Scalable Estimates of Concept Stability

Aleksey Buzmakov[1,2] Sergei O. Kuznetsov[2] Amedeo Napoli[1]

[1]INRIA-LORIA (CNRS-Université de Lorraine), Vandoeuvre-lès-Nancy, France

[2]National Research University Higher School of Economics, Moscow, Russia

12th International Conference on Formal Concept Analysis
June 10-13, 2014

Scalable Estimates of Concept Stability

Aleksey Buzmakov[1,2] Sergei O. Kuznetsov[2] Amedeo Napoli[1]

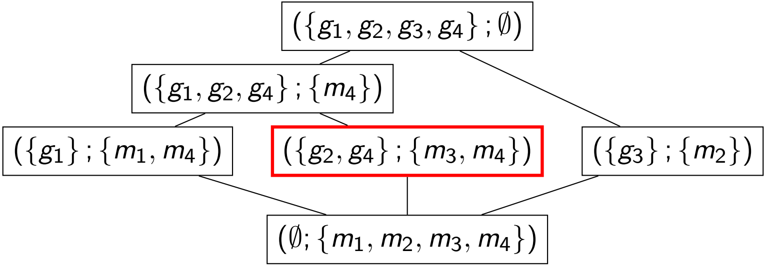
[1]INRIA-LORIA (CNRS-Université de Lorraine), Vandoeuvre-lès-Nancy, France

[2]National Research University Higher School of Economics, Moscow, Russia

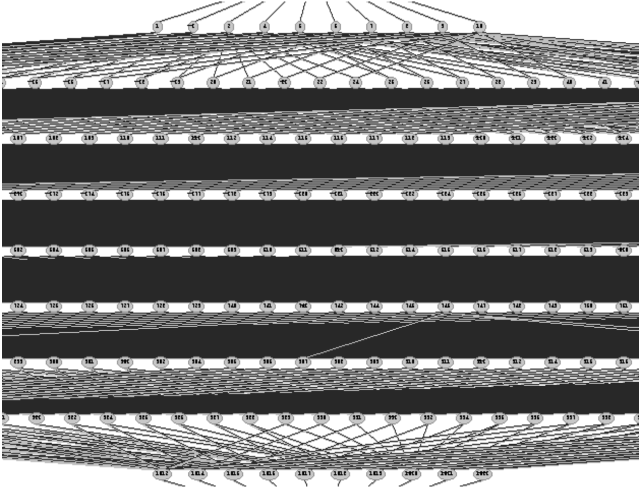
12th International Conference on Formal Concept Analysis
June 10-13, 2014

Outline

What you can find in presentations



Real Datasets



How to select the best patterns?

Stability [Kuznetsov, 1990, Kuznetsov, 2007, Roth et al., 2008]

Concept Probability and Separation [Klimushkin et al., 2010]

Basic Level of Concepts [Belohlavek and Trnecka, 2013]

Others from Data Mining [Hilderman and Hamilton, 1999, McGarry, 2005, Geng and Hamilton, 2006, Webb, 2010, Shaharanee and Hadzic, 2013]

The target selection tool in this paper is Stability.

Outline

1 Introduction

2 FCA and Stability

3 Experiment I: Behaviour of Stability

4 Experiment II: Estimate of Stability

5 Conclusion

Formal Concept Analysis (FCA) [Ganter and Wille, 1999]

(G, M, I)

	m_1	m_2	m_3	m_4
g_1	x			x
g_2			x	x
g_3		x		
g_4			x	x

Figure: A toy FCA context.

