Approaches to the Selection of Relevant Concepts in the Case of Noisy Data

Mikhail Klimushkin[1], Sergei Obiedkov[1], and Camille Roth[2]

[1] Higher School of Economics, Moscow, Russia,
klim.mikhail@gmail.com, sergei.obj@gmail.com,
[2] CAMS (CNRS/EHESS), Centre d'Analyse et de Mathématique Sociales
EHESS, 54 Bd Raspail, F-75006 Paris, France
roth@ehess.fr

Abstract. Concept lattices built on noisy data tend to be large and hence hard to interpret. We introduce several measures that can be used in selecting relevant concepts and discuss how they can be combined together. We study their performance in a series of experiments.

1 Introduction

Formal Concept Analysis (FCA) as a categorization method aims at grouping objects described by common attributes. In this framework, a category is more precisely defined as a maximal set of objects sharing a maximal set of attributes. Such groupings are then gathered in a hierarchical, lattice-based structure which straightforwardly exhibits various relationships between categories and their sub- and super-categories. As such, this approach provides an ideal formalization of categories in terms of concepts as they are traditionally defined philosophically, i.e., concepts extensionally described by sets of entities and intensionally described by attribute sets.

The taxonomical structure of a given domain is therefore frequently expected to be naturally revealed by applying FCA to an empirical description of its objects and their attributes. However, in spite of these strong theoretical foundations, the translation of empirical data into clean and relatively readable structures remains a common issue. Indeed, FCA induces a potentially dreadful combinatorial complexity and the structures obtained even from small-sized datasets can become prohibitively huge. In this respect, noise constitutes a primary factor of complication as it favors the existence of many similar but distinct concepts, which may excessively inflate the lattice with superfluous information to the cost of significantly impaired readability.

Hence, displaying interesting patterns while removing useless and cumbersome information constitutes a crucial task when assuming a noisy dataset. Such patterns are admittedly those which are or would be the only ones to be found in an ideally clean (non-noisy) dataset. This issue has seemingly received little attention in the FCA literature, apart from methods targeted at simplifying lattices [1, 2, 5, 6, 10]. In this paper, we design noise filtering criteria to specifically account for the likeliness of a concept to exist because of noise rather than to reflect essential features of the underlying taxonomy. To proceed, we essentially aim at appraising the diverse efficacy of such indices on basic datasets altered by simple noise effects.

The paper is organized as follows: in Sect. 2, we recall the principles and notations of FCA. Section 3 introduces the various indices and their rationale, while Sect. 4 describes their application on noisy contexts and comments the corresponding results.

2 FCA Definitions and Related Work

Before proceeding, we briefly recall the FCA terminology [3]. Given a *(formal) context* $\mathbb{K} = (G, M, I)$, where G is called a set of *objects*, M is called a set of *attributes*, and the binary relation $I \subseteq G \times M$ specifies which objects have which attributes, the derivation operators $(\cdot)^I$ are defined for $A \subseteq G$ and $B \subseteq M$ as follows:

$$A^I = \{m \in M \mid \forall g \in A : gIm\};$$

$$B^I = \{g \in G \mid \forall m \in B : gIm\}.$$

In words, A^I is the set of attributes common to all objects of A and B^I is the set of objects sharing all attributes of B.

If this does not result in ambiguity, $(\cdot)'$ is used instead of $(\cdot)^I$. The double application of $(\cdot)'$ is a closure operator, i.e., $(\cdot)''$ is extensive, idempotent, and monotonous. Therefore, sets A'' and B'' are said to be *closed*.

A *(formal) concept* of the context (G, M, I) is a pair (A, B), where $A \subseteq G$, $B \subseteq M$, $A = B'$, and $B = A'$. In this case, we also have $A = A''$ and $B = B''$. The set A is called the *extent* and B is called the *intent* of the concept (A, B).

A concept (A, B) is a *subconcept* of (C, D) if $A \subseteq C$ (equivalently, $D \subseteq B$). In this case, (C, D) is called a *superconcept* of (A, B). We write $(A, B) \leq (C, D)$ and define the relations $\geq$, $<$, and $>$ as usual. If $(A, B) < (C, D)$ and there is no (E, F) such that $(A, B) < (E, F) < (C, D)$, then (A, B) is a *lower neighbor* of (C, D) and (C, D) is an *upper neighbor* of (A, B); notation: $(A, B) \prec (C, D)$ and $(C, D) \succ (A, B)$.

