Concept Stability as a Tool for Pattern Selection

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

[1] LORIA (CNRS – Inria NGE – U. de Lorraine), Vandœuvre-lès-Nancy, France
[2] National Research University Higher School of Economics, Moscow, Russia
aleksey.buzmakov@inria.fr, amedeo.napoli@loria.fr, skuznetsov@hse.ru

Abstract. Data mining aims at finding interesting patterns from datasets, where "interesting" means reflecting intrinsic dependencies in the domain of interest rather than just in the dataset. Concept stability is a popular relevancy measure in FCA but its behaviour have never been studied on various datasets. In this paper we propose an approach to study this behaviour. Our approach is based on a comparison of stability computation on datasets produced by the same general population. Experimental results of this paper show that high stability of a concept in one dataset suggests that concepts with the same intent in other dataset drawn from the population have also high stability. Moreover, experiments shows some asymptotic behaviour of stability in such kind of experiments when dataset size increases.

Keywords: formal concept analysis, stability, pattern selection, experiments

1 Introduction

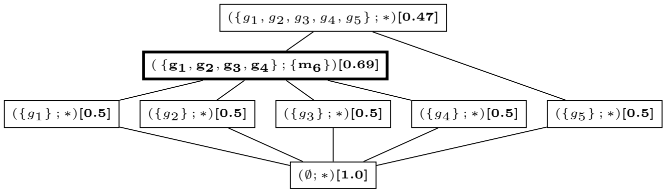
In data mining, many usefulness measures of patterns are introduced. For example, more than 30 statistical methods are enumerated and discussed in [1]. Such a high number of different approaches to pattern selection emphasizes the importance of the problem. In this paper we would like to focus on a measure which is introduced within Formal Concept Analysis (FCA). FCA is a mathematical formalism having many applications in data analysis [2]. Starting from the set of objects and the corresponding sets of attributes FCA tends to generalize the descriptions for any set of objects. Although this approach is less efficient than the statistical methods it is still feasible and ensures that no potentially interesting pattern is missed.

Within FCA there are several approaches for pattern selection. Two disjoint approaches can be distinguished. The first one is to introduce background knowledge into the procedure computing concepts [3–5]. These approaches allow one to find patterns which are likely to be useful for the current task. Although the number of resulting patterns can be significantly reduced, they are still numerous. The second approach can be applied in a composition with the first ones, ranking the resulting patterns w.r.t. a relevance measure.

The authors of [6] provide several measures for ranking concepts that stem from the algorithms possibly underlying human behavior. Stability is another

	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	

(a)


$(\{g_1, g_2, g_3, g_4, g_5\} ; *)$[0.47]
$(\{\mathbf{g_1}, \mathbf{g_2}, \mathbf{g_3}, \mathbf{g_4}\} ; \{\mathbf{m_6}\})$[0.69]
$(\{g_1\} ; *)$[0.5]
$(\{g_2\} ; *)$[0.5]
$(\{g_3\} ; *)$[0.5]
$(\{g_4\} ; *)$[0.5]
$(\{g_5\} ; *)$[0.5]
$(\emptyset; *)$[1.0]

(b)

Fig. 1: A toy formal context (a) and the correspnoding concept lattice with stability indexes (b).

measure for ranking concepts, introduced in [7] and later revised in [8–10]. Several other methods are considered in [11], where it is shown that stability is more reliable in noisy data. For the moment, stability seems to be the most widely used usefulness measure around the FCA community. Thus, in this paper we are going to focus on stability. Although this measure is often used, there is neither a reliable comparison nor a deep research on its usefulness. Consequently, the goal of this paper is to evaluate the usefulness of stability. Here we experimentally prove that the stability for a pattern is coherent with the stability computed for the same pattern but w.r.t. a different dataset coming from the same population (the similarly distributed dataset).

The rest of the paper is organised as follows. Section 2 introduces definition of stability and discusses known stability estimates. In Section 3 experiments on relevancy of stability are discussed.

2 Stability of a formal concept

2.1 Formal concept analysis (FCA)

FCA [2] is a formalism for data analysis. FCA starts with a formal context and builds a set of formal concepts organized within a concept lattice. A formal context is a triple (G, M, I), where G is a set of objects, M is a set of attributes and I is a relation between G and M, $I \subseteq G \times M$. In Figure 1a, a formal context is shown. A Galois connection between G and M is defined as follows:

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

The Galois connection maps a set of objects to the maximal set of attributes shared by all objects and reciprocally. For example, $\{g_1, g_2\}' = \{m_6\}$, while $\{m_6\}' = \{g_1, g_2, g_3, g_4\}$.

