Finding Robust Itemsets Under Subsampling

Nikolaj Tatti
University of Antwerp
Antwerp, Belgium
nikolaj.tatti@ua.ac.be

Fabian Moerchen
Siemens Corporate Research, Integrated Data Systems
755 College Road East, Princeton, NJ, 08540, USA
fabian.moerchen@siemens.com

***Abstract*—Mining frequent patterns is plagued by the problem of pattern explosion making pattern reduction techniques a key challenge in pattern mining. In this paper we propose a novel theoretical framework for pattern reduction. We do this by measuring the robustness of a property of an itemset such as closedness or non-derivability. The robustness of a property is the probability that this property holds on random subsets of the original data. We study four properties: closed, free, non-derivable and totally shattered itemsets, demonstrating how we can compute the robustness analytically without actually sampling the data. Our concept of robustness has many advantages: Unlike statistical approaches for reducing patterns, we do not assume a null hypothesis or any noise model and the patterns reported are simply a subset of all patterns with this property as opposed to approximate patterns for which the property does not really hold. If the underlying property is monotonic, then the measure is also monotonic, allowing us to efficiently mine robust itemsets. We further derive a parameter-free technique for ranking itemsets that can be used for top-k approaches. Our experiments demonstrate that we can successfully use the robustness measure to reduce the number of patterns and that ranking yields interesting itemsets.**

***Keywords*-pattern reduction; robust itemsets; closed itemsets; free itemsets; non-derivable itemsets; totally shattered itemsets**

I. Introduction

Frequent itemset mining was first introduced in the context of market basket analysis [1] and has been used since to address many data mining problems such as frequent pattern mining, association rule generation [2], clustering [3], classification [4], temporal data mining [5] and outlier detection [6]. The mining of itemsets is a core step in these methods that often dominates the overall complexity of the problem. The number of frequent itemsets can be extremely large even for moderately sized datasets complicating a manual analysis or further automated processing steps.

Researchers have proposed many solutions to reduce the number of patterns reported depending on the context in which the patterns are used or the process in which the data was generated, for example closed itemsets [7] to avoid redundant association rules, constrained itemsets [8] to incorporate prior knowledge, condensed representations [9] to answer frequency queries with limited memory, margin-closed itemsets [5] for exploratory analysis, and surprising itemsets [10], [11] or top-k patterns [12] for itemset ranking.

Many of reduction techniques have a drawback of being fragile. For example, given a non-closed itemset X, adding a single transaction to dataset containing only X will make X closed. In this paper we introduce a novel theoretical framework that uses this drawback to its advantage. Given a property of an itemset (closedness or non-derivability, for example) we can measure *robustness* of this property. A property of X is robust if it holds for many datasets subsampled from the original data. We demonstrate that we can compute this measure analytically for several important classes of itemsets: closed [7], free [13], non-derivable [14], and totally shattered itemsets [15]. Computing robust itemsets under subsampling turns out to be practical for free, non-derivable, and totally shattered itemsets, for closed itemsets the test for robustness is prohibitively expensive.

A possible drawback of our approach is that it depends on a parameter α, the probability of including a transaction in a subsample. In addition to providing reasonable guidelines to choose α we introduce a technique making us independent of α. We show that there is a neighborhood near 1 in which the ranking of itemsets does not depend on α. We further demonstrate how we can compute this ranking without actually discovering the exact neighborhood or computing the measure for the itemsets. We give exact solutions for free and totally shattered itemsets and provide practical heuristics for closed and non-derivable itemsets.

In the remainder of this paper we describe related work and motivate our approach in Section II. Itemsets robust under subsampling and algorithms to find them are described in Section III. Section V demonstrates how the subsampling approach can reduce the number of reported itemsets significantly. The results are discussed in comparison with approximate itemsets in Sections VI.

II. Related work and motivation

The design goal of condensed representations [9] of frequent itemsets is to be able to answer all possible frequency queries (approximately). For example, non-derivable itemsets [14] exclude any itemset whose support that can be derived from others in the condensed representation using logical rules exactly or approximately.