The set of all concepts ordered by $\leq$ forms a lattice, which is denoted by $\underline{\mathfrak{B}}(\mathbb{K})$ and called the *concept lattice* of the context $\mathbb{K}$. The relation $\prec$ defines edges in the *covering graph* of $\underline{\mathfrak{B}}(\mathbb{K})$. The meet and join in the lattice are denoted by $\wedge$ and $\vee$, respectively.

3 Indices for Concept Selection

3.1 Stability

The stability index describes the proportion of subsets of objects of a given concept whose closure is equal to the intent of this concept [6]. In other words, it is meant to capture how much a concept intent depends on particular objects of the extent: should some objects be removed from the concept extent, would the concept intent remain the same? In an extensional formulation, the index specifies how much a concept extent depends on intent attributes.

We thus distinguish between intensional and extensional stability indices σ_i and σ_e:

$$\sigma_i(A,B) = \frac{|\{C \subseteq A \mid C' = B\}|}{2^{|A|}}$$

$$\sigma_e(A,B) = \frac{|\{D \subseteq B \mid D' = A\}|}{2^{|B|}}$$

The intent of a concept with a high intensional stability index would be likely to exist even if we ignore several of its objects: it does not disappear if the intent of some of its objects is modified, e.g., if these objects lose some of the properties of this concept. Put differently, concepts relying on noisy objects and, therefore, not typical of a realistic category, are more likely to be unstable. Similarly, extensional stability helps isolating concepts that appear because of noisy attributes.

Given the covering graph of a concept lattice, computing stability for all concepts can be done using the algorithm presented in [9]. This algorithm is essentially quadratic in the number of concepts in the lattice, which may be prohibitively expensive for large lattices. On the other hand, this algorithm needs a lattice as input, and generation of lattices for very large datasets may be impractical, anyway. Hence, it would be useful to develop an algorithm that generates only stable concepts directly from the input context (perhaps, giving up on the exact computation of stability and computing only approximate estimates).

Below, we mainly rely on intensional stability, which we denote by σ for the sake of clarity.

3.2 Concept Probability

A concept that covers fewer objects is normally less intensionally stable than a concept covering more objects. Still, these rather specific concepts can correspond to interesting associations that should not be ignored in the analysis. To give such specific concepts a chance of surviving the stability test, we need to normalize the stability index. To this end, we introduce the notion of *concept probability*. The idea is that if a concept has low probability, but is still observed in the data, it may reflect an interesting dependency and should be taken into account.

The relation between the presence of some patterns and some simple features of 01-matrices has received some attention in the literature, mainly by appraising how some patterns could be artifactual with respect to a particular *a priori* knowledge on the structure of such matrices. These studies focus in particular on the effect of matrix marginals—i.e., the distributions of sums of rows and columns—on patterns. For instance, the so-called configuration model of [8] assumes a null-model of random matrices conserving only original marginals;

then, estimates the probability of some statistical features, generally related to the topology of the matrix interpreted as the adjacency matrix of a graph. Closer to FCA, [4] later proposed to estimate whether frequent itemsets could be due to chance by, again, relating their presence to marginals: patterns are subsequently said to be artifactual if they are found in comparable proportions in the original data and its randomized version.

The notion of concept probability essentially follows a relatively similar line of reasoning. For $m \in M$, denote by p_m the probability of an arbitrary object having the attribute m. For $B \subseteq M$, define p_B, the probability of an arbitrary object having all attributes from B, by

$$p_B = \prod_{m \in B} p_m,$$

thus assuming the mutual independence of attributes. By denoting $n = |G|$, we obtain the following formula for the probability of B being closed:

$$\begin{aligned} p(B = B'') &= \sum_{k=0}^{n} p(|B'| = k, B = B'') = \\ &= \sum_{k=0}^{n} \left[\binom{n}{k} p_B^k (1 - p_B)^{n-k} \prod_{m \notin B} (1 - p_m^k) \right] \end{aligned}$$

To see the reasoning behind the formula, note that, for $|B'| = k$ and $B = B''$, we need that

1. There are k objects that have all attributes from B;
2. Each of the other $n - k$ objects does not have at least one attribute from B;
3. No attribute outside B belongs to all the k objects.