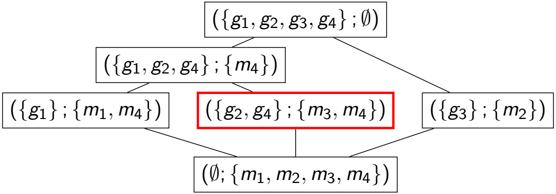


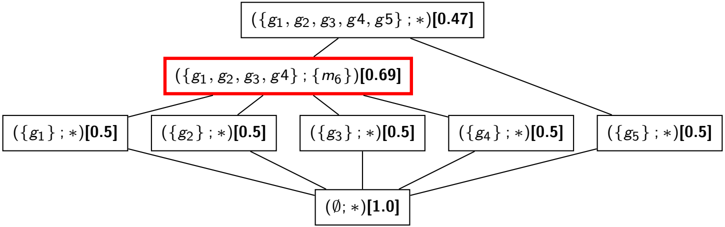
Figure: Concept Lattice for the toy context

$$A' = \{m \in M \mid \forall g \in A, (g, m) \in I\}, \qquad A \subseteq G$$
$$B' = \{g \in G \mid \forall m \in B, (g, m) \in I\}, \qquad B \subseteq M$$

Formal Concept

Formal Concept is a pair $\mathcal{C} = (A, B)$, $A \subseteq G$ and $B \subseteq M$. A is called an extent and denoted as $\mathtt{Ext}(\mathcal{C})$. B is called an intent, denoted as $\mathtt{Int}(\mathcal{C})$.

Stability of a Concept [Kuznetsov, 1990, Roth et al., 2008]



Definition

Given a concept $\mathcal{C}$, concept stability $\mathtt{Stab}(\mathcal{C})$ is defined as

$$\mathtt{Stab}(\mathcal{C}) := \frac{|\{s \in \wp(\mathtt{Ext}(\mathcal{C})) \mid s' = \mathtt{Int}(\mathcal{C})\}|}{2^{|\mathtt{Ext}(\mathcal{C})|}}$$

i.e. the relative number of subsets of the concept extent (denoted by $\mathtt{Ext}(\mathcal{C})$), whose description (i.e. the result of $(\cdot)'$) is equal to the concept intent (denoted by $\mathtt{Int}(\mathcal{C})$) where $\wp(P)$ is the power set of P.

Stability of a Concept

	m_1	m_2	m_3	m_4	m_5	m_6
g_1	X					X
g_2		X				X
g_3			X			X
g_4				X		X
g_5					X	

Figure: Stability toy context.

$$\texttt{Stab}((\{g_1, \cdots, g_4\}; \{m_6\})) = ?$$

Stability of a Concept

	m_1	m_2	m_3	m_4	m_5	m_6
g_1	X					X
g_2		X				X
g_3			X			X
g_4				X		X
g_5					X	

Figure: Stability toy context.

$(\{g_2, g_3, g_4\}, \{m_6\})$

$\texttt{Stab}((\{g_1, \cdots, g_4\}; \{m_6\})) = \frac{0+1}{1}$

Stability of a Concept

	m_1	m_2	m_3	m_4	m_5	m_6
g_1	X					X
g_2		X				X
g_3			X			X
g_4				X		X
g_5					X	

Figure: Stability toy context.

$(\{g_1, g_3, g_4\}, \{m_6\})$

$\mathtt{Stab}((\{g_1, \cdots, g_4\}; \{m_6\})) = \frac{1+1}{2}$

Stability of a Concept

	m_1	m_2	m_3	m_4	m_5	m_6
g_1	X					X
g_2		X				X
g_3			X			X
g_4				X		X
g_5					X	

Figure: Stability toy context.

~~$(\{g_2\}, \{m_2, m_6\})$~~

$\mathtt{Stab}((\{g_1, \cdots, g_4\}; \{m_6\})) = \frac{2+0}{3}$

Stability of a Concept

	m_1	m_2	m_3	m_4	m_5	m_6
g_1	X					X
g_2		X				X
g_3			X			X
g_4				X		X
g_5					X	

Figure: Stability toy context.

