Approximating Concept Stability

Mikhail A. Babin and Sergei O. Kuznetsov

National Research University Higher School of Economics,
Pokrovskii bd. 11, 109028 Moscow, Russia
mikleb@yandex.ru, skuznetsov@hse.ru

Abstract. Concept stability was used in numerous applications for selecting concepts as biclusters of similar objects. However, scalability remains a challenge for computing stability. The best algorithms known so far have algorithmic complexity quadratic in the size of the lattice. In this paper the problem of approximate stability computation is analyzed. An approximate algorithm for computing stability is proposed. Its computational complexity and results of computer experiments in comparing stability index and its approximations are discussed.

Keywords: concept stability, approximate counting, computational complexity.

1 Introduction

The approaches to data analysis and data mining using concept lattices for clustering and ontology engineering often encounter the problem of the large number of concepts of a formal context. There may be exponentially many formal concepts wrt. the size of the underlying context, the problem of computing the number of formal concepts given a context being #P-complete [5]. Several indices were proposed for measuring concept quality, such as concept stability [1,6,8,9], probability and separation [13]. Stability was used in numerous applications for selecting concepts as biclusters of similar objects, e.g., in technical diagnostics [1], in detecting scientific subcommunities [9,11,10], in planing medical treatment [12,17], or in grouping French verbs [14,16,15]. In [13] the authors compared filtration based on various indices and their linear combinations for data recovery. Linear index combinations that showed the best performance in computer experiments on concept filtration use stability with large weights. However, a potential constraint for applying stability for large data is the complexity of its computation, shown to be #P-complete in [1,8]. Sergei Obiedkov et al. proposed [11] an algorithm for computing stability index for all concepts using the concept lattice. This algorithm was quite good in practical applications so far, but in the worst case its complexity is quadratic in the size of the lattice (which itself can be exponential in the input context size). In this paper we consider the problem of approximate stability computation. We propose an approach to approximation, consider its computational complexity and discuss results of computer experiments in comparing stability index and its approximations. The rest of the paper is organized as follows. In the next section we recall

F. Domenach, D.I. Ignatov, and J. Poelmans (Eds.): ICFCA 2012, LNAI 7278, pp. 7–15, 2012.

main definitions related to FCA and concept stability, in Section 3 we discuss the complexity of approximations of the number of all closed and nonclosed sets, in Section 4 we consider computation of stability and in Section 5 we discuss results of computer experiments.

2 Main Definitions

2.1 FCA

Here we briefly recall the FCA terminology [3]. Let G and M be sets, called the set of objects and attributes, respectively. Let I be a relation $I \subseteq G \times M$ between objects and attributes: for $g \in G, m \in M, gIm$ holds iff the object g has the attribute m. The triple $\mathbb{K} = (G, M, I)$ is called a *(formal) context.* If $A \subseteq G, B \subseteq M$ are arbitrary subsets, then a *Galois connection* is given by the following *derivation operators*:

$$A' = \{m \in M \mid gIm \ \forall g \in A\}$$

$$B' = \{g \in G \mid gIm \ \forall m \in B\}$$

The pair (A, B), where $A \subseteq G$, $B \subseteq M$, $A' = B$, and $B' = A$ is called a *(formal) concept (of the context* $\mathbb{K}$*)* with *extent* A and *intent* B (in this case we have also $A'' = A$ and $B'' = B$).

The operation $(\cdot)''$ is a closure operator [3], i.e. it is idempotent $(X'''' = X'')$, extensive $(X \subseteq X'')$, and monotone $(X \subseteq Y \Rightarrow X'' \subseteq Y'')$. Sets $A \subseteq G$, $B \subseteq M$ are called *closed* if $A'' = A$ and $B'' = B$. Obviously, extents and intents are closed sets. The set of attributes B is *implied by the set of attributes* A, or the implication $A \to B$ holds, if all objects from G that have all attributes from the set A also have all attributes from the set B, i.e. $A' \subseteq B'$. Implications obey the Armstrong rules:

$$\frac{}{A \to A}, \quad \frac{A \to B}{A \cup C \to B}, \quad \frac{A \to B, B \cup C \to D}{A \cup C \to D}.$$

A subset $X \subseteq M$ *respects* an implication $A \to B$ if $A \subseteq X$ implies $B \subseteq X$. Every set of implications $\mathfrak{J}$ on the set M defines a closure operator $(\cdot)^{\mathfrak{J}}$ on M, where a subset of M is closed iff it respects all implications from $\mathfrak{J}$

2.2 Stability

The notion of stability of a formal concept was first introduced in [1,8] and now is used in a slightly revised form from [9,11].

