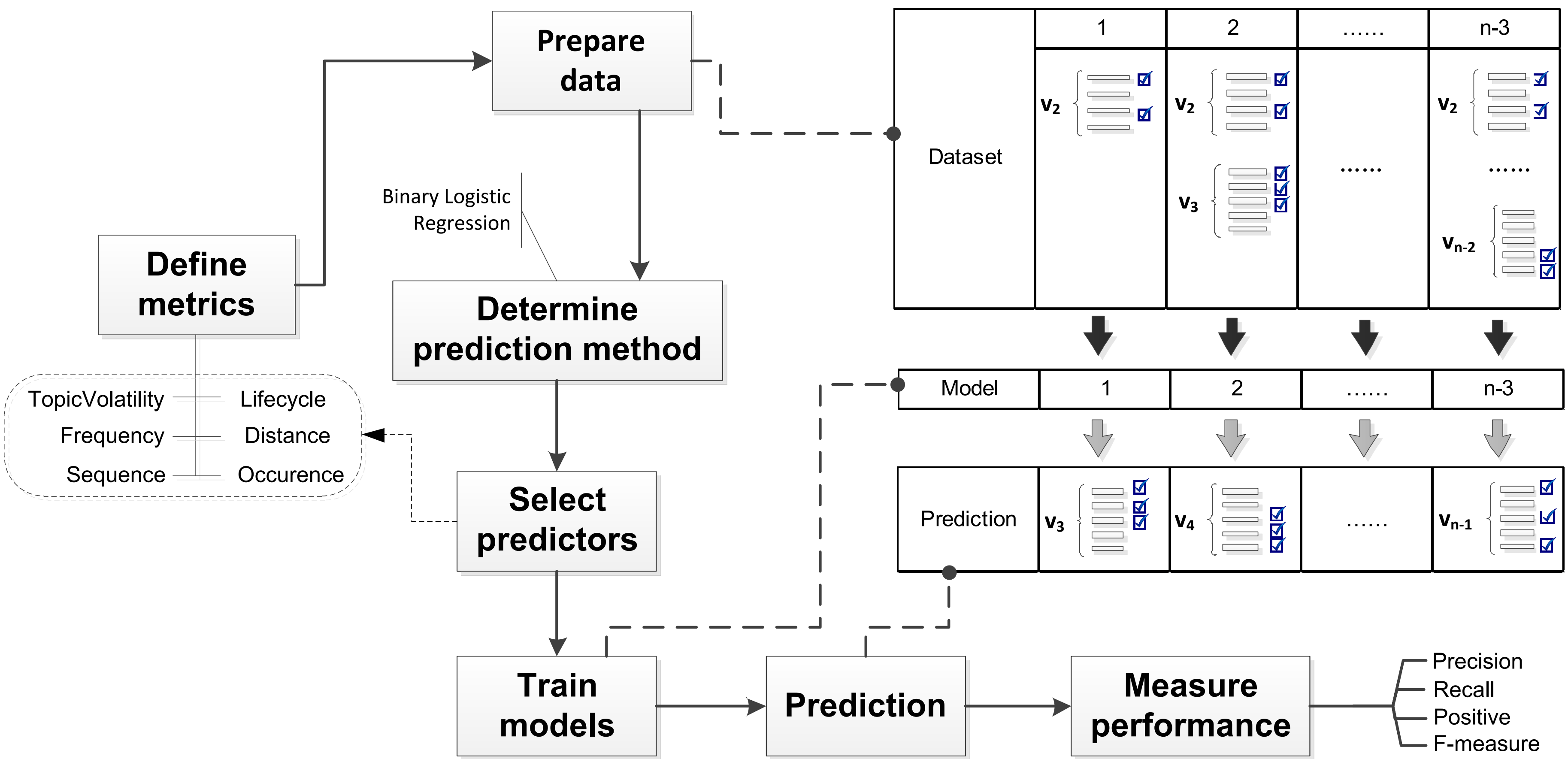
Taking lessons from HISTORY

Historic Requirements Evolution

Any relationship?