$(\{g_1, g_2, g_4\}, \{m_6\})$

$\texttt{Stab}((\{g_1, \cdots, g_4\}; \{m_6\})) = \frac{2+1}{4}$

Stability of a Concept

	m_1	m_2	m_3	m_4	m_5	m_6
g_1	X					X
g_2		X				X
g_3			X			X
g_4				X		X
g_5					X	

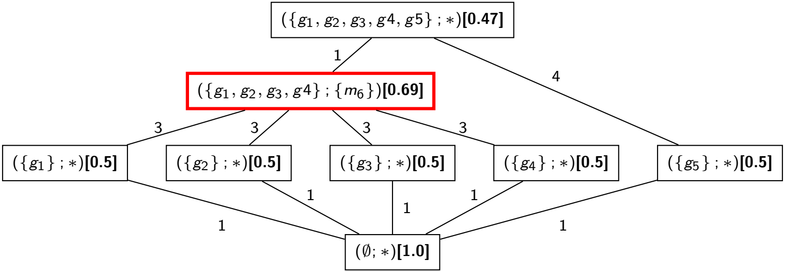
Figure: Stability toy context.

$(\{g_1, g_3\}, \{m_6\})$

$\mathtt{Stab}((\{g_1, \cdots, g_4\}; \{m_6\})) = \frac{3+1}{5}$

Estimate of Stability (Estimate Method)

Stability is #P-complete [Kuznetsov, 1990] $\Rightarrow$ estimates are important.



$$1 - \sum_{\mathcal{D} \in \mathtt{DD}(\mathcal{C})} \frac{1}{2^{\Delta(\mathcal{C},\mathcal{D})}} \leq \mathtt{Stab}(\mathcal{C}) \leq 1 - \max_{\mathcal{D} \in \mathtt{DD}(\mathcal{C})} \frac{1}{2^{\Delta(\mathcal{C},\mathcal{D})}}$$

$$0.5 = 1 - 4 \cdot \frac{1}{2^3} \leq 0.69 \leq 1 - \frac{1}{2^3} = 0.875$$

Previous Estimate of Stability

- Monte-Carlo approach by [Babin and Kuznetsov, 2012]
  - Tightness can be set.
  - Relatively slow: requires $N > \frac{1}{2\varepsilon^2} \ln \frac{2}{\delta}$ iterations, where ε is a required precision and δ is an error rate.
- Estimate Method
  - Tightness is undefined.
  - Fast: $O(|G| \cdot |M|^2)$ – for enumerating children.

Combined method

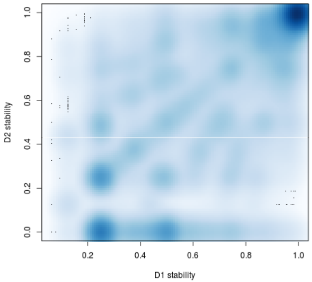
1. Compute an estimate by Estimate Method.
2. If tightness is to low, compute an estimate by Monte Carlo method.

Outline

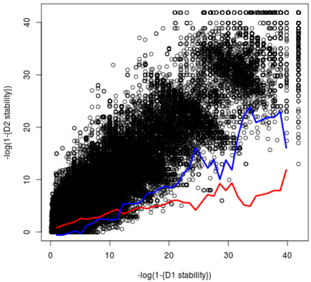
Scheme of the Experiment

1. Given a dataset $\mathbb{K}$, two disjoint dataset $\mathbb{K}_1$ and $\mathbb{K}_2$ are generated from $\mathbb{K}$ by sampling. $\mathbb{K}_1$ is called a reference dataset, $\mathbb{K}_2$ is a test dataset.
2. Two concept lattices ($\mathcal{L}_1$ and $\mathcal{L}_2$) are built for $\mathbb{K}_1$ and $\mathbb{K}_2$.
3. Concept $\mathcal{C}^1 \in \mathcal{L}_1$ is matched to $\mathcal{C}^2 \in \mathcal{L}_2$ if and only if $\mathtt{Int}(\mathcal{C}^1) = \mathtt{Int}(\mathcal{C}^2)$.
4. The stability of matched concepts can be further compared.