This is useful to support further mining tasks such as generation of association rules where the frequencies of all subsets of a (closed) itemset are needed to determine the confidence of all possible rules. For other tasks knowing the frequency of all frequent itemsets is less useful because there is a large redundancy in the set of frequent itemsets. All frequent itemsets can be grouped into equivalence classes where all itemsets in a class are observed in the same set of transactions. The maximal element of each equivalence class is a closed itemset [7]. No more items can be added to this set without losing some transactions. The minimal elements of the equivalence class are free itemsets [13] or generators. No items can be taken out without adding transactions.

However, even the number of closed and free itemsets can still be very large for low minimum support thresholds. It can be further reduced by clustering itemsets representing similar sets of transactions [16], enforcing itemsets to have a minimum margin of difference in support [5], or ranking itemsets by significance [10], [11], [17], [18].

The above approaches have in common that the complete dataset is considered and no assumption on potential noise are made. In fault tolerant approaches the strict definition of support, requiring all items of an itemset to be present in a transaction is relaxed, see [19]–[22], assuming that items can present or absent at random in the transactions.

These approaches can reveal important structures in noisy data that might otherwise get lost in a huge amount of fragmented patterns. One needs to be aware though that they report approximate support values and possibly list itemsets that are not observed as such in the collection at all or with much smaller support. Also the design goal is not to reduce the number of reported patterns, only [23] considers closedness in combination with fault tolerance.

Unlike the approaches based on significance [10], [11], [17], [18], we do not assume a statistical null hypothesis. We also do not assume any noise model, such as flipping the values of a matrix independently. Instead our goal is to study robustness of a given property based on subsampling transactions.

III. Robust itemsets

A. Notation and definitions

In this section we review the preliminaries and introduce the notation used in the paper.

A *binary dataset* D is a multiset of transactions, binary vectors of length K. The ith element of a transaction represents by an *item* a_i, a Bernoulli random variable. We denote the collection of all items by $A = \{a_1, \ldots, a_K\}$.

An itemset X is a subset of A. Given a transaction t and an itemset X, we define t_X to be the binary vector obtained by keeping only the items occurring in X.

Given an itemset $X = (x_1, \ldots, x_N)$ and a binary vector v of length N, we define the *support*

$$sp(X = v; D) = |\{t \in D \mid t_X = v\}|$$

to be the number of transactions in D, where items in X obtain values of v. We often omit D from notation, when it is clear from the context. In addition, if v contains only 1s, we simply write $sp(X)$. Note that $sp(X)$ coincides with the traditional definition of a support for X. Discovering frequent itemsets, that is, itemsets whose support exceeds some given threshold is a well-studied problem.

Example 1 *Throughout the paper we will use the following toy dataset*

$$D = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

containing 5 *items,* a*,* b*,* c*,* d*, and* e*, and* 6 *transactions as a running example. As an example, for this dataset we have* $sp(ab) = 2$*,* $sp(ab = [1, 0]) = 1$*.*

We say that a function f mapping an itemset X to a real number $f(X)$ is *monotonically decreasing* if for each $Y \subseteq X$ we have $f(Y) \geq f(X)$.

Closed Itemsets: An itemset X is *closed*, if there is no $Y \supsetneq X$ such that $sp(X) = sp(Y)$, i.e., they are maximal among the itemsets having the same support. We define a predicate

$$\sigma_c(X; D) = \begin{cases} 1 & \text{if } X \text{ is closed in } D, \\ 0 & \text{otherwise} \quad . \end{cases}$$

Free Itemsets: An itemset X said to be *free* if there is no $Y \subsetneq X$ such that $sp(X) = sp(Y)$, i.e., free itemsets are minimal among the itemsets having the same support. We define a predicate

$$\sigma_f(X; D) = \begin{cases} 1 & \text{if } X \text{ is free in } D, \\ 0 & \text{otherwise} \quad . \end{cases}$$

A vital property of free itemsets is that they constitute a downward closed collection allowing efficient mining with an Apriori-style algorithm (see Theorem 1 in [24]).