There are $\binom{n}{k}$ variants to choose k objects. The probability that each of the chosen k objects has all attributes from B is p_B^k, and the probability that each of the other $n - k$ objects does not have at least one attribute from B is $(1 - p_B)^{n-k}$. The probability that not all of the k chosen objects have an attribute m is $(1 - p_m^k)$; hence the probability that none of the attributes outside B belongs to all the k objects is $\prod_{m \notin B}(1 - p_m^k)$. Therefore, the joint probability of $|B'| = k$ and $B = B''$ is

$$\binom{n}{k} p_B^k (1 - p_B)^{n-k} \prod_{m \notin B} (1 - p_m^k).$$

By summing over k, we obtain the complete formula for the probability of an attribute set B being closed, which is given above.

Again, one can regard this as "intensional probability" of a concept and define "extensional probability" dually (as the probability of an object subset being closed).

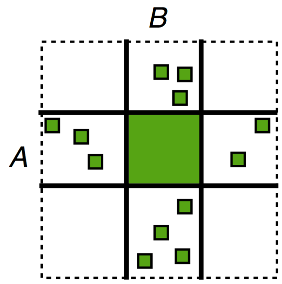


Fig. 1. Separation index $\mathfrak{s}(A, B)$ corresponds to the ratio between the green central area $|A| \times |B|$ and the total area covered by intents of every object of A and extents of every attribute of B.

It can be easily seen that it is possible to compute $\binom{n}{k+1} p_B^{k+1} (1 - p_B)^{n-(k+1)}$ from $\binom{n}{k} p_B^k (1 - p_B)^{n-k}$ using a constant number of multiplication operations, while $\prod_{m \notin B} (1 - p_m^k)$ requires $O(|G||M|)$ multiplications. Thus, computing the (intensional) probability of one concept involves $O(|G|^2|M|)$ multiplication operations.

3.3 Separation

The separation index is meant to describe how well a concept sorts out the objects it covers from other objects and, jointly, how well it sorts out the attributes it covers from other attributes of the context.[3]

Put differently, it indicates the significance of the difference between the objects covered by a given concept from other objects and, at the same time, between its attributes and other attributes.

To do so, the separation index $\mathfrak{s}$ is defined as the ratio between the area covered in the context by a concept (A, B) and the total area covered by its objects and attributes (see Fig. 1 for an illustration):

$$\mathfrak{s}(A, B) = \frac{|A||B|}{\sum_{g \in A} |g'| + \sum_{m \in B} |m'| - |A||B|}$$

Obviously, $\mathfrak{s}(A, B)$ can be computed in $O(|G| + |M|)$ time (assuming that object intents and attribute extents are pre-computed).

[3] A similar motivation is behind “relevant bi-sets” described in [2], but our approach is different.

4 Reconstruction of Noisy Datasets

4.1 Noisy Contexts

We understand noise in all generality as a measurement discrepancy between an empirical (real) setting and what a given dataset says about it. Empirical data may be noisy for various reasons: it can be due for instance to a lack of precision in collecting or building the dataset by mistakenly adding extra attributes to some objects or omitting some objects in describing the extent covered by some attribute. Noise can also be understood as exceptions to a rule, when attempting to exhibit clear-cut joint object-attribute categories, i.e., noisy objects or attributes. We therefore distinguish two different types of noise:

(Type I) — either by altering every cell value in the context with some probability;
(Type II) — or by adding to the original context a given number or proportion of completely random objects or attributes.

4.2 Example Contexts

We used four simple contexts with rather basic structures. This includes a chain-based lattice (300 objects, 6 attributes), an antichain-based lattice (300 objects, 12 attributes), and two more elaborate structures respectively built upon 300 and 400 objects and 6 and 4 attributes. The corresponding lattices are represented in Fig. 2.

Every concept of these lattices contains many identical objects, often characterized by only a handful of attributes, sometimes, only one attribute.