Stability or Logarithmic Stability?



(a) Mush4000.



(b) Mush4000 logarihtmic scale.

Figure: Stability in the test dataset w.r.t the reference one in Mush4000 in (a) plane scale (b) logarithmic scale.

Stability Threshold and Dataset Size

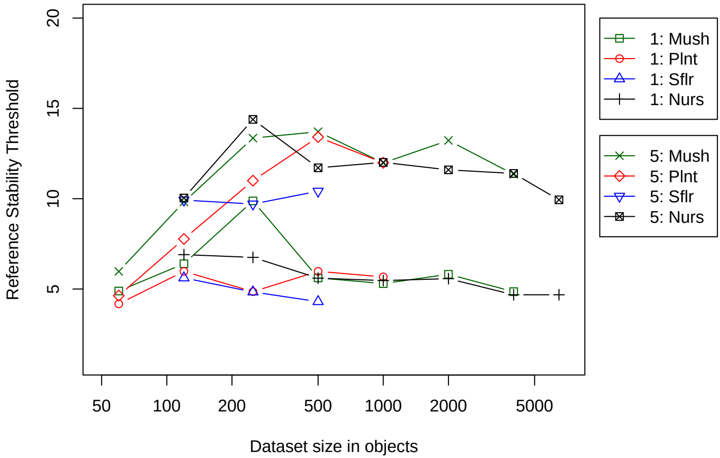


Figure: Stability threshold in the reference dataset ensuring that 99% of concepts in the test datasets corresponding to stable concepts are stable with stability thresholds 1 or 5.

Stability Ordering Ability

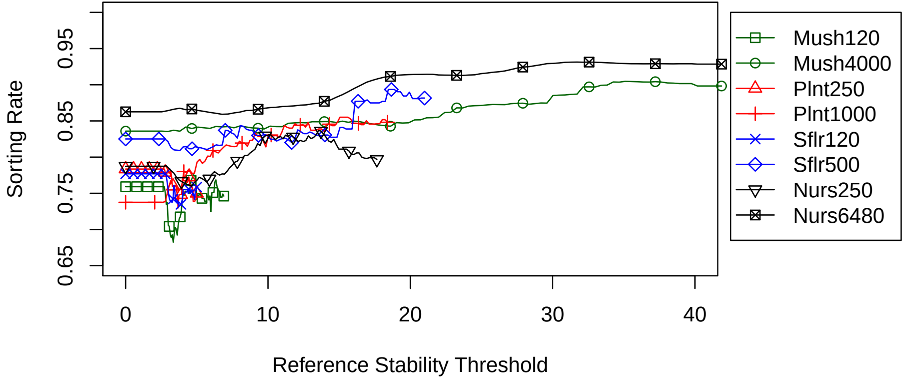


Figure: Global sorting rate for different datasets.

Outline

Computational Efficiency of the Estimate

Table: Execution time for different steps on different datasets. `Freq.` is the frequency threshold applied for big datasets; #MC is the number of calls to Monte Carlo Routine in Combined Method. The execution times are given in seconds.

Dataset	$\|\mathcal{L}\|$	$t_{\mathcal{L}}$	$t_{\texttt{Stab}}$	$t_{\texttt{FCbO}}$	Freq.	$t_{\texttt{Estimate}}$	$t_{\texttt{Comb.}}$	#MC
Mush8124	$2.3 \cdot 10^5$	324	57	0.7	0	$2 \cdot 10^3$	$6 \cdot 10^3$	$6 \cdot 10^4$
Plnt1000	$2 \cdot 10^6$	45	10^4	78	0	181	446	$3 \cdot 10^3$
Chss100	$2 \cdot 10^6$	46	10^4	3.5	0	90	192	$2.3 \cdot 10^3$
SFlr1066	2988	< 1	< 1	< 1	0	< 1	11	284
Nurs12960	$1.2 \cdot 10^5$	245	5	< 1	0	425	$1.2 \cdot 10^3$	$4 \cdot 10^4$
Chss3196	$4.4 \cdot 10^6$	–	–	42	1000	$2 \cdot 10^4$	$3.5 \cdot 10^4$ (2%)	?
Plnt34781	$5.8 \cdot 10^6$	–	–	795	1750	$4.1 \cdot 10^5$	$4.6 \cdot 10^5$ (4.7%)	?