Definition 1. *Let* $\mathbb{K} = (G, M, I)$ *be a formal context and* (A, B) *be a formal concept of* $\mathbb{K}$*. The* (intensional) stability $\sigma_{in}(A, B)$*, or* $\sigma_{in}(A)$*, is defined as follows:*

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

The *extentional stability* is defined in the dual way:

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

Usually, when it does not lead to misunderstanding, subscripts $_{in}$ and $_{ex}$ are omitted.

The numerator of intensional stability $\gamma(A, B) = |C \subseteq A \mid C' = B|$ is the number of all generators of the concept (A, B), so

$$2^{|A|} = \sum_{(C,D) \leq (A,B)} \gamma(C, D)$$

and

$$\gamma(A, B) = \sum_{(C,D) \leq (A,B)} 2^{|C|} \mu((C, D), (A, B)),$$

where $\mu(x, y)$ is the Möbius function of the concept lattice. Thus, stability nominator is dual to powersets of extents of the concept lattice wrt. the Möbius function of the concept lattice. This is reflected in the algorithm from [11] for computing stability, which is implicitly based on inclusion-exclusion principle, like standard algorithms for computing the Möbius function of a lattice.

3 Approximation of the Number of Closed and Nonclosed Sets

Many counting problems in FCA are known to be #P-complete but it does not imply that they cannot be solved approximately in polynomial time. For example, the problem of counting satisfying assignments for a DNF (unlike the dual problem for CNF) can be solved approximately using so-called FPRAS [2]. A *randomized approximation scheme* for a counting problem $f \colon \Sigma^* \to \mathbb{N}$ (e.g., the number of formal concepts of a context) is a randomized algorithm that takes as input an instance $x \in \Sigma^*$ (e.g. a formal context $\mathbb{K} = (G, M, I)$) and an error tolerance $\varepsilon > 0$, and outputs a number $N \in \mathbb{N}$ such that, for every input instance x,

$$Pr[(1 - \varepsilon) f(x) \leq N \leq (1 + \varepsilon) f(x)] \geq \frac{3}{4}$$

If the time of randomized approximation scheme is polynomial in $|x|$ and ε^{-1}, then this algorithm is called *fully polynomial randomized approximation scheme*, or FPRAS.

Below, for the problem *Problem* we will denote the number of solutions of *Problem* on corresponding input (which will be clear from the context) by $|\#Problem|$.

Given a hypergraph $G = (V, \mathcal{E}),\ \mathcal{E} = \{E_1, \ldots, E_m\}$, a subset $U \subseteq V$ is called *independent set* if $E_i \nsubseteq U$ for any $1 \leq i \leq m$ and is called *coindependent set* if $U \nsubseteq E_i$ for any $1 \leq i \leq m$.

Problem 1. Counting independent set (#IS)
INPUT: A hypergraph G.
OUTPUT: The number of independent sets of all sizes of G

It is known that there is no FPRAS for #IS unless $RP = NP$ (see [4]) when the hypergraph is a simple graph. So we can see this problem is hard even when V and $\{\emptyset\}$ are not edges of the hypergraph.

We also need the formulation of the following problem.

Problem 2. Counting coindependent sets (#CIS)
INPUT: A hypergraph $\mathcal{G} = (V, \mathcal{E}),\ \mathcal{E} = \{E_1, \ldots, E_m\},\ E_i \subseteq V$.
OUTPUT: The number of coindependent sets of G.

Note that set $U \subset V$ is an independent set of a hypergraph $G = (V, \mathcal{E}),\ \mathcal{E} = \{E_1, \ldots, E_m\}$ iff $V \backslash U$ is a coindependent set of the hypergraph $G' = (V, \mathcal{E}'),\ \mathcal{E}' = \{V \setminus E_1, \ldots, V \setminus E_m\}$. Thus there is no FPRAS for #CIS, unless $RP = NP$.

