Concept interestingness measures: a comparative study

Sergei O. Kuznetsov[1] and Tatiana P. Makhalova[1,2]

[1]National Research University Higher School of Economics, Kochnovsky pr. 3,
Moscow 125319, Russia
[2]ISIMA, Complexe scientifique des Cézeaux, 63177 Aubière Cedex, France

skuznetsov@hse.ru,t.makhalova@gmail.com

Abstract. Concept lattices arising from noisy or high dimensional data have huge amount of formal concepts, which complicates the analysis of concepts and dependencies in data. In this paper, we consider several methods for pruning concept lattices and discuss results of their comparative study.

1 Introduction

Formal Concept Analysis (FCA) underlies several methods for rule mining, clustering and building taxonomies. When constructing a taxonomy one often deals with high dimensional or/and noisy data which results in a huge amount of formal concepts and dependencies given by implications and association rules. To tackle this issue different approaches were proposed for selecting most important or interesting concepts. In this paper we consider existing approaches which fall into the following groups: pre-processing of a formal context, modification of the closure operator, and concept filtering based on interestingness indices (measures). We mostly focus on comparison of interestingness measures and study their correlations.

2 FCA framework

Here we briefly recall FCA terminology [20]. A formal context is a triple (G, M, I), where G is called a set objects, M is called a set attributes and $I \subseteq G \times M$ is a relation called incidence relation, i.e. $(g, m) \in I$ if the object g has the attribute m. The derivation operators $(\cdot)'$ are defined for $A \subseteq G$ and $B \subseteq M$ as follows:

$$A' = \{m \in M | \forall g \in A : gIm\}$$
$$B' = \{g \in G | \forall m \in B : gIm\}$$

A' is the set of attributes common to all objects of A and B' is the set of objects sharing all attributes of B. The double application of $(\cdot)'$ is a closure operator,

© paper author(s), 2015. Published in Sadok Ben Yahia, Jan Konecny (Eds.): CLA 2015, pp. 59–72, ISBN 978–2–9544948–0–7, Blaise Pascal University, LIMOS laboratory, Clermont-Ferrand, 2015. Copying permitted only for private and academic purposes.

i.e. $(\cdot)''$ is extensive, idempotent and monotone. Sets $A \subseteq G$, $B \subseteq M$, such that $A = A''$ and $B = B''$ are said to be closed.

A (formal) concept is a pair (A, B), where $A \subseteq G$, $B \subseteq M$ and $A' = B$, $B' = A$. A is called the (formal) extent and B is called the (formal) intent of the concept (A, B). A partial order $\leqslant$ is defined on the set of concepts as follows: $(A, B) \leq (C, D)$ iff $A \subseteq C\,(D \subseteq B)$, a pair (A, B) is a subconcept of (C, D), while (C, D) is a superconcept of (A, B).

3 Methods for simplifying a lattice structure

With the growth of the dimension of a context the size of a lattice can increase exponentially, it becomes almost impossible to deal with the huge amount of formal concepts. With this respect a wide variety of methods have been proposed. Classification of them was presented in [16]. Authors proposed to divide techniques for lattice pruning into three classes: redundant information removal, simplification, selection. In this paper, we consider also other classes of methods and their application to concept pruning.

3.1 Pre-processing

Algorithms for concept lattice are time consuming. To decrease computation costs one can reduce the size of a formal context. Cheung and Vogel [13] applied Singular Value Decomposition (SVD) to obtain a low-rank approximation of Term-Document matrix and construct concept lattice using pruned concepts. Since this method is also computationally complex [25], alternative methods such as spherical k-Means [14] and fuzzy k-Means [17], Non-negative Matrix Decomposition [33] were proposed.

Dimensionality reduction can dramatically decrease the computational load and simplify the lattice structure, but in most cases it is very difficult to interpret the obtained results.

Another way to solve described problems without changing the dimension of the context was proposed in [18], where an algorithm that significantly improves the lattice structure by making small changes of context was presented. The central notion of the method is the concept incomparability w.r.t. $\leq$ relation. The goal of the proposed method is to diminish total incomparability of the concepts in the lattice.

The authors note that the result is close to that of fuzzy k-Means, but the former is achieved with fewer context changes than required by the latter. However, such transformations do not always lead to the decrease of a number of formal concepts, the transformations of a context are aimed at increasing the share of comparable concepts, thus this method does not ensure a significant simplification of the lattice structure.

Context pruning by clustering objects was introduced in [15]. The similarity of objects is defined as the weighted sum of shared attributes. Thus, the original context is replaced by the reduced one. Firstly, we need to assign weights w^m

for each attribute $m \in M$. The similarity between objects is defined as weighted sum of shared attributes.