Example 2 *Closed itemsets in our running example are* a*,* e*,* bde*, and* $abcde$*. On the other hand, itemsets* a*,* b*,* c*,* d*,* e*,* ab*,* ad*, and* ae *are free.*

Non-derivable Itemsets: An itemset X is said to be *derivable*, if we can derive its support from the supports of proper subsets of X, otherwise an itemset is called *non-derivable*. We define a predicate

$$\sigma_n(X; D) = \begin{cases} 1 & \text{if } X \text{ is non-derivable in } D, \\ 0 & \text{otherwise} \quad . \end{cases}$$

Proposition 3 *An itemset* X *is derivable if and only if there are two vectors* v *and* w *with* v *having odd number of* 0*s and* w *having even number of* 0*s and* $sp(X = v) = sp(X = w) = 0$*.*

Proof: To verify whether an itemset is derivable, we compute bounds for the frequency by using the inclusion-exclusion principle. An itemset is derivable if and only if the upper and lower bounds are equal. We can show that the upper bound is equal to $u = sp(X) + \min_v sp(X = v)$, where v has odd number of 0s. Similarly, the lower bound is equal to $l = sp(X) - \min_w sp(X = w)$, where w has even number of 0s (see [14]). Itemset is derivable if and only $0 = u - l = \min_v sp(X = v) + \min_w sp(X = w)$. □
Corollary 3.4 in [14] states that non-derivable itemsets are downward closed, hence we can mine them using an Apriori-style approach.

We say that an itemset X is *totally shattered* if $sp(X = v) > 0$ for all possible binary vectors v. In other words, every possible combination of values for X occur in D. Again, we define a predicate

$$\sigma_s(X; D) = \begin{cases} 1 & \text{if } X \text{ is totally shattered in } D, \\ 0 & \text{otherwise} \quad . \end{cases}$$

Totally shattered itemsets are related to VC-dimension [15],and we can show that a totally shattered itemset is always free and non-derivable (but not vice versa).

EXAMPLE 4 *Itemset ab in the running example is totally shattered. Itemset ac is non-derivable but not totally shattered because $sp(ac = [0, 1]) = 0$.*

It is easy to see from the definition that totally shattered itemsets constitute a downward collection, hence they are easy to mine using an Apriori-style approach.

B. Measuring robustness

In this section we propose a measure of robustness for itemsets with a predicate σ. Intuitively we consider an itemset robust if the predicate is true for many subsets of the database.

In order to define the measure formally, we first define a probability for a subset of D.

DEFINITION 5 Given a binary dataset D, and a real number α, $0 \leq \alpha \leq 1$, we define a random dataset D_α obtained from D by keeping each transaction with probability α, or otherwise discarding it.

Let S be a subset of D. The probability of $D_\alpha = S$ is equal to

$$p(D_\alpha = S) = \alpha^{|S|}(1 - \alpha)^{|D|-|S|} \quad . \tag{1}$$

DEFINITION 6 Given a binary dataset D, a real number α, and an itemset predicate σ, we define the robustness to be the probability that $\sigma(X; D_\alpha) = 1$, that is,

$$r(X; \sigma, D, \alpha) = p(\sigma(X; D_\alpha) = 1) = \sum_{\sigma(X;S)=1} p(D_\alpha = S) \quad .$$

For notational clarity, we will omit D and α when they are clear from the context.

EXAMPLE 7 *Consider itemset ab in our running example. Let $\alpha = 1/3$. Note that $sp(ab = [0, 0]) = sp(ab = [1, 0]) = 1$ and $sp(ab = [0, 1]) = sp(ab = [1, 1]) = 2$. In order for ab to still be totally shattered on a subset each of these supports needs to stay greater than zero. The probability of this event is equal to*

$$1/3 \times 1/3 \times (1 - 2/3 \times 2/3) \times (1 - 2/3 \times 2/3) = 25/729,$$

because for the first two cases we need to sample the single transaction upholding the property and for the other two cases we need to make sure we do not skip both of the two transactions we need to uphold the property.

Our main goal is to mine itemsets for which the robustness measure exceed some given threshold, that is, find all itemsets for which $r(X; \sigma, D, \alpha) \geq \rho$.

Let us first consider the effect of α. If we set $\alpha = 1$, then $r(X; \sigma, D, \alpha) = \sigma(X; D)$. Naturally, we expect that when we lower α then the robustness would decrease. This holds for predicates that satisfy a specific property.