Tightness of the Estimate Method

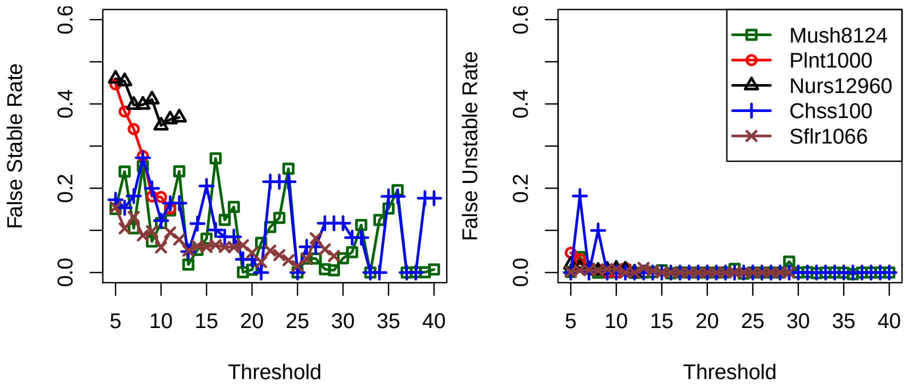


Figure: Over- and under- estimation rate for selecting stable concepts w.r.t. upper and lower bound of stability.

Tightness Mean and Std. Deviation

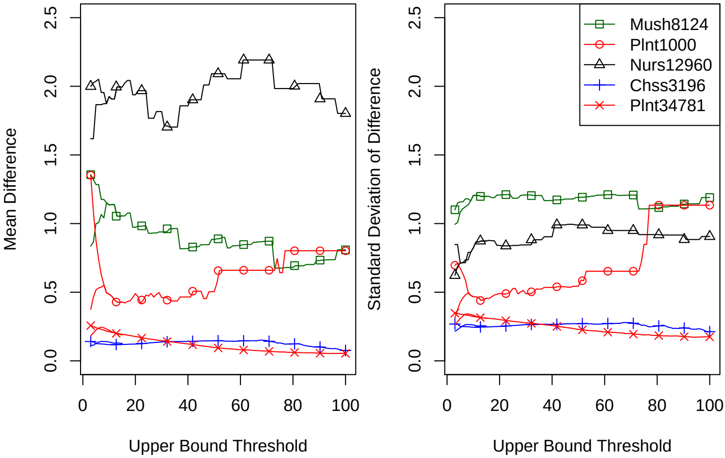


Figure: The mean and the standard deviation of the stability estimate interval.

Concept Ranking with the Estimate

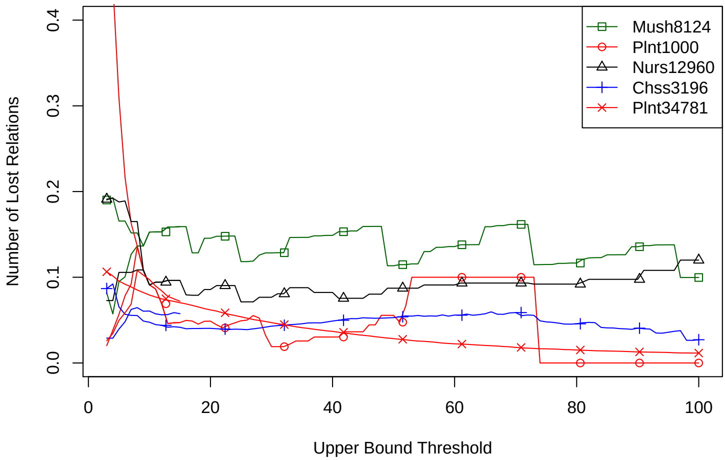


Figure: Losing rate of relations for stability estimate.