Objects are considered similar if $sim(g, h) \geq \varepsilon$, where ε is a predefined threshold. In order to avoid the generation of large clusters another threshold α was proposed. Thus, the algorithm is an agglomerative clustering procedure, such that at each step clusters are brought together if the similarity between them is less than ε and the volume of clusters is less than $\alpha|G|$ objects.

3.2 Reduction based on a background knowledge or predefined constraints

Another approach to tackle computation and representation issues is to determine constraints on the closure operator. It can be done using background knowledge of attributes. In [8] the extended closure operator was presented. It is based on the notion of AD-formulas (attribute-dependency formulas), which establish dependence of attributes and their relative importance. Put differently, the occurrence of certain attributes implies that more important ones should also occur. Concepts which do not satisfy this condition are not included in the lattice.

In [5] a numerical approach to defining attribute importance was proposed. The importance of a formal concept can be defined by various aggregation functions (average, minimum, maximum) and different intent subsets (generator, minimal generator or intent itself). It was shown [5] that there is a correspondence between this numerical approach and AD-formulas.

Carpineto and Romano [12] considered document-term relation and proposed to use a thesaurus of terms to prune the lattice. Two different attributes are considered as same if there is a common ancestor in the hierarchy. To enrich the set of attributes they used a thesaurus, but in general, it may be quite difficult to establish such kind of relationship between arbitrary attributes.

Computing concepts with extents exceeding a threshold was proposed in [26] and studied in relation to frequent itemset mining in [34]. The main drawback of this approach, called "iceberg lattice" mining, is missing rare and probably interesting concepts.

Several polynomial-time algorithms for computing Galois sub-hierarchies were proposed, see [9, 3].

3.3 Filtering concepts

Selecting most interesting concepts by means of interestingness measures (indices) is the most widespread way of dealing with the huge number of concepts. The situation is aggravated by complexity of computing some indices. However, this approach may be fruitful, since it provides flexible tools for exploration of a derived taxonomy. In this section we consider different indices for filtering formal concepts. These indices can be divided into the following groups: measures designed to assess closed itemsets (formal concepts), arbitrary itemsets and measures for assessing the membership in a basic level (a psychology-motivated approach).

Indices for formal concepts

Stability Stability indices were introduced in [27, 28] and modified in [29]. One distinguishes intensional and extensional stability. The first one allows estimating the strength of dependence of an intent on each object of the respective extent. Extensional stability is defined dually.

$$Stab_i\left(A,B\right)=\frac{\left|\left\{C\subseteq A|C'=B\right\}\right|}{2^{|A|}}$$

The problem of computing stability is $\#P$-complete [28] and hence it makes this measure impractical for large contexts. In [4] its Monte Carlo approximation was introduced, a combination of Monte Carlo and upper bound estimate was proposed in [10]. Since for large contexts the stability is close to 1 [21] the logarithmic scale of stability (inducing the same ranking as stability) [10] is often used:

$$LStab\left(c\right)=-log_2\left(1-Stab\left(c\right)\right)$$

The bounds of stability are given by

$$\Delta_{min}\left(c\right)-log_2\left(|M|\right)\leq -log_2\sum_{d\in DD(c)}2^{-\Delta(c,d)}\leq LStab\left(c\right)\leq \Delta_{min}\left(c\right),$$

where $\Delta_{min}\left(c\right)=min_{d\in DD(c)}\Delta\left(c,d\right)$, $DD\left(c\right)$ is a set of all direct descendants of c in the lattice and $\Delta\left(c,d\right)$ is the size of the set-difference between extents of formal concepts c and d.

In our experiments we used the bounds of logarithmic stability, because the combined method is still computationally demanding.

Concept Probability Stability of a formal concept may be interpreted as probability of retaining its intent after removing some objects from the extent, taking that all subsets of the extent have equal probability. In [24] it was noticed that some interesting concepts with small number of object usually have low stability value. To ensure selection of interesting infrequent closed patterns, the concept probability was introduced. It is equivalent to the probability of a concept introduced earlier by R. Emilion [19].

The probability that an arbitrary object has all attributes from the set B is defined as follows

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

Concept probability is defined as the probability of B being closed:

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

where $n = |G|$.

The concept probability has the following probabilistic components: the occurrence of each attribute from B in all k objects, the absence of at least one attribute from B in other objects and the absence of other attributes shared by all k objects.

Robustness Another probabilistic approach to assessing a formal concept was proposed in [35]. Robustness is defined as the probability of a formal concept intent remaining closed while deleting objects, where every object of a formal context is retained with probability α. Then for a formal concept $c = (A, B)$ the robustness is given as follows:

$$r(c, \alpha) = \sum_{d \preceq c} (-1)^{|B_d| - |B_c|} (1 - \alpha)^{|A_c| - |A_d|}$$

Separation The separation index was considered in [24]. The main idea behind this measure is to describe the area covered by a formal concept among all nonzero elements in the corresponding rows and columns of the formal context. Thus, the value characterizes how specific is the relationship between objects and attributes of the concept with respect to the formal context.