Definition 1. *A formal concept is a pair (A, B), where A is a subset of objects, B is a subset of attributes, such that $A' = B$ and $A = B'$, where A is called the extent of the concept, and B is called the intent of the concept.*

For example, a pair $(\{g_1, g_2, g_3, g_4\}; \{m_6\})$ is a formal concept. Formal concepts can be partially ordered w.r.t. the extent inclusion (dually, intent inclusion). For example, $(\{g_3\}; \{m_3, m_6\}) \leq (\{g_1, g_2, g_3, g_4\}; \{m_6\})$. This partial order of concepts is shown in Figure 1b.

2.2 The definition of stability

Stability is an interestingness measure of a formal concept introduced in [7] and later revised in [8, 10].

Definition 2. *Given a concept c, concept stability $Stab(c)$ is defined as*

$$Stab(c) := \frac{|\{s \in \wp(Ext(c)) \mid s' = Int(c)\}|}{2^{|Ext(c)|}} \tag{1}$$

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

Example 1. Figure 1b shows the concept lattice of the context in Figure 1a, for simplicity some intents are not given. The extent of the highlighted concept c is $Ext(c) = \{g_1, g_2, g_3, g_4\}$, thus, its power set contains 2^4 elements. The descriptions of 5 subsets of $Ext(c)$ ($\{g_1\}, \ldots, \{g_4\}$ and $\emptyset$) are different from $Int(c) = \{m_6\}$, while all other subsets of $Ext(c)$ have a description equal to $\{m_6\}$. So, $Stab(c) = \frac{2^4-5}{2^4} = 0.69$.

Stability measures the dependence of a concept intent on objects of the concept extent. In [10] it is shown that stability of a concept c is the relative number of subcontexts where there exists the concept c with intent $Int(c)$. A stable concept can be found in many such subcontexts, and therefore is likely to be found in an unrelated context built from the population under study.

In some papers it is noticed that in large datasets most of the concepts tends to have stability close to 1 [12, 13]. Thus, in order to distinguish between them we use the following logarithmic stability:

$$LStab(c) = -\log_2(1 - Stab(c)) \tag{2}$$

Stability computation is #P-complete [7, 8]. In this paper we rely on the algorithm from [10], with a worst-case complexity of $O(L^2)$, where L is the size of the concept lattice. However, generally it is quite efficient on real data.

3 Experiment on relevancy of stability

Experiments on behaviour of stability are carried out on public datasets available from the UCI repository [14]. These datasets are shown in Table 1. With their different size and complexity, these datasets provide a rich experimental basis. Complexity here stands for the size of the concept lattice given the initial number

Table 1: Datasets used in the experiments. Column 'Shortcut' refers to the short name of the dataset used in the rest of the paper; 'Size' is the number of objects in the dataset; 'Max. Size' is the maximal number of objects in a random subset of the dataset the concept lattice can be computed for; 'Max. Lat. Size' is the size of the corresponding concept lattice; 'Lat. Time' is the time in seconds for computing this lattice; 'Stab. Time' is the time in seconds to compute stability for every concept in the maximal lattice.

Dataset	Shortcut	Size	Max. Size	Max. Lat. Size	Lat. Time	Stab. Time
Mushrooms[1]	Mush	8124	8124	$2.3 \cdot 10^5$	324	57
Plants[2]	Plants	34781	1000	$2 \cdot 10^6$	45	10^4
Chess[3]	Chess	3198	100	$2 \cdot 10^6$	30	$7.4 \cdot 10^3$
Solar Flare (II)[4]	Flare	1066	1066	2988	0	0
Nursery[5]	Nurs	12960	12960	$1.2 \cdot 10^5$	245	5

[1] `http://archive.ics.uci.edu/ml/datasets/Mushroom`

[2] `http://archive.ics.uci.edu/ml/machine-learning-databases/plants/`

[3] `http://archive.ics.uci.edu/ml/datasets/Chess+(King-Rook+vs.+King-Pawn)`

[4] `http://archive.ics.uci.edu/ml/datasets/Solar+Flare`

[5] `http://archive.ics.uci.edu/ml/datasets/Nursery`

of objects in the corresponding context. For example, `Chess` is the most complex dataset as for only 100 objects in the context there are already $2 \cdot 10^6$ of concepts in the concept lattice.