DEFINITION 8 We say that a predicate σ is *monotonic w.r.t. deletion* if for each itemset X, each dataset D, and each transaction $t \in D$ it holds that if $\sigma(X; D) = 0$, then $\sigma(X; D - t) = 0$.

PROPOSITION 9 *Let σ be a predicate monotonic w.r.t. deletion. Then $r(X; \sigma, D, \alpha) \leq r(X; \sigma, D, \beta)$, for $\alpha \leq \beta$.*

Proof: We will prove the proposition by induction over $|D|$. Proposition holds trivially for $|D| = 0$. Assume that theorem holds $|D| = N$ and let D be a dataset with $|D| = N + 1$.

Fix $t \in D$ and define a new predicate $\sigma_t(X; S) = \sigma(X; S \cup \{t\})$, where S is a dataset. σ_t is monotonic w.r.t deletion. Otherwise, if there is a dataset S, a transaction $u \in S$ an itemset Y violating the monotonicity, then $S \cup \{t\}$, the same transaction u and the itemset Y will violate the monotonicity for σ.

Moreover, since σ is monotonic w.r.t deletion, it holds that $\sigma(X; S) \leq \sigma_t(X; S)$. This in turns implies that

$$r(X; \sigma, S, \alpha) \leq r(X; \sigma_t, S, \alpha)\,. \tag{2}$$

Let us write $D' = D - \{t\}$. Then we have,

$$\begin{aligned} r(X; \sigma, D, \alpha) &= (1 - \alpha) r(X; \sigma, D', \alpha) + \alpha r(X; \sigma_t, D', \alpha) \\ &\leq (1 - \beta) r(X; \sigma, D', \alpha) + \beta r(X; \sigma_t, D', \alpha) \\ &\leq (1 - \beta) r(X; \sigma, D', \beta) + \beta r(X; \sigma_t, D', \beta) \\ &= r(X; \sigma, D, \beta)\,, \end{aligned}$$

where the first inequality holds because of Equation 2 and the second inequality holds because of induction assumption. This proves the proposition. □

It turns out that all the predicates we considered in Section III-A are monotonic w.r.t. deletion.

PROPOSITION 10 *Predicates σ_c, σ_f, σ_n, and σ_s are monotonic w.r.t. deletion.*

Proof: An itemset is not totally shattered if there is a binary vector v such that $sp(X = v; D) = 0$. This immediately implies that $sp(X = v; D - \{t\}) = 0$. Thus σ_s is monotonic w.r.t. deletion. Similarly, Proposition 3 implies that σ_n is monotonic w.r.t. deletion.

An itemset X is not free, if there is $x \in X$ such that there is no transaction $u \in D$ for which $u_x = 0$ and $u_y = 1$ for all $y \in X - \{x\}$. If this holds in D, then it holds for $D - \{t\}$. This makes σ_f monotonic w.r.t. deletion. Similarly, an itemset X is not closed, if there is $x \notin X$ such that there is no transaction $u \in D$ for which $u_x = 0$ and $u_y = 1$ for all $y \in X$. If this holds in D, then it holds for $D - \{t\}$. This makes σ_c monotonic w.r.t. deletion. □

EXAMPLE 11 *The itemset bd is not closed because its superset bde is always observed when bd is observed. No matter which transaction we delete (one with or without bde) this will not change. Note, however, that bde can become non-closed if transactions 2 and 4 are deleted because then $abcde$ will have the same support of 2.*

In order to mine all significant patterns we need to show that the robustness measure is monotonically decreasing. This is indeed the case if the underlying predicate is monotonically decreasing.

PROPOSITION 12 *Let σ be a monotonically decreasing predicate. Then $r(X; \sigma, D, \alpha)$ is also monotonically decreasing.*

Proof: Let Y and X be itemsets such that $Y \subset X$. Then $r(X; \sigma, D, \alpha)$ is

$$\sum_{\sigma(X;S)=1} p(D_\alpha = S) \leq \sum_{\sigma(Y;S)=1} p(D_\alpha = S) = r(Y; \sigma, D, \alpha)\,,$$

which proves the proposition. □

C. Computing the measure

In this section we demonstrate how to compute the robustness measure for the predicates. Computing the measure directly from the definition is impractical since there are $2^{|D|}$ number of subsets of D. It turns out that computing free, non-derivable, and totally shattered itemsets has practical formulas while the robustness measure for closed itemsets has no practical formulation (see Table I).