Now we are ready to discuss complexity of the counting problems for nonclosed sets of a formal context, closed sets of the closure system given by implication base, and nonclosed sets of the closure system given by implication base.

Problem 3. Counting nonclosed sets (#NC)
INPUT: A formal context $\mathbb{K} = (G, M, I)$
OUTPUT: The number of sets $A \subset M$ that $A'' \neq A$

Proposition 2. *There is no FPRAS for #NC, unless* $RP = NP$

Proof. Consider any input instance $(V, \mathcal{E}),\ V = \{v_1, \ldots, v_n\}\ \mathcal{E} = \{E_1, \ldots, E_m\}$ of #CIS. From this instance we construct the formal context $\mathbb{K} = (G, V, I)$ with the set of object intents $\bigcup_{1 \leq i \leq m} E_i \cup \{E_i \setminus \{v_1\}\} \cup \{E_i \setminus \{v_2\}\} \cup \ldots \cup \{E_i \setminus \{v_n\}\}$. Obviously, the set $A \subseteq V$ is a coindependent set of hypergraph $(V, \mathcal{E})$ iff $A'' \neq A$ or $A = V$ for the context $\mathbb{K}$. Hence $|\#CIS| = |\#NC| + 1$. □

Problem 4. Counting closed sets of implication base ($\#C_{\mathfrak{J}}$)
INPUT: An inplication base $\mathfrak{J} = \{A_1 \to B_1, \ldots, A_m \to B_m\},\ A_i, B_i \subseteq M$
OUTPUT: The number of closed sets of $\mathfrak{J}$.

Proposition 3. *There is no FPRAS for* $\#C_{\mathfrak{J}}$*, unless* $RP = NP$

Proof. Consider any input instance $(V, \mathcal{E})$, $\mathcal{E} = \{E_1, \dots, E_m\}$ of #IS. From this instance let us construct the implication base $\mathfrak{J} = \{E_1 \to V, \dots, E_m \to V\}$ (implications are defined on the set V). Obviously, a set U is an independent set of hypergraph $(V, \mathcal{E})$ iff U is closed set of $\mathfrak{J}$ and $U \neq V$. Hence $|\#IS| = |\#C_{\mathfrak{J}}| - 1$. □

Since a closed set wrt. an implication base can be represented as a satisfying assignment of a Horn CNF, we immediately get

Corollary 1. There is no FPRAS for the counting problem of Horn CNF satisfiability ($\#Horn\ SAT$), unless $NP = RP$.

Problem 5. Counting nonclosed set of implication base ($\#NC_{\mathfrak{J}}$)
INPUT: An inplication base $\mathfrak{J} = \{A_1 \to B_1, \dots, A_m \to B_m\}$, $A_i, B_i \subseteq M$
OUTPUT: The number of nonclosed sets of $\mathfrak{J}$.

Proposition 4. *There is FPRAS for $\#NC_{\mathfrak{J}}$*

Proof. Consider an instance $\mathfrak{J} = \{A_1 \to B_1, \dots, A_m \to B_m\}$, $A_i, B_i \subseteq X$ of $\#NC_{\mathfrak{J}}$. Closed sets of implication base $\mathfrak{J}$ are in one-to-one correspondence with the satisfying truth assignments of the corresponding Horn CNF $f_{\mathfrak{J}}$. Thus nonclosed sets of $\mathfrak{J}$ are in one-to-one correspondence with satisfying truth assignments of DNF $\neg f_{\mathfrak{J}}$. There is a known FPRAS for the counting problem of satisfying assignments of a DNF [2]. □

It is worth to note that exact complexity of approximate counting of a closed set of a formal context is open, but it is known that this problem is complete in class $\#RH\Pi_1$ [4]. All of the above results of this section can be summarized in the following table.

Table 1. Complexity of closed/nonclosed sets counting

	#closed sets	#nonclosed sets
$cs(\mathbb{K})$	$\#RH\Pi_1$-complete	no FPRAS, unless $RP = NP$
$cs(\mathfrak{J})$	no FPRAS, unless $RP = NP$	FPRAS

$cs(\mathbb{K})$ denotes the case where a closure system is given by context $\mathbb{K}$.
$cs(\mathfrak{J})$ denotes the case where a closure system is given by implication base $\mathfrak{J}$.