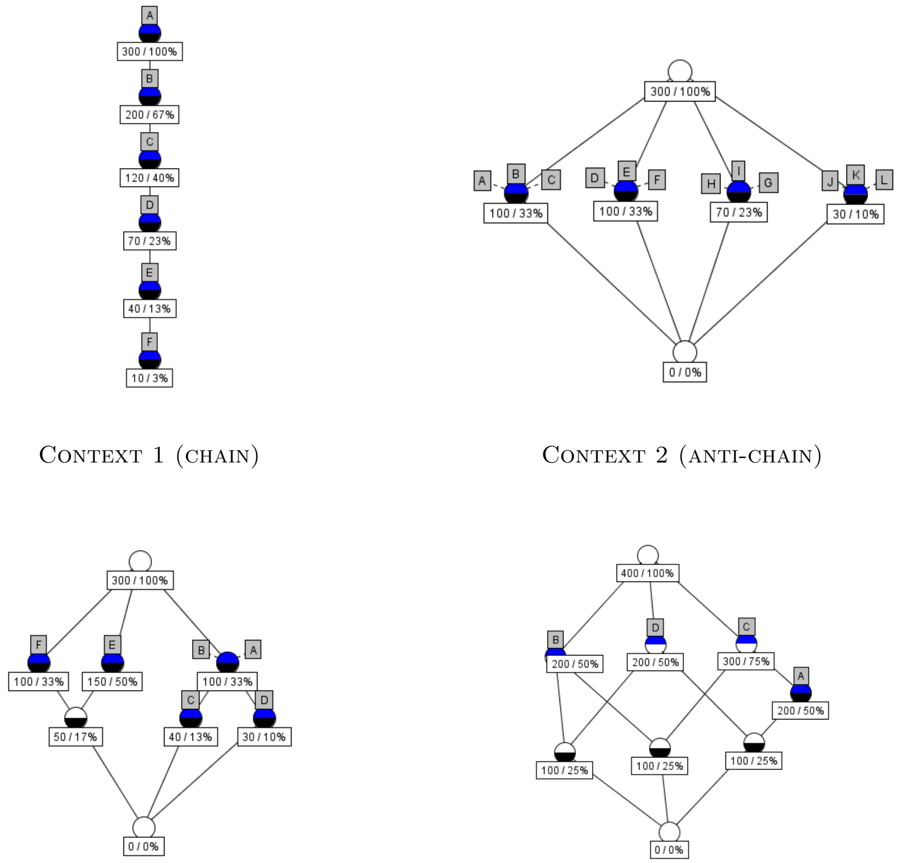
In each of these contexts, we uniformly introduce noise (type I or type II), which leads to the appearance of many new concepts, and then try to find the concepts of the original context among the concepts of the noisy context using various combinations of the indices discussed above.

4.3 Results

Context 1—Chain Lattice. Stability is relatively successful at dealing with type I-noise, with only a few discrepancies with the original lattice and up to 20% of perturbation (correctly selecting the original six intents among 56 intents of the noisy lattice)—see Fig. 3. On the contrary, separation indices are much less effective, even with only 2% of noise.

The best result (not shown in Fig. 3) is achieved with the combination of intensional stability and probability $\sigma(A, B) - k \cdot p(B = B'')$. With appropriately chosen value of $k \geq 0$, it makes it possible to completely restore the original structure even with 20% of noise (in this case, we had $k = 0.0005$).

The relevance of stability is confirmed when examining contexts altered with type II-noise, both when random objects or random attributes are introduced (using intensional stability in the former case and extensional stability in the



Context 1 (chain)

Context 2 (anti-chain)

Context 3

Context 4

Fig. 2. Original contexts

latter case). This should not come as a surprise, since intensional (extensional) stability is defined specifically to address the case of noisy objects (attributes) added to the otherwise clean context (that is, exactly type II-noise).

Context 2—Antichain Lattice. As regards type-I noise, when the noise is relatively limited (10%, which yields 324 concepts), separation is quite equivalent to stability: separation produces more faithful lattices, while stability yields more readable structures. When using thresholds on both stability and separation, the lattice exhibits exactly the same structure as the original. At 20% of noise (786 concepts), this does not work anymore. Yet, stability normalized by probability remarkably yields the original structure again (see Fig. 4).