Requirements evolution in the future

Methodology Overview



Document the History

- Identify the evolution history
  - Compare the requirements of two adjacent versions
  - Build the evolution matrix

Req. ID \ Version	2	3	4	5	6	7	8
1	0	0	2	1	0	-1	0
…	…	…	…	…	…	…	…
r	0	2	0	0	1	1	0

(Where added = 2, modified = 1, deleted = -1, unchanged = 0, number of historic versions = 8)

Metrics Definition - Intuition

Hypotheses

- Requirements that are frequently changed in historic versions are also likely to change in the future.
- Requirements that belong to topics that are prone to change in historic versions are also prone to change in the future.

The metric should

1. measure how frequently a requirement changes.
2. describe the length of the time interval between two changes.
3. indicate the volatility of the topic of a requirement.

Metrics Definition

Category	Metric names	Description
Content	TopicVolatility (TV)	Average percentage of changed requirements over all requirements in the same topic per version
Change degree	Frequency (F)	Sum of the times of changes for a requirement over all historic versions
	Sequence (S)	Maximum number of versions that a requirement continuously changed
Length of change time	Distance (D)	Sum of number of versions between two changes.
	Lifecycle (LC)	The number of versions between the version that a requirement is added and the latest version
	Occurrence (OC)	The position of the center of changes towards the lifecycle of a requirement

$$TopicVolatility\ (r) = \frac{1}{n-2}\sum_{i=2}^{n-1}\frac{|R_C(i,t)|}{|R(i,t)|}$$

$$Frequency\ (r) = \frac{1}{n-1}\sum_{i=2}^{n} isChanged(r,i)$$

$$Sequence\ (r) = max(N(r,i)),\ (2 \leq i \leq n)$$

$$Distance\ (r) = \sum_{i=2}^{n}(n-i)$$

$$Lifecycle\ (r) = n - VA + 1$$

$$Occurrence\ (r) = \frac{1}{Lifecycle\ (r)} * \frac{Distance\ (r)}{Frequency\ (r)}$$

Prediction Method - Logistic Regression

- Logistic Regression (LR) is a prediction approach that can be used when the target variable is a categorical variable with two categories.
- LR aims to determine whether there is some form of functional dependency between the explanatory variables and the dependent variable.

$$Logit(P(change)) = w_1*\alpha_1 + w_2*\alpha_2 + \ldots\ldots + w_n*\alpha_n + w_0,$$

Where $Logit(p) = \log\{p/(1-p)\}$, α_i *denotes the predictors selected in the model,* w_0 *is the intercept.*

Training and Predicting

- Example - Requirements of the v_i version

Req.ID	**TV**	**F**	**S**	**D**	**LC**	**OC**	***Change***
1	0.24	0.25	0.33	3	5	2.4	0
2	0.44	0.5	0.33	3	4	1.5	1
3	0.36	0.75	0.18	4	5	1.1	1

Training and Predicting

- Example - Requirements of the v_i version

Req.ID	**TV**	**F**	**S**	**D**	**LC**	**OC**	***Change***
1	0.24	0.25	0.33	3	5	2.4	0
2	0.44	0.5	0.33	3	4	1.5	1
3	0.36	0.75	0.18	4	5	1.1	1

The requirement actually change in v_{i+1}

Prediction Performance Measurement

		Predicted	
		Not change	***change***
Actual	***Not change***	TN = True Negative	FP = False Positive
	change	FN = False Negative	TP = True Positive

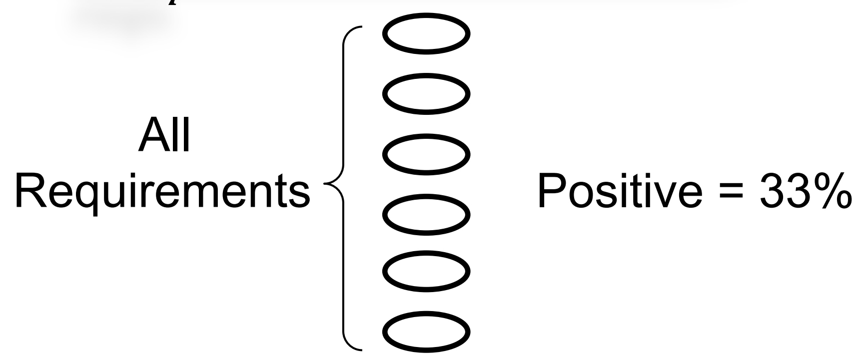
Precision = $TP/TP+FP$

Recall = $TP/TP+FN$

F-Measure = $2*Precision*Recall/Precision+Recall$

Positive = $TP+FP/TP+FN+TN+FP$

#Reqs predicted to be changed/Total number of reqs.