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

Basic Level Metrics The group of so-called "basic level" measures was considered by Belohlavek and Trnecka [6, 7]. These measures were proposed to formalize the existing psychological approach to defining basic level of a concept [31].

Similarity approach (S) A similarity approach to basic level was proposed in [32] and subsequently formalized and applied to FCA in [6]. The authors defined basic level as combination of three fuzzy functions that correspond to formalized properties outlined by Rosch: high cohesion of concepts, considerably greater cohesion with respect to upper neighbor and a slightly less cohesion with respect to lower neighbors. The membership degree of a basic level is defined as follows:

$$BL_S = coh^{**}(A, B) \otimes coh^{**}_{un}(A, B) \otimes coh^{**}_{ln}(A, B),$$

where α_i is a fuzzy function that corresponds to the conditions defined above, $\otimes$ is t-norm [23].

A cohesion of a formal concept is a measure of pairwise similarity of all object in the extent. Various similarity measures can be used for cohesion functions:

$$sim_{SMC}(B_1, B_2) = \frac{|B_1 \cap B_2| + |M - (B_1 \cup B_2)|}{|M|}$$

$$sim_J(B_1, B_2) = \frac{|B_1 \cap B_2|}{|B_1 \cup B_2|}$$

The first similarity index sim_{SMC} takes into account the number of common attributes, while Jaccard similarity sim_J takes exactly the proportion of attributes shared by two sets. There are two ways to compute cohesion of formal concepts: taking average or minimal similarity among sets of attributes of the concept extent, the formulas are represented below (for average and minimal similarity respectively).

$$coh^{a}_{\dots}(A,B) = \frac{\sum_{x_1,x_2 \subseteq A, x_1 \neq x_2} sim_{\dots}(x'_1, x'_2)}{|A|(|A|-1)/2}$$

$$coh^{m}_{\dots}(A,B) = \min_{x_1,x_2 \in A} sim_{\dots}(x'_1, x'_2)$$

The Rosch's properties for upper and lower neighbors take the following forms:

$$coh^{a*}_{\dots,un}(A,B) = 1 - \frac{\sum_{c \in UN(A,B)} coh^{*}_{\dots}(c)/coh^{*}_{\dots}(A,B)}{|UN(A,B)|}$$

$$coh^{a*}_{\dots,ln}(A,B) = \frac{\sum_{c \in LN(A,B)} coh^{*}_{\dots}(A,B)/coh^{*}_{\dots}(c)}{|LN(A,B)|}$$

$$coh^{m*}_{\dots,un}(A,B) = 1 - \max_{c \in UN(A,B)} coh^{*}_{\dots}(c)/coh^{*}_{\dots}(A,B)$$

$$coh^{m*}_{\dots,ln}(A,B) = \min_{c \in LN(A,B)} coh^{*}_{\dots}(A,B)/coh^{*}_{\dots}(c)$$

where $UN(A,B)$ and $LN(A,B)$ are upper and lower neighbors of a formal concept (A,B) respectively.

As the authors noted, experiments revealed that the type of cohesion function does not affect the result, while the choice of similarity measure can greatly affect the outcome. More than that, in some cases upper (lower) neighbors may have higher (lower) cohesion than the formal concept itself (for example, some boundary cases, when a neighbors's extent (an intent) consists of identical rows (columns) of a formal context). To tackle this issue of non-monotonic neighbors w.r.t. similarity function authors proposed to take $coh^{**}_{\dots,ln}$ and $coh^{**}_{\dots,un}$ as 0, if the rate of non-monotonic neighbors is larger that a threshold.

In our experiments we used the following notation: SMC** and J**, where the first star is replaced by a cohesion type, the second one is replaced by the type of a similarity function. Below, we consider another four metrics that were introduced in [7].

Predictability approach (P) Predictability of a formal concept is computed in a quite similar way to BL_S. A cohesion function is replaced by a predictability function:

$$P(A,B) = pred^{**}(A,B) \otimes pred^{**}_{un}(A,B) \otimes pred^{**}_{ln}(A,B)$$

The main idea behind this approach is to assign high score to concept (A,B) with low conditional entropy of the presence of attributes not in B in intents of objects from A (i.e., requiring few attributes outside B in objects from A)[7]:

$$E\left(\mathbb{I}\left[\langle x,y\rangle \in I\right]|\mathbb{I}\left[x \in A\right]\right) = -\frac{|A \cap y'|}{|A|} log \frac{|A \cap y'|}{|A|}$$

$$pred\left(A,B\right) = 1 - \sum_{y \in M-B} \frac{E\left(\mathbb{I}\left[\langle x,y\rangle \in I\right]|\mathbb{I}\left[x \in A\right]\right)}{|M-B|}.$$