Context 1, With Noise

type I			type II
separation $\mathfrak{s}$	stability σ		
2%	10%	20%	

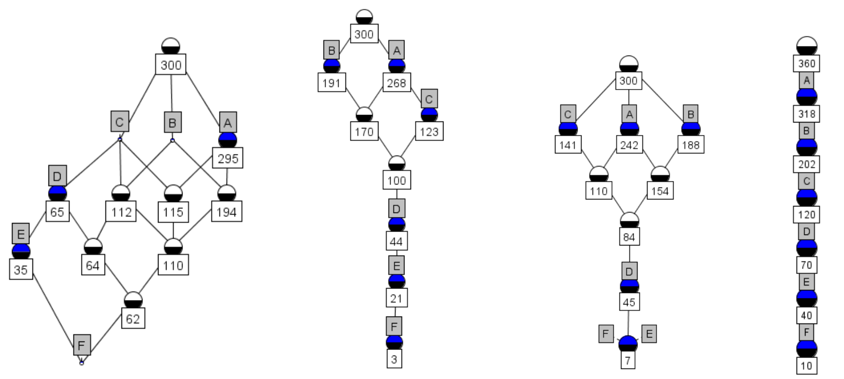


Fig. 3. Noisy instances of context 1 (chain lattice) with various filtering strategies. *Three first lattices, from left to right:* type-I noise using respectively 2, 10 and 20% noise, together with separation (left) or stability (middle lattices). The ideal result (not shown in the picture) is achieved by the combination of stability and probability $\sigma(A, B) - k \cdot p(B = B'')$. *Rightmost lattice:* type-II noise using 20% object addition and stability-based pruning.

As regards type-II noise, as with the chain lattice case, stability is robust, even with 40% of noise (268 concepts in the noisy lattice).

Context 3. This context features several concepts which are either super- or sub-categories in distinct parts of the lattice. Context 3 therefore significantly diverges from the previous prototypical cases.

As regards type I-noise, stability performs relatively well: it perfectly yields the original structure at 5% noise (36 concepts) and remains close to this benchmark at 10% (52 concepts) and even 20% (63 concepts) noise (see Fig. 5, top).

On the other hand, separation yields comparatively unconvincing results, from just 5% of noise (see Fig. 5, middle). Combining stability with separation and probability allows us to achieve better results (see Fig. 5, bottom).

As usual, stability works just fine for type II-noise, even with 40% of noise.

Context 4. This lattice eventually offers a mix of sub- and super-concepts intertwined in a somewhat balanced manner. Stability, again, is the only criterion that yields the exact original structure. Separation, when used alone or in various

Context 2, With Noise

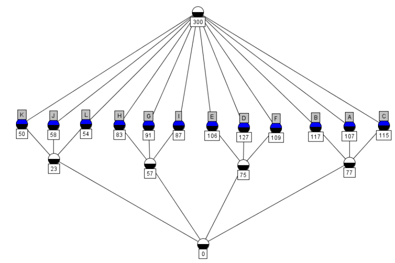
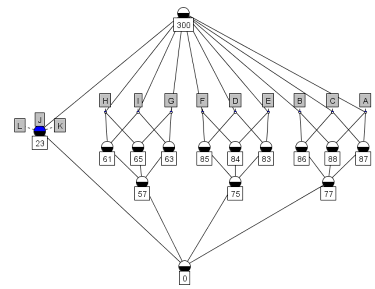
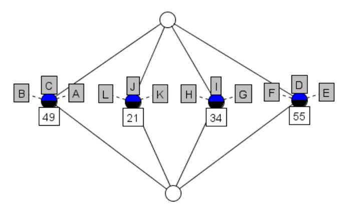
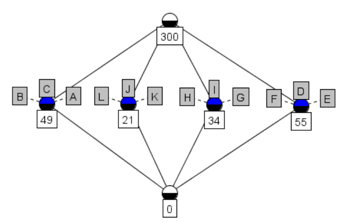
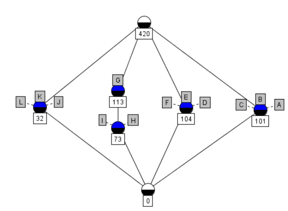
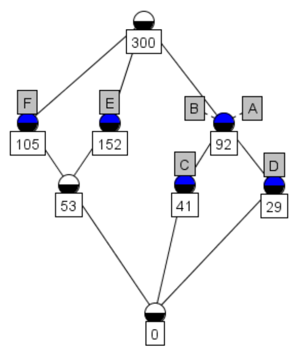
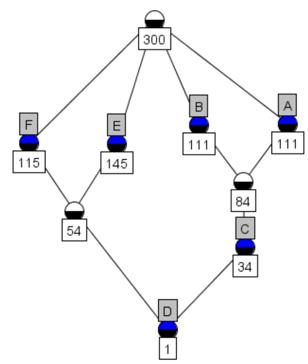
type I	stability σ	10%	
	separation $\mathfrak{s}$	10%	
	stability and separation $\sigma \cdot \mathfrak{s}$	20%	
	stability, separation, and probability $\sigma \; + k_1 \cdot \mathfrak{s} - k_2 \cdot p$	20%	
type II	stability σ	40%	