Indicate the workload of RV analysis

Agenda

1 Motivation

2 Methodology

3 Case Study

4 Conclusion

Case Study

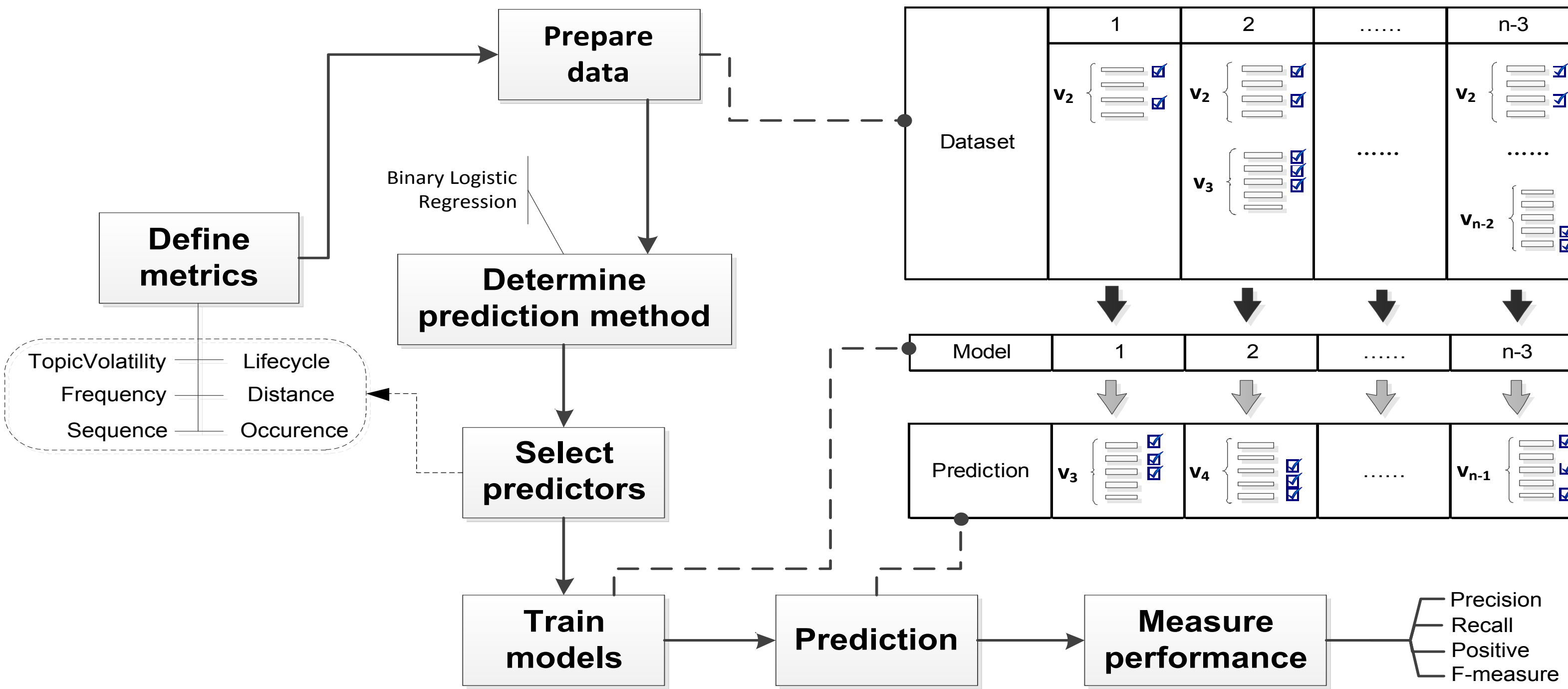
- Data Sources
- Selected Predictors
- Building Models
- Prediction Results
  - Preliminary results
  - Cutoff point selection
  - Ultimate results

Data Sources

- An industrial product
  - – 8 year on-going
  - – enterprise software process management
  - – 13 historic versions
  - – 4,044 requirements with 800 changes

#version	1	2	3	4	5	6	7	8	9	10	11	12	13
Release time	2004	2005	2006	2007	2007	2008	2009	2009	2010	2010	2011	2011	2012
# req.	186	189	195	221	276	310	308	334	355	389	394	426	461
	A	3	6	26	60	52	11	38	21	34	5	32	35
	M	10	22	60	43	57	26	28	36	77	20	28	32
	D	0	0	0	5	18	13	2	0	0	0	0	0