Cue Validity (CV), Category Feature Collocation (CFC), Category Utility (CU) The following measures based on the conditional probability of object $g \in A$ given that $y \subseteq g'$ were introduced in [7]:

$$CV\left(A,B\right) = \sum_{y \in B} P\left(A|y'\right) = \sum_{y \in B} \frac{|A|}{|y'|}$$

$$CFC\left(A,B\right) = \sum_{y \in M} p\left(A|y'\right) p\left(y'|A\right) = \sum_{y \in M} \frac{|A \cap y'|}{|y'|}\frac{|A \cap y'|}{|A|}$$

$$CU\left(A,B\right) = p\left(A\right) \sum_{y \in M} \left[p\left(y'|A\right)^2 - p\left(y'\right)^2\right] = \frac{|A|}{|G|} \sum_{y \in M} \left[\left(\frac{|A \cap y'|}{|y'|}\right)^2 - \left(\frac{|y'|}{|G|}\right)^2\right]$$

The main intuition behind CV is to express probability of extent given attributes from intent, CFC index takes into account the relationship between all attributes of the concept and intent of the formal concept, while CU evaluates how much an attribute in an intent is characteristic for a given concept rather than for the whole context [36].

Metrics for arbitrary itemsets

Frequency(support) It is one of the most popular measures in the theory of pattern mining. According to this index the most "interesting" concepts are frequent ones (having high support). For an arbitrary formal concept the support is defined as follows

$$supp\left(A,B\right) = \frac{|A|}{|G|}$$

The support provides efficient level-wise algorithms for constructing semilattices since it exhibits anti-monotonicity (a priori property [2, 30]):

$$B_1 \subset B_2 \rightarrow supp\left(B_1\right) \geq supp\left(B_2\right)$$

Lift In the previous section different methods with background knowledge were considered. Another way to add additional knowledge to data is proposed in [11]. Under assumption of attributes independence it is possible to compute individual frequencies of attributes and take their product as the expected frequency. The ratio of the observed frequency to its expectation is defined as *lift*. The lift of a formal concept (A, B) is defined as follows:

$$lift\left(B\right) = \frac{P\left(A\right)}{\prod_{b \in B} P\left(b'\right)} = \frac{|A|/|G|}{\prod_{b \in B} |b'|/|G|}$$

Collective Strength The collective strength [1] combines ideas of comparing the observed data and expectation under the assumption of independence of attributes. To calculate this measure for a formal concept (A, B) one needs to define for B the set of objects V_B that has at least one attribute in B, but not all of them at once. Denote $q = \prod_{b \in B} supp\left(b'\right)$ and $supp\left(V_B\right) = v$, the collective strength of a formal concept has the following form:

$$cs\left(B\right) = \frac{1-v}{v}\frac{q}{1-q}$$

4 Experiments

In this section, we compare measures with respect to their ability to help selecting most interesting concepts and filtering concepts coming from noisy datasets. For both goals, one is interested in a ranking of concepts rather than in particular values of the measures.

4.1 Formal Concept Mining

Usually concept lattices constructed from empirical data have huge amount of formal concepts, many of them being redundant, excessive and useless. In this connection the measures can be used to estimate how meaningful a concept is. Since the “interestingness” of a concept is a fairly subjective measure, the correct comparison of indices in terms of ability to select meaningful ones is impossible. With this respect we focus on similarity of indices described above. To identify how similar indices are, we use the Kendall tau correlation coefficient [22]. Put differently, we consider pairwise similarity of two lists of the same concepts that are ordered by values of the chosen indices. A set of strongly correlated measures can be replaced by one with the lowest computational complexity.

We randomly generated 100 formal contexts of random sizes. The number of attributes was in range between 10 and 40, while the number of objects varied from 10 to 70. For generated contexts we calculated pairwise Kendall tau for all indices of each context.The averaged values of correlations coefficients are represented in Table 1.

In [7] it was shown that the CU, CFC and CV are correlated, while S and P are not strongly correlated to other metrics. The results of our simulations allow us to conclude that CU, CFC and CV are also pairwise correlated to separation and support. Moreover, support is strongly correlated to separation and probability. Since the computational complexity of support is less than that of separation and probability, it is preferable to use support. It is worth noting that predictability (P) and robustness are not correlated to any other metrics and hence they can not be replaced by the metrics introduced so far.

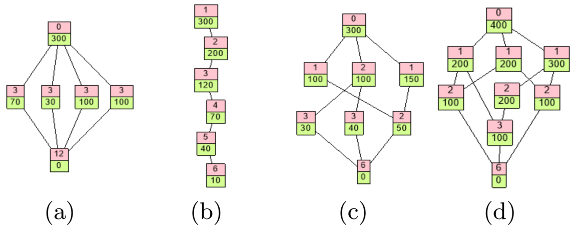
Thus, based on the correlation analysis, it is possible to reduce computationally complexity by choosing the most easily computable index within the class of correlated metrics.