To facilitate the analysis we introduce the following function: Given an itemset X and a set of binary vectors $V \subseteq \{0,1\}^{|X|}$ we define

$$o(X, V, \alpha) = \prod_{v \in V} 1 - (1 - \alpha)^{sp(X=v)} \quad .$$

Table I
COMPUTATIONAL COMPLEXITY OF ROBUSTNESS AND ORDERS. COMPUTING MEASURES IS EXPLAINED IN SECTION III-C. COMPUTING ORDERS IS EXPLAINED IN SECTION IV. K IS THE NUMBER OF ITEMS, $|\mathcal{C}|$ IS THE NUMBER OF FREQUENT CLOSED ITEMSETS.

predicate	measure	order	order estimate
free	$O(\lvert X\rvert)$	$O(\lvert X\rvert)$	–
totally shattered	$O(2^{\lvert X\rvert})$	$O(2^{\lvert X\rvert})$	–
closed	$O(2^{K-\lvert X\rvert})$	$O(2^{K-\lvert X\rvert})$	$O(\lvert\mathcal{C}^2\rvert)$
non-derivable	$O(2^{\lvert X\rvert})$	$O(2^{2^{\lvert X\rvert}})$	$O(4^{\lvert X\rvert})$

PROPOSITION 13 *Given an itemset X, let V be the set of $|X|$ vectors having $|X| - 1$ ones and one 0. The robustness of a free itemset is $r(X; \sigma_f, \alpha) = o(X, V, \alpha)$.*

Proof: Given an item $x \in X$, define a random variable $T_x = sp(X - \{x\}; D_\alpha) > sp(X; D_\alpha)$. X is still free in D_α if T_x is true for all $x \in X$. T_x is true if and only if D_α contains a transaction t with $t_x = 0$ and $t_y = 1$ for $y \in X - \{x\}$. There are $sp(X = v; D)$ such transactions, where $v \in V$ is the vector for which $v_x = 0$. $p(T_x)$ is the probability of not removing all these transactions, thus

$$p(T_x) = 1 - (1 - \alpha)^{sp(X=v;D)} \quad .$$

Since each of these transaction is missing only one $x \in X$, there is no common transactions between different events T_x, making them independent. Thus, we can conclude $r(X; \sigma_f, \alpha) = \prod_{x \in X} p(T_x) = o(X, V, \alpha)$. □

PROPOSITION 14 *Given an itemset X, let V be the set of all binary vectors of length $|X|$. The robustness of a totally shattered itemset is $r(X; \sigma_s, \alpha) = o(X, V, \alpha)$.*

Proof: Given a binary vector $v \in V$, define a random variable $T_v = sp(X = v; D_\alpha) > 0$. X is still totally shattered in D_α if T_v is true for all $v \in V$. $p(T_v)$ is the probability of not removing all these transactions, thus $p(T_v) = 1 - (1 - \alpha)^{sp(X=v;D)}$. Again, since no transaction can contribute to different T_v being true, the random variables are independent and we obtain $r(X; \sigma_s, \alpha) = \prod_{v \in V} p(T_v) = o(X, V, \alpha)$. □

Note that this formula in this proposition corresponds directly to Example 7.

We will now consider closed itemsets. Unlike with free/totally shattered itemsets, there is an exponential number of terms. The key problem is that closure depends on the items outside the itemset whereas other predicates consider only the items inside the itemset.