4 Computation of Stability

Recall that exact computing of concept stability is an #P-complete problem[1,6,8]. Moreover, there is no FPRAS for computing stability, unless $RP = NP$. In order to show this fact consider the context from the proof of proposition 2. Clearly $\sigma(M) = (|\#NC|+1)/2^{|M|}$. Here we discuss how to approximate stability with bounded absolute error. By definition of stability, stability of an intent A of a formal context $\mathbb{K} = (G, M, I)$ equals to the probability that a closure of a random subset of A is equal to A, i.e. $\sigma(A) = Pr(X'' = A)$, when X is chosen uniformly and random from 2^A. Thus to estimate $\sigma(A)$ we can use a Monte Carlo method.

```
GetStability($A, N$)
answer ← 0
for $i \leftarrow 1$ to $N$
        do pick random subset $X$ of $A$
            if $X'' = A$
                then answer ← answer +1
answer ← $\frac{answer}{N}$
return answer
```

Recall Chernoff-Hoeffding theorem with simplified bounds [2].

Theorem 5 (Chernoff-Hoeffding). *Let $X_1, X_2, \ldots, X_N$ be independent and identically distributed random variables with $p = E(X_i)$. Then*

$$Pr(\frac{1}{N}\sum X_i \leq p - \varepsilon) \leq \exp{(-2\varepsilon^2 N)}$$

It is easy to get the following proposition which states that for sufficiently large $N = N(\varepsilon, \delta)$, the probability of $|\,answer - \sigma(A)| \geq \varepsilon$ is not greater than δ.

Proposition 6. *The Monte Carlo method yields an approximation to $\sigma(A)$ with probability at least $1 - \delta$ and absolute error ε provided*

$$N > \frac{1}{2\varepsilon^2}\ln\frac{2}{\delta}$$

Proof. If we take random variables to be $1 - X_i$, and substitute them in the inequality of the Chernoff-Hoeffding theorem, then we have $p = E(1 - X_i)$ and we get

$$Pr(\frac{1}{N}\sum X_i \geq p + \varepsilon) \leq \exp{(-2\varepsilon^2 N)}.$$

Hence

$$Pr(|\frac{1}{N}\sum X_i - p| \geq \varepsilon) \leq$$

$$\leq Pr(\frac{1}{N}\sum X_i \leq p - \varepsilon) + Pr(\frac{1}{N}\sum X_i \geq p + \varepsilon) \leq$$

$$\leq 2\exp{(-2\varepsilon^2 N)}.$$

Consider random variables X_i such that $X_i = 1$ iff $X'' = A$ in the i-th iteration of GetStability and $X_i = 0$ otherwise. Thus $\frac{1}{N}\sum X_i = answer$, where *answer* is returned by GetStability(A,N) at the ith iteration. Absolute error probability is $Pr(|\,answer - p| \geq \varepsilon) \leq 2\exp{(-2\varepsilon^2 N)} \leq \delta$. Hence $2\varepsilon^2 N \geq \ln\frac{2}{\delta}$. □

We can use results of this algorithm to select top approximate stable concepts using the following straightforward algorithm.

```
TopStableConcepts($\mathbb{K}, \gamma_0$)
answer $\leftarrow \emptyset$
for every concept $C = (A, A')$ of $\mathbb{K}$
    do if approxStability$(A) > \sigma_\theta$
        then answer $\leftarrow$ answer $\cup \{(A, A')\}$
return answer
```

5 Experimental Results

In this section we discuss experimental results in computing stability approximations for random contexts of various sizes and density. The results of the approximate stability computation on random contexts are presented in Figure 1 and Figure 2. The Y-axis (labeled as *Error*) gives the relative error

$$|S(\mathbb{K}, \tilde{\sigma}, \sigma_\theta) \mathit{\Delta} S(\mathbb{K}, \sigma, \sigma_\theta)|/|S(\mathbb{K}, \sigma, \sigma_\theta)|.$$

Here $S(\mathbb{K}, \sigma, \sigma_\theta)$ denotes the set of all concepts with stability $\sigma \geq \sigma_\theta$; $S(\mathbb{K}, \tilde{\sigma}, \sigma_\theta)$ denotes the set of all concepts with approximate stability $\tilde{\sigma} \geq \sigma_\theta$, where σ_θ is a parameter (*stability threshold*). For every pair $g \in G$, $m \in M$ of a random context $\mathbb{K} = (G, M, I)$ one has $(g, m) \in I$ with probability d called *context density*.