Datasets for training and testing



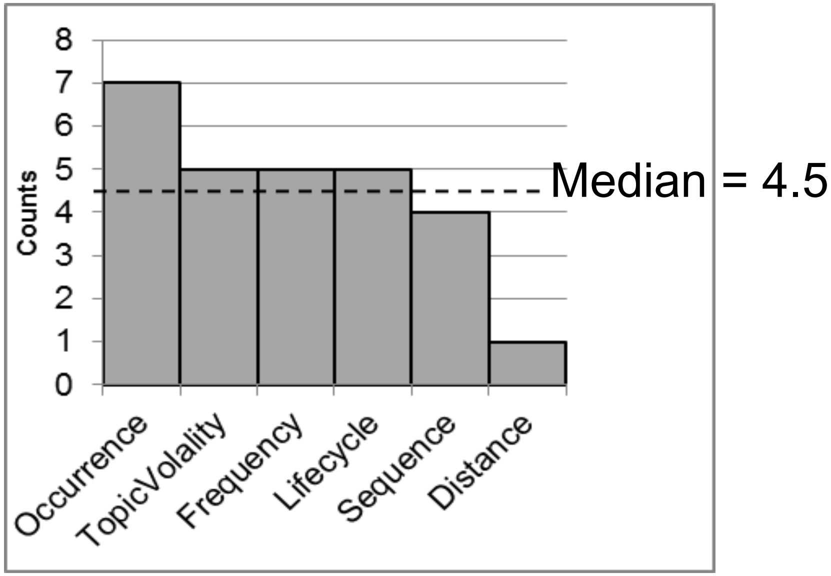
#	datasets for training	datasets for testing
1	$\{v_3\}$	$\{v_4\}$
2	$\{v_3, v_4\}$	$\{v_5\}$
3	$\{v_3, v_4, v_5\}$	$\{v_6\}$
……	……	……
9	$\{v_3, v_4, \ldots\ldots, v_{11}\}$	$\{v_{12}\}$

Selected Predictors

- Stepwise regression
  - iteratively deleting predictors from the full model until no further improvement could provide

# Regression	1	2	3	4	5	6	7	8	9
TopicVolality					√	√	√	√	√
Frequency					√	√	√	√	√
Distance			√						
Lifecycle	√	√					√	√	√
Occurrence			√	√	√	√	√	√	√
Sequence	√	√	√	√					

Selected Predictors



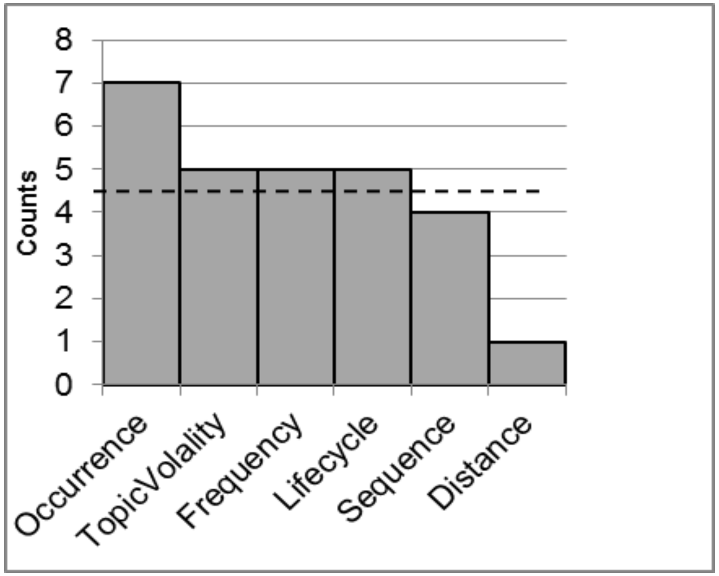
$$Logit\,(P(change)) = w_1 * topicVolatility + w_2 * frequence + w_3 * lifecycle + w_4 * occurrence + w_0,$$

where $Logit(p) = \log\{p/(1-p)\}$, w_0 *is the intercept.*

# Regression	1	2	3	4	5	6	7	8	9
TopicVolality					√	√	√	√	√
Frequency					√	√	√	√	√
Distance			√						
Lifecycle	√	√					√	√	√
Occurrence			√	√	√	√	√	√	√
Sequence	√	√	√	√					