When computing stability, one wants to know if the intent of a stable concept is a general characteristic rather than an artefact specific for a dataset. For that it is necessary to evaluate stability w.r.t. a test dataset different from the reference one. Reference and test datasets are two names of disjoint datasets on which the stability behaviour is evaluated. In order to do that the following scheme of experiment is developed:

1. Given a dataset $\mathbb{K}$ of size K objects, experiments are performed on dataset subsets whose size in terms of number of objects is N. This size is required to be at least half the size of K. For example, for a dataset of size $K = 10$ the size of it subset can be $N = 4$.
2. Two disjoint dataset subsets $\mathbb{K}_1$ and $\mathbb{K}_2$ of size N (in terms of objects) of dataset $\mathbb{K}$ are generated by sampling, e.g., $\mathbb{K}_1 = \{g_2, g_5, g_6, g_9\}$ and $K_2 = \{g_3, g_7, g_8, g_{10}\}$. Later, $\mathbb{K}_1$ is used as a reference dataset for computing stability, while $\mathbb{K}_2$ is a test dataset for evaluating stability computed in $\mathbb{K}_1$.
3. The corresponding sets of concepts $\mathcal{L}_1$ and $\mathcal{L}_2$ with their stability are built for both datasets $\mathbb{K}_1$ and $\mathbb{K}_2$.
4. The concepts with the same intents in $\mathcal{L}_1$ and $\mathcal{L}_2$ are declared as corresponding concepts.
5. Based on this list of corresponding concepts, a list of pairs $S = \{\langle X, Y \rangle, \dots\}$ is built, where X is the stability of the concept in $\mathcal{L}_1$ and Y is the stability of the corresponding concept in $\mathcal{L}_2$. If an intent exists only in one dataset, its stability is set to zero in the other dataset (following the definition of

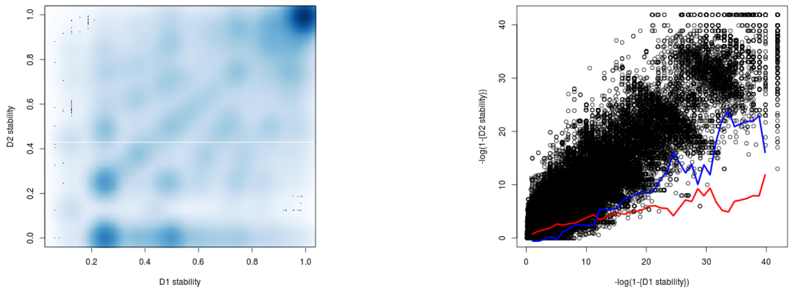


Fig. 2: Stability in the test dataset w.r.t the reference one in Mush4000 in (a) plane scale (b) logarithmic scale.

stability). Finally, the list $LS = \{\langle X_{\log}, Y_{\log}\rangle, \dots\}$ includes the stability pairs from S in logarithmic scale as stated by Eq. (2). We study here the sets S and LS.

The idea of evaluating stability computed on a reference dataset w.r.t. a test dataset comes from the supervised classification methods. Moreover, this idea is often used to evaluate statistical measures for pattern selection and can be found as a part of pattern selection algorithms with a good performance [15].

Sets of pairs S and LS can be drawn by matching every point $\langle X, Y\rangle$ to a point in a 2D-plot. The best case is $y = x$. It means that stability computed for dataset part $\mathbb{K}_1$ is exactly the same as stability computed for the dataset part $\mathbb{K}_2$. However, this is hardly the case in real-world experiments. For example, Figure 2a shows the corresponding diagram for the dataset `Mush4000`.[1] This figure also highlights the fact that many concepts have stability close to 1. However, when the logarithmic set LS is used, a blurred line $y = x$ can be perceived in Figure 2b. Moreover, selecting the concepts which are stable w.r.t. a high threshold in the reference dataset $\mathbb{K}_1$, the corresponding concepts in $\mathbb{K}_2$ are stable w.r.t. a lower threshold. Thus, we can conclude that stability is more tractable in the logarithmic scale, and, thus, we only consider this logarithmic scale in the rest of the paper.