Fig. 4. Noisy instances of context 2 (antichain lattice) with various filtering strategies. *Four top lattices, from top to bottom:* type I noise using respectively 10, 10, 20, and 20% noise; together with stability, separation, the product of stability and separation, and the sum of weighted stability, separation, and probability. *Bottommost lattice:* type II-noise using 40% object addition and stability-based pruning.

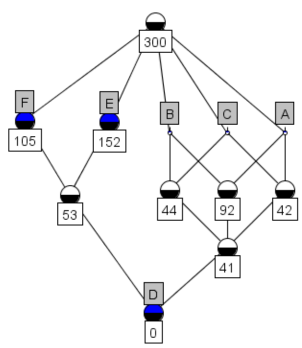
Context 3, With Noise of Type I

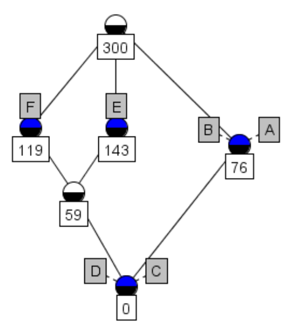
stability σ	
5%	10%



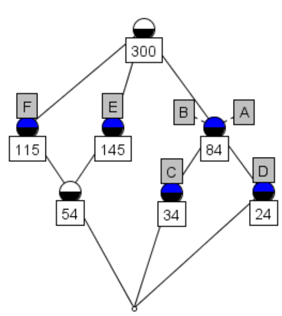


separation $\mathfrak{s}$	
5%	20%





$\sigma \ + k_1 \cdot s - k_2 \cdot p$	
10%	20%



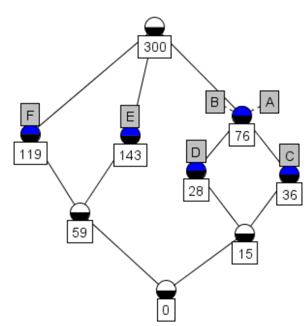


Fig. 5. Instances of context 3 with type I-noise and various filtering strategies. *Top lattices:* filtering lattices with contexts at 5% and 10% noise, using stability. *Middle lattices:* filtering with separation with 5% and 20% noise level. *Bottom lattices:* filtering with a combination of the three criteria with 10% and 20% noise level.

combinations with stability, produces a slightly different structure, as shown in Fig. 6 (this applies, e.g., when using the sum or the product of σ and $\mathfrak{s}$).

Context 4, With Noise of Type I	
stability σ	separation $\mathfrak{s}$
10%	

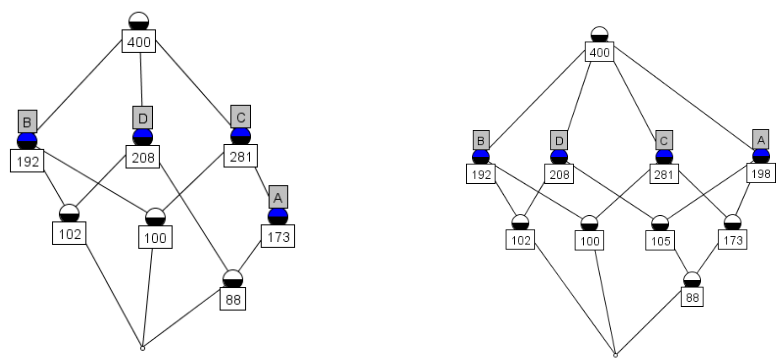


Fig. 6. Instances of context 4 with type I-noise at 10%, filtering with stability (*left*) or separation (*right*).

Again, type II noise can be perfectly filtered out by stability alone.