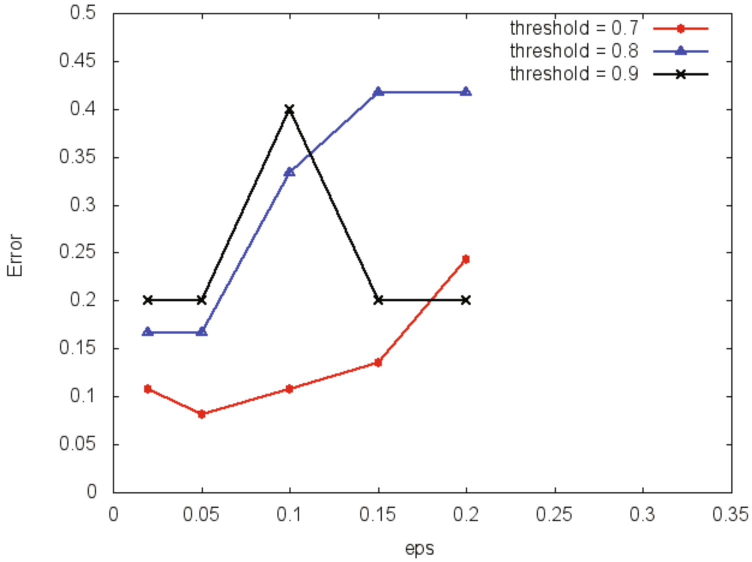

Fig. 1. Approximation quality for random contexts 100 × 30 with density 0.3

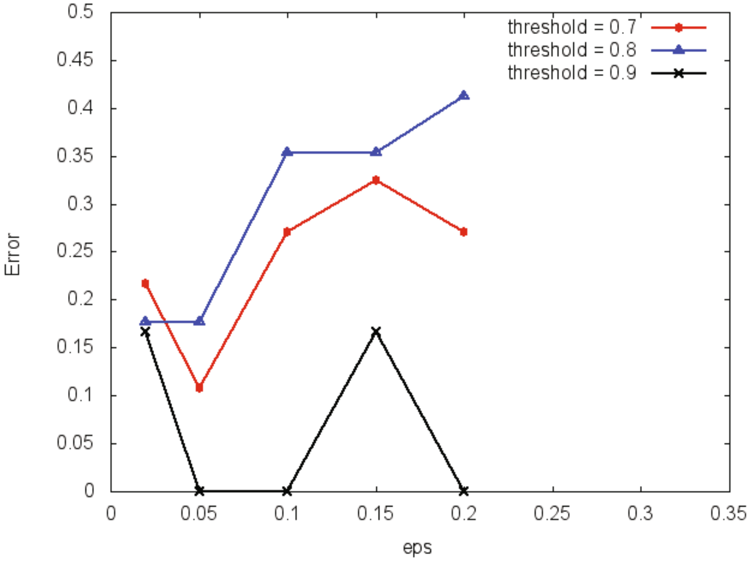

Fig. 2. Approximation quality for random contexts 150 × 30 with density 0.2

The results of computer experiments show that the algorithm for computing approximated stability algorithm has better precision when stability threshold is lower. This behaviour is consistent with theory, since when the stability threshold is high the number of stable concept is small and a small deviation of this threshold can result in significant change of the relative number of ”stable” concepts (i.e. concepts with approximate stability larger than threshold).

6 Conclusion

The problem of approximate stability computation was analyzed. Approximate solution of the problem was shown to be hard: the existence of FPRAS solving this problem would imply NP = RP. An approximate algorithm for computing stability, which can run in reasonable time for approximations with bounded absolute error was proposed. Its computational complexity and results of computer experiments in comparing stability index and its approximations were discussed. The results show that the approximations are better when stability threshold is low. Further study will be related to comparing approximate stability to other concept interestingness measures, such as independence, probability, wrt. computation time and selectiveness. Another challenging task would be the generation of interesting concepts without generating the set of all concepts.