Table 1. Kendall tau correlation coefficient for indices

	S_J^{mm}	S_J^{ma}	S_J^{am}	S_J^{aa}	S_{SMC}^{mm}	S_{SMC}^{ma}	S_{SMC}^{am}	S_{SMC}^{aa}	P	CU	CFC	CV	$Rob_{0.8}$	$Rob_{0.5}$	$Rob_{0.3}$	
Prob	0.18	0.15	0.14	0.14	0.04	0.03	0.00	-0.02	0.04	0.30	0.49	-0.01	-0.07	-0.11	-0.14	
Sep	0.20	0.20	0.18	0.18	0.07	0.07	0.14	0.12	0.05	0.36	0.45	0.54	-0.11	-0.12	-0.13	
CS	-0.08	-0.05	-0.06	-0.05	-0.07	-0.07	0.02	0.04	-0.09	0.04	-0.12	0.29	0.00	0.02	0.04	
Lift	-0.16	-0.13	-0.08	-0.07	-0.09	-0.08	0.02	0.03	-0.15	-0.07	-0.25	0.25	0.07	0.10	0.11	
Sup	0.17	0.17	0.21	0.21	-0.01	-0.02	0.03	0.00	-0.06	0.54	0.80	0.31	-0.10	-0.15	-0.18	
Stab	0.08	0.08	0.11	0.11	0.01	0.01	-0.02	-0.02	-0.18	-0.05	0.08	0.12	0.23	0.14	0.06	
$Stab_l$	0.06	0.06	0.11	0.11	0.02	0.02	0.01	0.01	-0.17	-0.16	-0.05	0.07	0.24	0.21	0.14	
$Stab_h$	0.15	0.14	0.15	0.14	0.02	0.01	-0.04	-0.05	-0.11	0.24	0.45	0.23	0.13	0.00	-0.09	
$Rob_{0.1}$	-0.09	-0.09	-0.02	-0.02	0.00	0.00	-0.01	0.00	-0.02	-0.11	-0.16	-0.09	0.56	0.73	0.86	
$Rob_{0.3}$	-0.10	-0.10	-0.03	-0.02	0.00	0.00	-0.02	0.00	-0.03	-0.12	-0.18	-0.09	0.68	0.86		
$Rob_{0.5}$	-0.08	-0.08	-0.02	-0.02	0.02	0.02	-0.02	-0.01	-0.03	-0.12	-0.15	-0.07	0.82			
$Rob_{0.8}$	-0.06	-0.06	-0.03	-0.02	0.03	0.03	-0.03	-0.02	-0.03	-0.11	-0.12	-0.06				
CV	0.08	0.09	0.15	0.15	-0.04	-0.04	0.05	0.05	-0.14	0.50	0.52					
CFC	0.09	0.08	0.15	0.15	-0.13	-0.13	-0.05	-0.06	-0.18	0.72						
CU	0.03	0.04	0.10	0.11	-0.13	-0.13	-0.06	-0.07	-0.17					-0.11		$Stab_h$
P	0.43	0.42	0.28	0.27	0.50	0.50	0.40	0.41					0.39	0.09		$Stab_l$
S_{SMC}^{aa}	0.39	0.39	0.56	0.56	0.49	0.50	0.92					0.86	0.59	0.03		Stab
S_{SMC}^{am}	0.39	0.38	0.58	0.57	0.48	0.49					0.18	0.02	0.58	-0.17		Sup
S_{SMC}^{ma}	0.51	0.50	0.37	0.37	0.96					-0.47	-0.04	0.05	-0.29	0.10		Lift
S_{SMC}^{mm}	0.51	0.48	0.36	0.36					0.64	-0.32	-0.09	-0.04	-0.25	0.03		CS
S_J^{aa}	0.41	0.42	0.95					0.14	0.01	0.42	0.03	-0.02	0.20	-0.13		Sep
S_J^{am}	0.42	0.41					0.17	-0.53	-0.73	0.76	0.15	0.02	0.48	-0.14		Prob
S_J^{ma}	0.90						Sep	CS	Lift	Sup	Stab	$Stab_l$	$Stab_h$	$Rob_{0.1}$		

4.2 Noise Filtering

In practice, we often have to deal with noisy data. In this case, the number of formal concepts can be very large and the lattice structure becomes too complicated [24]. To test the ability to filter out noise we took 5 lattices of different structure. Four of them are quite simple (Fig. 1) and the fifth one is the binarized fragment of the Mushroom data set [1] on 500 objects and 14 attributes, its concept lattice consists of 54 formal concepts.