Selected Predictors

# Regression	1	2	3	4	5	6	7	8	9
TopicVolality					√	√	√	√	√
Frequency					√	√	√	√	√
Distance			√						
Lifecycle	√	√					√	√	√
Occurrence			√	√	√	√	√	√	√
Sequence	√	√	√	√					



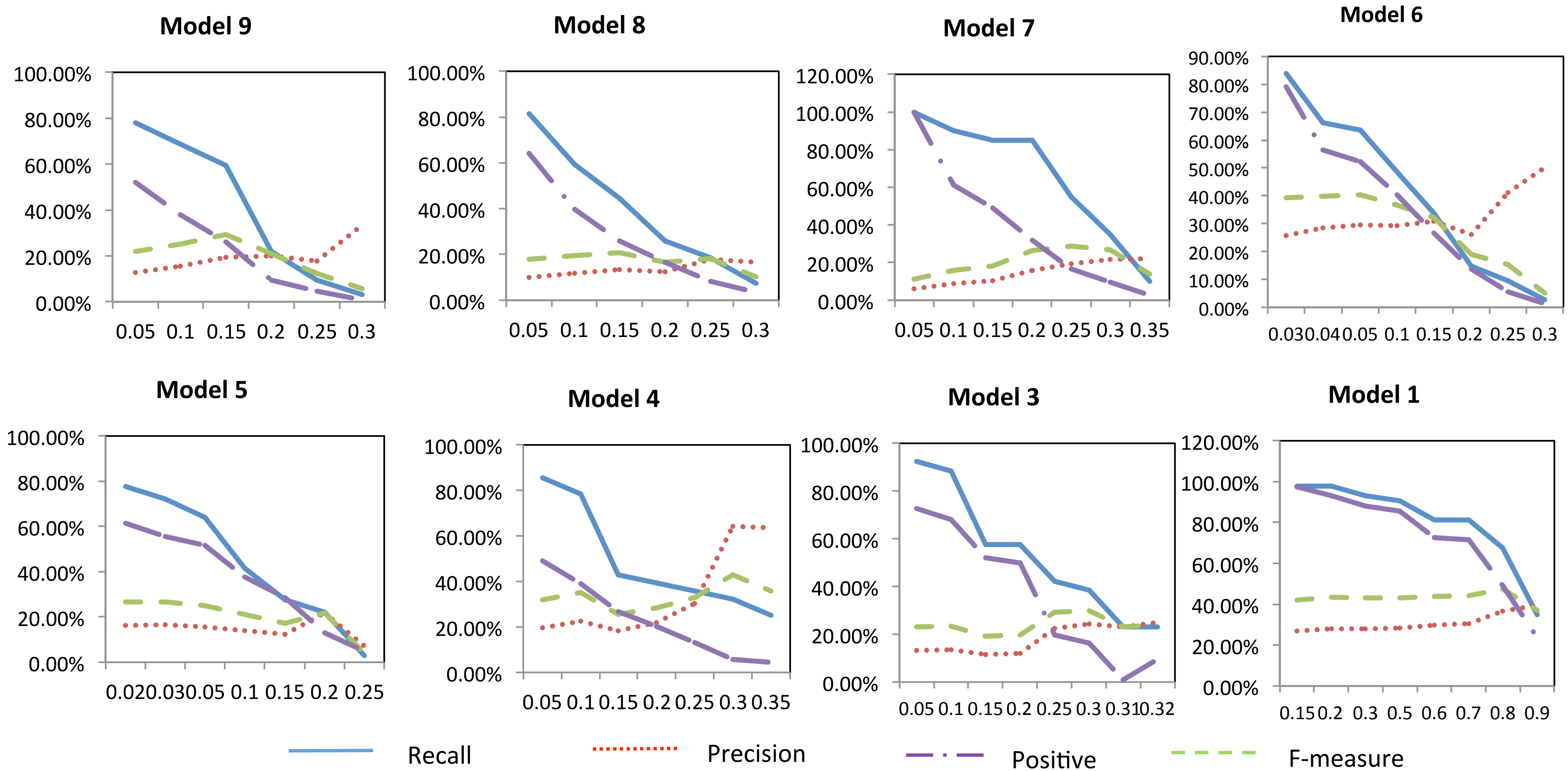
$$Logit\ (P(change)) = w_1 * topicVolatility + w_2 * frequence + w_3 * lifecycle + w_4 * occurrence + w_0,$$