Acknowledgments. This work was partially supported by Russian Foundation for Basic Research, project no. 08-07-92497-NTsNILa.

References

1. Kuznetsov, S.O.: Stability as an estimate of the degree of substantiation of hypotheses derived on the basis of operational similarity. Autom. Document. Math. Ling. (Nauchn. Tekh. Inf., Ser. 2) (12), 21–29 (1990)
2. Motwani, R., Raghavan, P.: Randomized Algorithms. Cambridge University Press (1995)
3. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Berlin (1999)
4. Dyer, M., Goldberg, L.A., Greenhill, C., Jerrum, M.: On the Relative Complexity of Approximate Counting Problems. In: Jansen, K., Khuller, S. (eds.) APPROX 2000. LNCS, vol. 1913, pp. 108–119. Springer, Heidelberg (2000)
5. Kuznetsov, S.O.: On Computing the Size of a Lattice and Related Decision Problems. Order 18(4), 313–321 (2001)
6. Kuznetsov, S.O.: Stability of a Formal Concept. In: San-Juan, E. (ed.) Proc. 4th Journee d'Informatique Messine (JIM 2003), Metz (2003)
7. Roth, C., Obiedkov, S., Kourie, D.: Towards Concise Representation for Taxonomies of Epistemic Communities. In: Yahia, S.B., Nguifo, E.M., Belohlavek, R. (eds.) CLA 2006. LNCS (LNAI), vol. 4923, pp. 240–255. Springer, Heidelberg (2008)
8. Kuznetsov, S.O.: On Stability of a Formal Concept. Annals of Mathematics and Artificial Intelligence 49, 101–115 (2007)
9. Kuznetsov, S.O., Obiedkov, S., Roth, C.: Reducing the Representation Complexity of Lattice-Based Taxonomies. In: Priss, U., Polovina, S., Hill, R. (eds.) ICCS 2007. LNCS (LNAI), vol. 4604, pp. 241–254. Springer, Heidelberg (2007)
10. Roth, C., Obiedkov, S., Kourie, D.: Towards Concise Representation for Taxonomies of Epistemic Communities. In: Yahia, S.B., Nguifo, E.M., Belohlavek, R. (eds.) CLA 2006. LNCS (LNAI), vol. 4923, pp. 240–255. Springer, Heidelberg (2008)
11. Obiedkov, S.A., Roth, C., Kourie, D.G.: On Succinct Representation of Knowledge Community Taxonomies with Formal Concept Analysis. Int. J. Found. Comput. Sci. 19(2), 383–404 (2008)
12. Jay, N., Kohler, F., Napoli, A.: Analysis of Social Communities with Iceberg and Stability-Based Concept Lattices. In: Medina, R., Obiedkov, S. (eds.) ICFCA 2008. LNCS (LNAI), vol. 4933, pp. 258–272. Springer, Heidelberg (2008)
13. Klimushkin, M., Obiedkov, S., Roth, C.: Approaches to the Selection of Relevant Concepts in the Case of Noisy Data. In: Kwuida, L., Sertkaya, B. (eds.) ICFCA 2010. LNCS, vol. 5986, pp. 255–266. Springer, Heidelberg (2010)
14. Falk, I., Gardent, C., Lorenzo, A.: Using Formal Concept Analysis to Acquire Knowledge about Verbs. In: Proc. 7th of the International Conference on Concept Lattices and Their Applications (CLA 2010), pp. 151–162 (2010)
15. Falk, I., Gardent, C.: Bootstrapping a Classification of French Verbs Using Formal Concept Analysis. In: Interdisciplinary Workshop on Verbs, Pisa, Italy (2011)
16. Falk, I., Gardent, C.: Combining Formal Concept Analysis and Translation to Assign Frames and Thematic Role Sets to French Verbs. In: Proc. International Conference Concept Lattices and Their Applications (CLA 2011), Nancy, France (2011)
17. Egho, E., Jay, N., Raissi, C., Napoli, A.: A FCA-based Analysis of Sequential Care Trajectories. In: Proc. 8th of the International Conference on Concept Lattices and Their Applications (CLA 2011), Nancy, France, pp. 362–376 (2011)