(a) (b) (c) (d)

Fig. 1. Concept lattices for formal contexts with 300 objects and 6 attributes (a - c),with 400 objects and 4 attributes (d)

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

For a generated 0-1 datatable we changed table elements (0 to 1 and 1 to 0) with a given probability. The rate of noise (the probability of replacement) varied in the range from 0.05 to 0.5. We test the ability of a measure to filter redundant concepts in terms of precision and recall. For top-n (w.r.t. a measure) formal concepts, the recall and precision are defined as follows:

$$recall_{top-n} = \frac{|original \quad concepts_{top-n}|}{|original \quad concepts|}$$

$$precision_{top-n} = \frac{|original \quad concepts_{top-n}|}{|top-n \quad concepts|}$$

Table 2. Precision of indices with $recall = 0,6$

	Noise rate	Prob	Sep	Stab$_l$	Stab$_h$	CV	CFC	CU	Freq	$Rob_{0.5}$
Antichain	0.1	0.03	1	1	1	1	0.15	0.25	0.13	0.05
	0.3	0.03	1	1	1	1	0.09	0.20	0.10	0.02
	0.5	0.02	0.20	0.12	0.13	0.29	0.07	0.10	0.06	0.02
Chain	0.1	0.80	0.44	1	1	0.67	0.27	0.13	0.27	0.80
	0.3	0.21	0.18	0.67	1	0.22	0.18	0.17	0.19	1
	0.5	0.29	0.13	0.25	0.57	0.21	0.14	0.16	0.14	0.57
Context 3	0.1	0.20	1	1	1	0.36	0.33	0.44	0.67	0.40
	0.3	0.16	0.67	0.80	0.80	0.44	0.33	0.44	0.50	0.40
	0.5	0.19	0.50	0.50	0.50	0.44	0.27	0.33	0.50	0.57
Context 4	0.1	0.44	1.00	1.00	1.00	1.00	0.80	0.57	0.80	0.50
	0.3	0.22	1.00	1.00	1.00	1.00	0.80	0.57	0.80	0.57
	0.5	0.14	0.67	1.00	1.00	0.44	0.80	0.57	0.80	0.67
Mushroom	0.1	0.28	0.29	0.84	0.84	0.32	0.28	0.32	0.31	0.30
	0.3	0.16	0.16	0.36	0.39	0.25	0.18	0.20	0.22	0.09
	0.5	0.08	0.10	0.17	0.17	0.14	0.11	0.16	0.11	0.06

Figures 2 show the ROC curve for the measures. The curves that are close to the left upper corner correspond to the most powerful measures.

The best and most stable results correspond to the high estimate of stability (stability$_h$). The similar precision has the lower estimate of stability (Table 2), whereas precision of separation and probability depends on the proportion of noise and lattice structure as well. The measures of basic level that utilize similarity and predictability approaches become zero for some concepts. The rate of vanished concepts (including original ones) increases as the noise probability gets bigger. In our study we take such concepts as “false negative”, so in this case ROC curves do not pass through the point (1,1). More than that, recall and

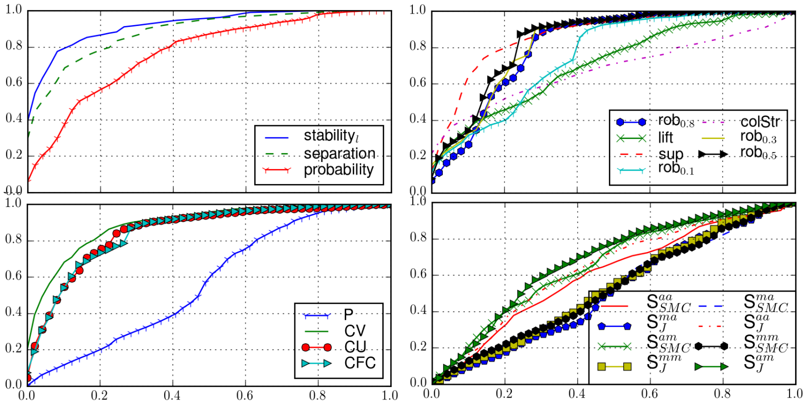


Fig. 2. Averaged ROC curves of indices among contexts 1 - 5 with different noise rate (0.1 - 0.5)

precision are unstable with respect to the noise rate and lattice structure. This group of measures is inappropriate for noise filtering.

The other basic level measures, such as CU, CFC and CV, demonstrate much better recall compared to previous ones. However, in general the precision of CU, CFC and CV is determined by lattice structure (Table 2).

Frequency has the highest precision among the indices that are applicable for the assessment of arbitrary sets of attributes. Frequency is stable with respect to the noise rate, but can vary under different lattice structures. For the lift and the collective strength precision depends on the lattice structure, and the collective strength also has quite unstable recall.