where $Logit(p) = \log\{p/(1-p)\}$, w_0 *is the intercept.*

Building Models

#Model	Model Coefficients		TopicVolatility		Frequency		Lifecycle		Occurrence	
	Chi-square	*Sig.*	*B*	*Sig.*	*B*	*Sig.*	*B*	*Sig.*	*B*	*Sig.*
1	19.146	0.001	7.008	0.004	2.317	0.022	1.070	0.032	-1.430	0.152*
2	4.377	0.357*	-0.466	0.687*	1.024	0.083*	0.189	0.168*	-0.327	0.411*
3	23.579	0.000	-0.570	0.516*	1.225	0.010	0.238	0.001	-0.737	0.001
4	55.927	0.000	-1.656	0.042	1.338	0.003	0.155	0.004	-0.774	0.000
5	98.167	0.000	-1.898	0.017	1.518	0.000	0.134	0.003	-0.683	0.000
6	95.062	0.000	-2.448	0.001	1.463	0.000	0.101	0.005	-0.414	0.000
7	67.416	0.000	-2.060	0.004	1.070	0.003	0.117	0.000	-0.207	0.000
8	109.668	0.000	-1.925	0.007	1.297	0.000	0.110	0.000	-0.270	0.000
9	134.481	0.000	-1.751	0.012	1.323	0.000	0.090	0.000	-0.268	0.000