3.1 Setting a stability threshold

In the previous subsection it is mentioned that concepts stable in the reference dataset are stable in the test dataset with a smaller threshold. *But what is "smaller"?* Imagine that in the reference dataset $\mathbb{K}_1$ we have the threshold θ_1, i.e., if $Stab(c) \geq \theta_1$ then c is stable, while in the $\mathbb{K}_2$ we have θ_2. Then, we want to know the threshold θ_1 such that at least 99% of stable concepts in $\mathbb{K}_1$

[1] From here, the name of a dataset followed by a number such as '$NameN$' refers to an experiment based on the dataset $Name$ where $\mathbb{K}_1$ and $\mathbb{K}_2$ are of the size N.

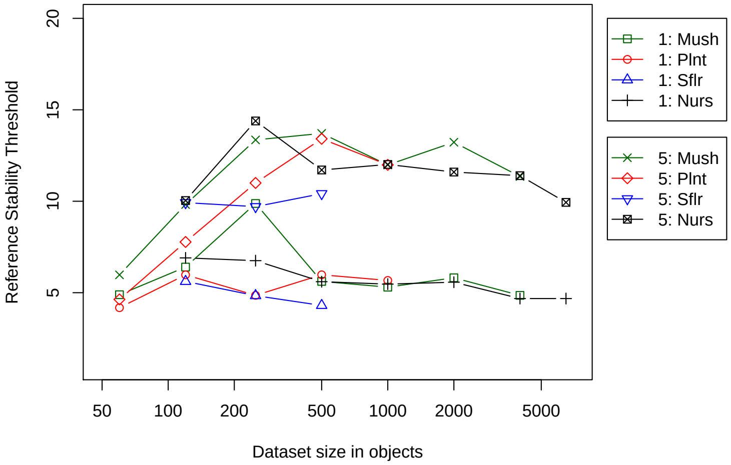


Fig. 3: Stability threshold in the reference dataset ensuring that 99% of concepts in the test datasets corresponding to stable concepts are stable with stability thresholds 1 or 5.

corresponds to stable concepts in $\mathbb{K}_2$. Figure 3 shows the reference threshold θ_1 (x-axis) w.r.t. the size of the datasets (y-axis) for $\theta_2 = 1$ and $\theta_2 = 5$. For example, the line '5: Mush' corresponds to the line of θ_1, where θ_2 is fixed to 5 w.r.t. to the size of the dataset built from dataset Mushrooms. The value $\theta_2 = 1$ means that any stable concept is just found in the test dataset, while $\theta_2 = 5$ requires that they are quite stable in the test dataset. We can see that for large datasets the stability threshold is independent of the dataset, while for small datasets the diversity is higher. We can see that the value of θ_1 should be set to 5–6 in order to ensure that 99% of stable concepts have corresponding concepts in another dataset.

3.2 Stability and ranking

Another way of using usefulness measures is pattern ranking. Thus, it is an interesting question if the order of patterns could be preserved by using stability. A way to study an order of an array ar is to compute its sorting rate r, i.e., the relative number of pairs in the array sorted in the ascending order: $r = 2 \cdot \frac{\{(i,j) | i<j \text{ and } ar_i \leq ar_j\}}{|ar| \cdot (|ar|-1)}$. A sorting rate equal to 1 means that the array is in the ascending order, while 0 means that it is in the descending order; the value 0.5 means that there is no order at all. Figure 4 shows the sorting rate (SR) for different datasets, i.e., the sorting rate of concept stabilities in $\mathbb{K}_2$, ordered w.r.t. stabilities of the corresponding stable concepts in K_1. We can see that SR for all datasets is slowly increasing preserving nearly the same value along the stability threshold in $\mathbb{K}_1$. And, thus, concept stability can be used to rank concepts.

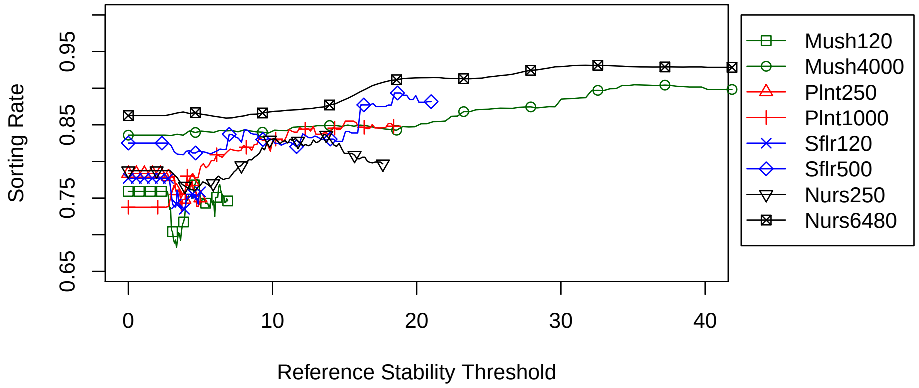


Fig. 4: Global sorting rate for different datasets.