Precision of robustness depends on both lattice structure and value of α (Fig. 2). In our study we have got the highest precision for α close to 0.5.

Thus, the most preferred metrics for noise filtering are stability estimates, CV, frequency and robustness (where α is greater than 0.4).

In [24] it was noticed that the combination of the indices can improve the filtering power of indices. In this regard, we have studied top-n concepts selected by pairwise combination of measures. As it was shown by the experiments, the combination of measures may improve recall of the top-n set, while precision gets lower with respect to a more accurate measure. Figure 3 shows recall and precision of different combination of measures. In the best case it is possible to improve the recall, the precision on small sets of top-n concepts is lower than the precision of one measure by itself.

5 Conclusion

In this paper we have considered various methods for selecting interesting concepts and noise reduction. We focused on the most promising and well interpretable approach based on interestingness measures of concepts. Since "inter-

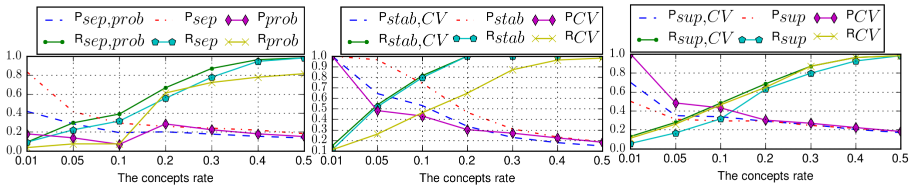


Fig. 3. Recall and precision of metrics and their combination on a Mushroom dataset fragment with the noise probability 0.1

estingness" of a concept is a subjective measure, we have compared several measures known in the literature and identified groups of most correlated ones. CU, CFC, CV, separation and frequency make up the first group. Frequency is correlated to separation and probability.

Another part of our experiments was focused on the noise filtering. We have found that the stability estimates work perfectly with data of various noise rate and different structure of the original lattice. Robustness and 3 of basic level metrics (cue validity, category utility and category feature collocation approaches) could also be applied to noise reduction. The combination of measures can also improve the recall, but only in the case of high noise rate.

Acknowledgments

The authors were supported by the project "Mathematical Models, Algorithms, and Software Tools for Mining of Structural and Textual Data" supported by the Basic Research Program of the National Research University Higher School of Economics.

References

1. Aggarwal, C.C., Yu, P.S.: A new framework for itemset generation. In: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. pp. 18–24. ACM (1998)
2. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proc. 20th int. conf. very large data bases, VLDB. vol. 1215, pp. 487–499 (1994)
3. Arévalo, G., Berry, A., Huchard, M., Perrot, G., Sigayret, A.: Performances of galois sub-hierarchy-building algorithms. In: Formal Concept Analysis, pp. 166–180. Springer (2007)
4. Babin, M.A., Kuznetsov, S.O.: Approximating concept stability. In: Domenach, F., Ignatov, D., Poelmans, J. (eds.) Formal Concept Analysis. Lecture Notes in Computer Science, vol. 7278, pp. 7–15. Springer Berlin Heidelberg (2012)
5. Belohlavek, R., Macko, J.: Selecting important concepts using weights. In: Valtchev, P., Jschke, R. (eds.) Formal Concept Analysis, Lecture Notes in Computer Science, vol. 6628, pp. 65–80. Springer Berlin Heidelberg (2011)
6. Belohlavek, R., Trnecka, M.: Basic level of concepts in formal concept analysis. In: Domenach, F., Ignatov, D., Poelmans, J. (eds.) Formal Concept Analysis, Lecture Notes in Computer Science, vol. 7278, pp. 28–44. Springer Berlin Heidelberg (2012)