Prediction Results - Preliminary results

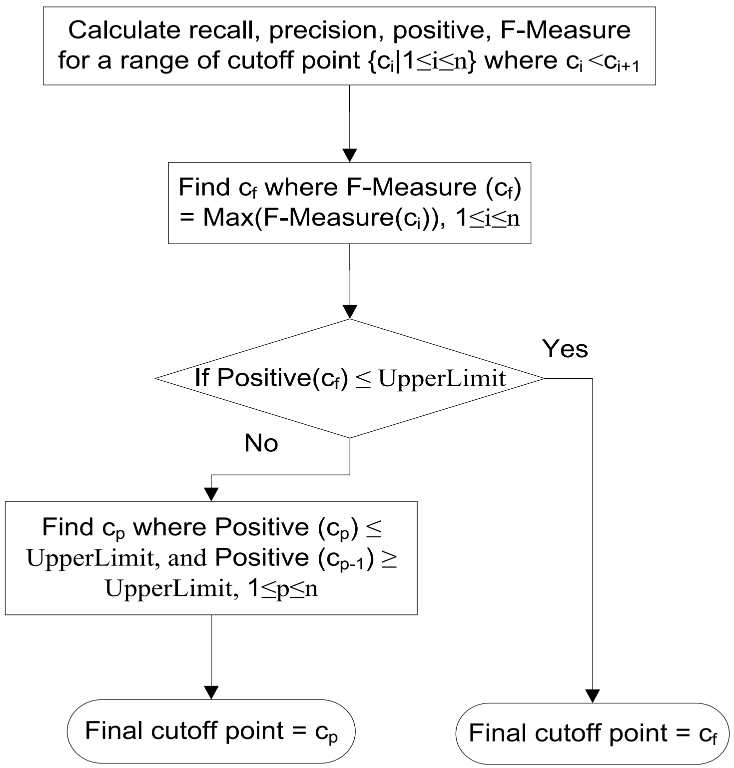


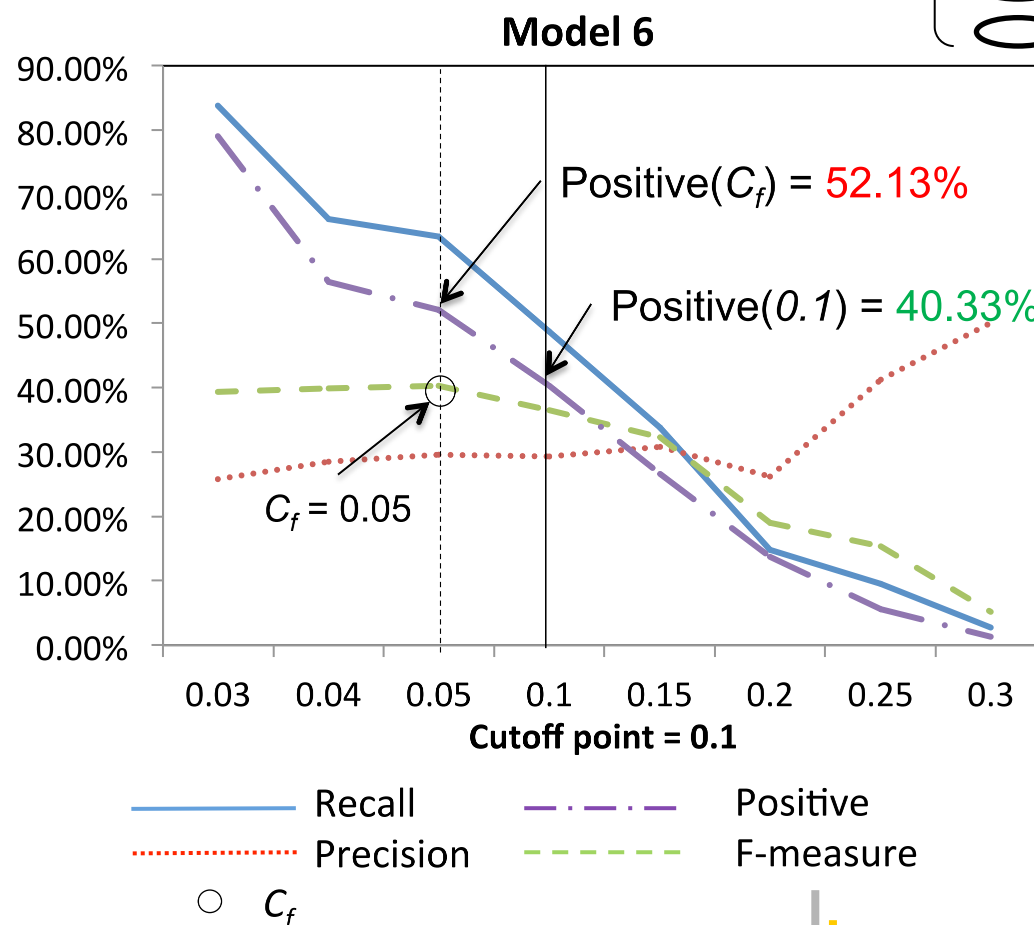
- Find the optimal cutoff point
  - Classify the probability of a requirement to be changed or not

Prediction Results - Cutoff point selection

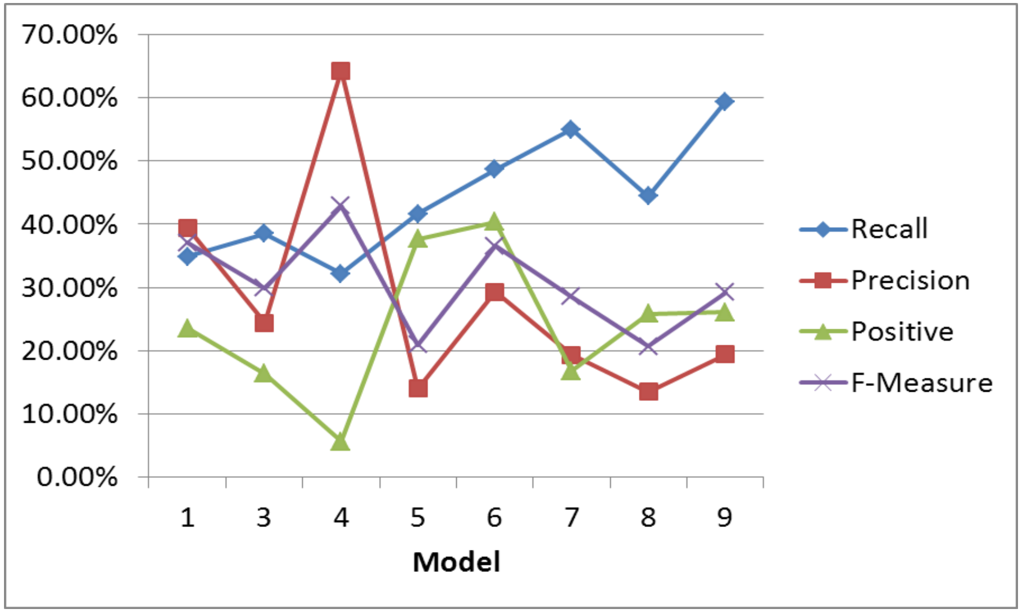
All Requirements

Constrain: positive < 50%





Prediction Results - Ultimate results



	Max	Min	Mean	Std. Dev.
Positive	40.33%	5.65%	$\boxed{24.05\%}$	0.114
Recall	59.38%	32.14%	$\boxed{44.33\%}$	0.095
Precision	64.29%	13.48%	$\boxed{27.95\%}$	0.170
F-Measure	42.86%	20.69%	30.72%	0.078

- Retrieve nearly half of future requirement changes with the cost of 2 % effort on analysis in average.