4 Conclusion

In this paper we study concept stability as an efficient measure for pattern selection. It is shown that stability computed in the logarithmic scale is more convenient since it allows one to better distinguish stable concepts. Given a threshold of stability, patterns whose stability are above a threshold in a given dataset are likely to have stability above a smaller threshold in another dataset coming from the same distribution. However, independently of a dataset, as found experimentally, a concept should have logarithmic stability more than 5 in order to reflect any property of the population. We also show that stability is able to sort concepts in two independent datasets with nearly the same order by selecting concepts with stability above a certain threshold.

There are many future research directions. The found properties of stability suggest that interesting concepts can be found by resampling, i.e., analyzing many small parts of a large dataset, thus providing a key to an efficient processing of datasets with Formal Concept Analysis. The second important direction is to develop a methodology for comparison of stability and other known approaches for pattern selection.

Acknowledgements

This research was supported by the Basic Research Program at the National Research University Higher School of Economics (Moscow, Russia) and by the BioIntelligence project (France).

References

1. Masood, A., Soong, S.: Measuring Interestingness – Perspectives on Anomaly Detection. Computer Engineering and Intelligent Systems **4**(1) (2013) 29–40

2. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. 1st edn. Springer (1999)
3. Bělohlávek, R., Vychodil, V.: Formal Concept Analysis with Constraints by Closure Operators. In Schärfe, H., Hitzler, P., Ohrstrom, P., eds.: Conceptual Structures: Inspiration and Application. Volume 4068 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2006) 131–143
4. Belohlavek, R., Vychodil, V.: Formal Concept Analysis With Background Knowledge: Attribute Priorities. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **39**(4) (July 2009) 399–409
5. Buzmakov, A., Egho, E., Jay, N., Kuznetsov, S.O., Napoli, A., Raïssi, C.: On Projections of Sequential Pattern Structures (with an application on care trajectories). In: Proc. 10th International Conference on Concept Lattices and Their Applications. (2013) 199–208
6. Belohlavek, R., Trnecka, M.: Basic Level in Formal Concept Analysis: Interesting Concepts and Psychological Ramifications. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. IJCAI'13, AAAI Press (August 2013) 1233–1239
7. Kuznetsov, S.O.: Stability as an Estimate of the Degree of Substantiation of Hypotheses on the Basis of Operational Similarity. Automatic Documentation and Mathematical Linguistics (Nauch. Tekh. Inf. Ser. 2) **24**(6) (1990) 62–75
8. Kuznetsov, S.O.: On stability of a formal concept. Annals of Mathematics and Artificial Intelligence **49**(1-4) (2007) 101–115
9. Kuznetsov, S., Obiedkov, S., Roth, C.: Reducing the Representation Complexity of Lattice-Based Taxonomies. In Priss, U., Polovina, S., Hill, R., eds.: Conceptual Structures: Knowledge Architectures for Smart Applications. Volume 4604 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2007) 241–254
10. Roth, C., Obiedkov, S., Kourie, D.G.: On succinct representation of knowledge community taxonomies with formal concept analysis. International Journal of Foundations of Computer Science **19**(02) (April 2008) 383–404
11. Klimushkin, M., Obiedkov, S.A., Roth, C.: Approaches to the Selection of Relevant Concepts in the Case of Noisy Data. In: Proc. of the 8th International Conference on Formal Concept Analysis. ICFCA'10, Springer (2010) 255–266
12. Jay, N., Kohler, F., Napoli, A.: Analysis of Social Communities with Iceberg and Stability-Based Concept Lattices. In Medina, R., Obiedkov, S., eds.: Formal Concept Analysis. Volume 4933 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2008) 258–272
13. Buzmakov, A., Kuznetsov, S.O., Napoli, A.: Scalable Estimates of Concept Stability. In Glodeanu, C.V., Kaytoue, M., Sacarea, C., eds.: Formal Concept Analysis. Volume 8478 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2014) 157–172
14. Frank, A., Asuncion, A.: UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. University of California, Irvine, School of Information and Computer Sciences (2010)
15. Webb, G.I.: Discovering Significant Patterns. Machine Learning **68**(1) (2007) 1–33