7. 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. pp. 1233–1239. IJCAI '13, AAAI Press (2013)
8. Belohlavek, R., Vychodil, V.: Formal concept analysis with background knowledge: attribute priorities. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 39(4), 399–409 (2009)
9. Berry, A., Huchard, M., McConnell, R., Sigayret, A., Spinrad, J.: Efficiently computing a linear extension of the sub-hierarchy of a concept lattice. In: Ganter, B., Godin, R. (eds.) Formal Concept Analysis, Lecture Notes in Computer Science, vol. 3403, pp. 208–222. Springer Berlin Heidelberg (2005)
10. Buzmakov, A., Kuznetsov, S.O., Napoli, A.: Scalable estimates of concept stability. In: Glodeanu, C., Kaytoue, M., Sacarea, C. (eds.) Formal Concept Analysis, Lecture Notes in Computer Science, vol. 8478, pp. 157–172. Springer International Publishing (2014)
11. Cabena, P., Choi, H.H., Kim, I.S., Otsuka, S., Reinschmidt, J., Saarenvirta, G.: Intelligent miner for data applications guide. IBM RedBook SG24-5252-00 173 (1999)
12. Carpineto, C., Romano, G.: A lattice conceptual clustering system and its application to browsing retrieval. Machine Learning 24(2), 95–122 (1996)
13. Cheung, K., Vogel, D.: Complexity reduction in lattice-based information retrieval. Information Retrieval 8(2), 285–299 (2005)
14. Dhillon, I., Modha, D.: Concept decompositions for large sparse text data using clustering. Machine Learning 42(1-2), 143–175 (2001)
15. Dias, S.M., Vieira, N.: Reducing the size of concept lattices: The JBOS approach. In: Proceedings of the 7th International Conference on Concept Lattices and Their Applications, Sevilla, Spain, October 19-21, 2010. pp. 80–91 (2010)
16. Dias, S.M., Vieira, N.J.: Concept lattices reduction: Definition, analysis and classification. Expert Systems with Applications 42(20), 7084 – 7097 (2015)
17. Dobša, J., Dalbelo-Bašić, B.: Comparison of information retrieval techniques: latent semantic indexing and concept indexing. Journal of Inf. and Organizational Sciences 28(1-2), 1–17 (2004)
18. Düntsch, I., Gediga, G.: Simplifying contextual structures. In: Kryszkiewicz, M., Bandyopadhyay, S., Rybinski, H., Pal, S.K. (eds.) Pattern Recognition and Machine Intelligence, Lecture Notes in Computer Science, vol. 9124, pp. 23–32. Springer International Publishing (2015)
19. Emilion, R.: Concepts of a discrete random variable. In: Brito, P., Cucumel, G., Bertrand, P., de Carvalho, F. (eds.) Selected Contributions in Data Analysis and Classification, pp. 247–258. Studies in Classification, Data Analysis, and Knowledge Organization, Springer Berlin Heidelberg (2007)
20. Ganter, B., Wille, R.: Contextual attribute logic. In: Tepfenhart, W., Cyre, W. (eds.) Conceptual Structures: Standards and Practices, Lecture Notes in Computer Science, vol. 1640, pp. 377–388. Springer Berlin Heidelberg (1999)
21. 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, Lecture Notes in Computer Science, vol. 4933, pp. 258–272. Springer Berlin Heidelberg (2008)
22. Kendall, M.G.: A new measure of rank correlation. Biometrika pp. 81–93 (1938)
23. Klement, E.P., Mesiar, R., Pap, E.: Triangular norms. Springer Netherlands (2000)

24. 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.) Formal Concept Analysis, Lecture Notes in Computer Science, vol. 5986, pp. 255–266. Springer Berlin Heidelberg (2010)
25. Kumar, C.A., Srinivas, S.: Latent semantic indexing using eigenvalue analysis for efficient information retrieval. Int. J. Appl. Math. Comput. Sci 16(4), 551–558 (2006)
26. Kuznetsov, S.O.: Interpretation on graphs and complexity characteristics of a search for specific patterns. Automatic Documentation and Mathematical Linguistics 24(1), 37–45 (1989)
27. Kuznetsov, S.O.: Stability as an estimate of degree of substantiation of hypotheses derived on the basis of operational similarity. Nauchn. Tekh. Inf., Ser. 2 (12), 21–29 (1990)
28. Kuznetsov, S.O.: On stability of a formal concept. Annals of Mathematics and Artificial Intelligence 49(1-4), 101–115 (2007)
29. Kuznetsov, S.O., Obiedkov, S., Roth, C.: Reducing the representation complexity of lattice-based taxonomies. In: Conceptual Structures: Knowledge Architectures for Smart Applications, pp. 241–254. Springer Berlin Heidelberg (2007)
30. Mannila, H., Toivonen, H., Verkamo, A.I.: Efficient algorithms for discovering association rules. In: KDD-94: AAAI workshop on Knowledge Discovery in Databases. pp. 181–192 (1994)
31. Murphy, G.L.: The big book of concepts. MIT press (2002)
32. Rosch, E.: Principles of categorization pp. 27–48 (1978)
33. Snasel, V., Polovincak, M., Abdulla, H.M.D., Horak, Z.: On concept lattices and implication bases from reduced contexts. In: ICCS. pp. 83–90 (2008)
34. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing iceberg concept lattices with titanic. Data Knowl. Eng. 42(2), 189–222 (Aug 2002)
35. Tatti, N., Moerchen, F., Calders, T.: Finding robust itemsets under subsampling. ACM Transactions on Database Systems (TODS) 39(3), 20 (2014)
36. Zeigenfuse, M.D., Lee, M.D.: A comparison of three measures of the association between a feature and a concept. In: Proceedings of the 33rd Annual Conference of the Cognitive Science Society. pp. 243–248 (2011)