4.4 Conclusion

On the whole, stability is remarkably effective at sorting out type II-noise. Since the most stable concepts are usually those which are the least affected by individual objects or attributes, the success of this criterion should be relatively unsurprising. Still, stability proves to be significantly helpful for type I-noise as well. Unlike type II-noise, type I-noise might make some original concepts disappear (in which case, none of the methods described in the paper is sufficient to reconstruct such concepts). However, the probability for a stable concept to disappear or become significantly less stable because of type I-noise is quite low, especially if the noise is relatively limited.[4] Additional experiments could help characterize this probability more precisely, thus explaining why stability behaves well even for type I-noise.

Separation is founded on a different rationale and, therefore, can be used only as an auxiliary criterion when dealing with noise. A notable exception is given by anti-chain-like parts of lattices, which can be efficiently reconstructed with separation. In other cases, it is less effective than stability. One of its shortcomings

[4] We are grateful to the anonymous reviewer for this remark.

is that concepts with similar extents and intents tend to have similar separation indices. As a result, separation-based filtering prunes groups of similar concepts (Fig. 4). It seems promising to combine separation and stability, as they reduce each other's drawbacks.

Similarly, probability is not a self-sufficient criterion for filtering noise. Low or high probability of a concept does not say much of its importance. It can only correct other indices, serving a normalizing measure for stability or separation. It seems that the most promising combination is $\sigma + k_1 \cdot \mathfrak{s} - k_2 \cdot p$, where $k_2 < 0.2$. Such criterion seems able to reconstruct the original structure where stability and separation alone are not sufficient. Learning appropriate values for k_1 and k_2 can be done separately for different datasets. It remains to study the best ways to automatically learn these coefficients.

A next step would consist in validating these approaches on empirical data and theoretical interpretation of the performance of various combinations of the indices on different structures. It would also be interesting to see how the indices behave when noise is introduced in a non-uniform way. Automated learning of the best performing combinations (e.g., with evolutionary programming) is another research direction. This would most likely require defining a distance between two lattices, so as to be able to precisely evaluate the quality of reconstructing original lattices.

Acknowledgements

The first two authors were supported by the Scientific Foundation of the State University–Higher School of Economics (projects 08-04-0022 and 10-04-0017). The second author was also supported by the Russian Foundation for Basic Research (project 08-07-92497-NTsNIL_a). Line diagrams were produced with Concept Explorer (http://conexp.sourceforge.net/), a software tool developed by Serhiy Yevtushenko.

References

1. Bayardo, Jr., R., Goethals, B., Zaki, M. (eds.): Proc. of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI 2004). CEUR-WS.org (2004)
2. Boulicaut, J.F., Besson, J.: Actionability and formal concepts: A data mining perspective. In: Medina and Obiedkov [7], pp. 14–31
3. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Berlin (1999)
4. Gionis, A., Mannila, H., Mielikäinen, T., Tsaparas, P.: Assessing data mining results via swap randomization. ACM Trans. Knowl. Discov. Data 1(3), 14 (2007)
5. Jay, N., Kohler, F., Napoli, A.: Analysis of social communities with iceberg and stability-based concept lattices. In: Medina and Obiedkov [7], pp. 258–272
6. 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: 15th Intl Conf on Conceptual Structures, ICCS 2007, Sheffield, UK. LNCS/LNAI, vol. 4604, pp. 241–254. Springer (July 2007)

7. Medina, R., Obiedkov, S.A. (eds.): Formal Concept Analysis, 6th International Conference, ICFCA 2008, Montreal, Canada, February 25-28, 2008, Proceedings, Lecture Notes in Computer Science, vol. 4933. Springer (2008)
8. Newman, M.E.J., Strogatz, S., Watts, D.: Random graphs with arbitrary degree distributions and their applications. Physical Review E 64(026118) (2001)
9. Roth, C., Obiedkov, S., Kourie, D.G.: Towards concise representation for taxonomies of epistemic communities. In: Yahia, S.B., Nguifo, E.M. (eds.) Proc. CLA 4th Intl. Conf. on Concept Lattices and their Applications. LNCS/LNAI, vol. 4923, pp. 240–255. Springer (2006)
10. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing iceberg concept lattices with TITANIC. Data & Knowledge Engineering 42, 189–222 (2002)