Conclusion

- Stability has similar behaviour on the datasets produced by the same general population.
- Logarithmic scale is more suitable for stability.
- Stability should be at least 5 in order to ensure that concept is found in a test dataset.
- Ordering of concepts by stability in different datasets is similar.
- The introduced estimates of stability are efficient in terms of computations and tightness and can be used for computing stability in big datasets.
    - Estimate Method is fast.
    - Combined Method ensures a required level of tightness.
- The estimates of stability lose less than 20% of ordering relation.

Future Work

- A formal proof of the experimental finding is necessary.
- Finding stable concepts by resampling is an important question for efficient FCA.
- Stability should be objectively compared w.r.t. other relevancy measures.
- Efficient realisation of the estimates (e.g. by parallelization) is an important task.

Scalable Estimates of Concept Stability

Aleksey Buzmakov[1,2] Sergei O. Kuznetsov[2] Amedeo Napoli[1]

[1]INRIA-LORIA (CNRS-Université de Lorraine), Vandoeuvre-lès-Nancy, France

[2]National Research University Higher School of Economics, Moscow, Russia

12th International Conference on Formal Concept Analysis
June 10-13, 2014

Bibliography

- Ganter, B. and Wille, R. (1999).
  Formal Concept Analysis: Mathematical Foundations.
  Springer, 1st edition
- Kuznetsov, S. O. (1990). Stability as an Estimate of the Degree of Substantiation of Hypotheses on the Basis of Operational Similarity.
  Autom. Doc. Math. Linguist. (Nauch. Tekh. Inf. Ser. 2), 24(6):62–75

Bibliography: Stability

- Kuznetsov, S. O. (1990). Stability as an Estimate of the Degree of Substantiation of Hypotheses on the Basis of Operational Similarity.
  Autom. Doc. Math. Linguist. (Nauch. Tekh. Inf. Ser. 2), 24(6):62–75

- Kuznetsov, S. O. (2007). On stability of a formal concept.
  Ann. Math. Artif. Intell., 49(1-4):101–115

- Kuznetsov, S., Obiedkov, S., and Roth, C. (2007). Reducing the Representation Complexity of Lattice-Based Taxonomies.
  In Priss, U., Polovina, S., and Hill, R., editors, Concept. Struct. Knowl. Archit. Smart Appl., volume 4604 of Lecture Notes in Computer Science, pages 241–254. Springer Berlin Heidelberg

- Roth, C., Obiedkov, S., and Kourie, D. G. (2008). On succinct representation of knowledge community taxonomies with formal concept analysis A Formal Concept Analysis Approach in Applied Epistemology.
  Int. J. Found. Comput. Sci., 19(02):383–404

- Babin, M. and Kuznetsov, S. (2012). Approximating Concept Stability.
  In Domenach, F., Ignatov, D., and Poelmans, J., editors, Form. Concept Anal., volume 7278 of Lecture Notes in Computer Science, pages 7–15. Springer Berlin Heidelberg

Datasets

Frank, A. and Asuncion, A. (2010).

UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].
University of California, Irvine, School of Information and Computer Sciences

- **Mushrooms:**
  http://archive.ics.uci.edu/ml/datasets/Mushroom
- **Plants:**
  http://archive.ics.uci.edu/ml/machine-learning-databases/plants/
- **Chess:**
  http://archive.ics.uci.edu/ml/datasets/Chess+(King-Rook+vs.+King-Pawn)
- **Solar Flare (II):**
  http://archive.ics.uci.edu/ml/datasets/Solar+Flare
- **Nursery:**
  http://archive.ics.uci.edu/ml/datasets/Nursery