Agenda

1 Motivation

2 Methodology

3 Case Study

4 Conclusion

Conclusion

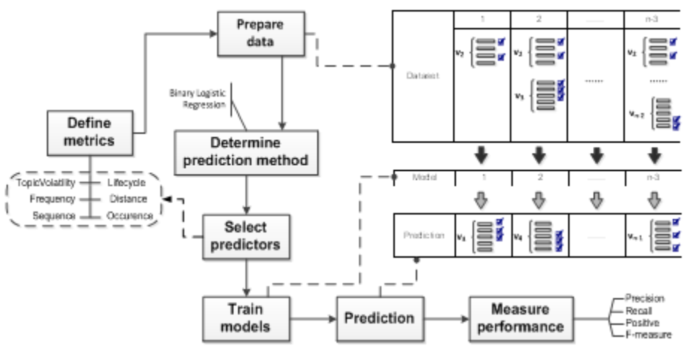
– A novel solution helps to downsize the workload of requirements volatility analysis by recommending a converging subset of change-prone requirements.

– The proposed method can achieve a tradeoff between analysis costs and evolution risks to help practitioners manage the requirements evolution.

– Our study could contribute to quantitative requirements management for companies that devote to high maturity process improvement.

Questions?

Methodology Overview



ISCAS

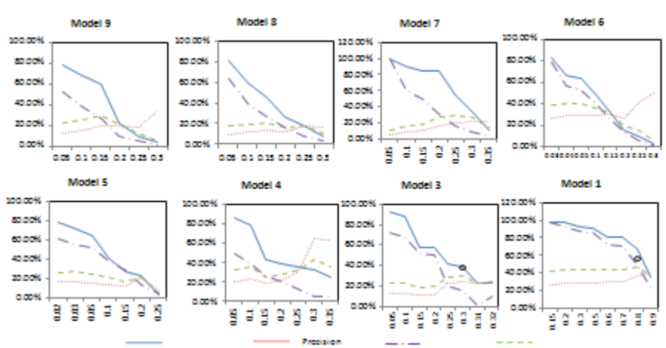
Metrics Definition

- Considerations
  - The metric should indicate the volatility of the topic that a requirement belongs to.
  - The metric should measure how frequently a requirement changes.
  - The metric should describe the length of the time interval between two changes.

Category	Metric names	Description
Content	TopicVolality (TV)	Average percentage of changed requirements over all requirements in the same topic per version
Change degree	Frequency (F)	Sum of the times of changes for a requirement over all historic versions
	Sequence (S)	Maximum number of versions that a requirement continuously changed
Length of change time	Distance (D)	Sum of number of versions between the version that a requirement is changed and the latest version.
	Lifecycle (LC)	The number of versions between the version that a requirement is added and the latest version
	Occurrence (OC)	The ratio of the difference of center of change to the latest version to the lifecycle for a requirement

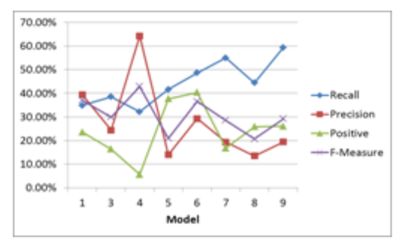
ISCAS

Prediction Results - Preliminary results



ISCAS

Prediction Results - Ultimate results



	Max	Min	Mean	Std. Dev.
Recall	59.38%	32.14%	44.33%	0.095
Precision	64.29%	13.48%	27.95%	0.170
Positive	40.33%	5.65%	24.05%	0.114
F-Measure	42.86%	20.69%	30.72%	0.078

- There are 24.05% requirements predicted to change in average.
- The average precision for predicted possible changing requirements is 27.95%.
- Prediction models can retrieve 44.33% actual changed requirements in average.

ISCAS