PROPOSITION 15 *The robustness of a closed itemset is*

$$r(X; \sigma_c, \alpha) = \sum_{Y \supseteq X} (-1)^{|Y|-|X|} (1 - \alpha)^{sp(X)-sp(Y)} \quad .$$

Proof: Given an item $y \notin X$, define a random variable $T_y = sp(X \cup \{y\}; D_\alpha) = sp(X; D_\alpha)$. X is still closed in

D_α if all T_y are false, thus $r(X;\sigma_c,\alpha)$ is equal to

$$1 - p(\bigvee_{y \notin X} T_y) = \sum_{Y \cap Z = \emptyset} (-1)^{|Z|} p(\bigwedge_{y \in Z} T_y),$$

where the equality follows from the inclusion-exclusion principle. Through this transformation we now need to determine the probability of all T_y being true. For this all $sp(X) - sp(Y \cup X)$ transactions containing X but not Z must have been excluded from D_α, hence

$$p(\bigwedge_{y \in Z} T_y) = (1-\alpha)^{sp(X)-sp(Z \cup X)} \quad .$$

Substituting this above and writing $Y = X \cup Z$ leads to the proposition. □

EXAMPLE 16 *In our running example, we have $sp(bde) = 4$. This itemset has 3 superitemsets having the supports $sp(abde) = sp(bcde) = sp(abcde) = 2$. Hence, the measure $r(bde;\sigma_c,\alpha)$ is equal to*

$$1 - (1-\alpha)^{4-2} - (1-\alpha)^{4-2} + (1-\alpha)^{4-2} = 1 - (1-\alpha)^2,$$

where itemsets bde, $abde$, $bcde$, and $abcde$ correspond to the terms on the left side in the given order.

PROPOSITION 17 *Given an itemset X, write V to be the set of binary vectors of length $|X|$ having odd number of ones. Similarly let W be the set of binary vectors of length $|X|$ having even number of ones. The robustness of a non-derivable itemset is*

$$r(X;\sigma_n,\alpha) = 1 - (1 - o(X,\alpha,V))(1 - o(X,\alpha,W)) \quad .$$

Proof: Let us define an event T_V to be the lack of $v \in V$ such that $sp(X = v) = 0$. Similarly, let T_W be an event corresponding to the lack of $w \in W$ such that $sp(X = w) = 0$. According to Proposition 3, an itemset X is derivable if T_V and T_W are both false.

Using the same argument as with Proposition 14, we see that $p(T_V) = o(X,\alpha,V)$. Similarly, $p(T_W) = o(X,\alpha,W)$. Since $V \cap W = \emptyset$, events A and B are independent. Hence, $r(X;\sigma_n,\alpha)$ is equal to

$$1 - p(\neg T_V \wedge \neg T_W) = 1 - (1 - p(T_V)(1 - p(T_W)) \quad .$$

This completes the proof. □

IV. ORDERING PATTERNS

The robustness measure depends on the parameter α. In this section we propose a parameter-free approach. The idea is to study how measure is behaving when α is *close* to 1. We can show that there is a (small) neighborhood close to 1, where the *ranking* of itemsets does not depend on α. We can compute a ranking that can use to select top-k itemsets by robustness without actually computing the measure or determining the neighborhood.

We will first introduce the general idea and then demonstrate how can we compute the ranking for free and totally shattered itemsets and how can we estimate the ranking for closed and non-derivable itemsets. For computational complexity see Table I.

A. Measuring robustness when α approaches 1

When $\alpha = 1$ then $D_\alpha = D$ with probability 1 and the measure is equivalent to the underlying predicate, providing only a crude ranking: itemsets that satisfy the predicate vs. itemsets that do not. If we make α slightly smaller the measure will decrease a little bit for each itemset. The amount of this change will vary from one itemset to another based on how likely removing only very few transactions will break the predicate for this itemset. We can use the magnitude of this change to obtain a more fine-grained ranking by robustness. The key result for this is that there is a small neighborhood below 1 in which the ranking of itemsets based on the measure does not depend on α.

PROPOSITION 18 *Given a predicate σ and a dataset D, there exists a number $\beta < 1$ such that*

$$r(X;\sigma,D,\alpha) \leq r(Y;\sigma,D,\alpha) \text{ if and only if}$$
$$r(X;\sigma,D,\alpha') \leq r(Y;\sigma,D,\alpha'),$$

for any itemset X and Y and $\beta \leq \alpha \leq 1$, $\beta \leq \alpha' \leq 1$.

Proof: Fix X and Y and consider

$$f(\alpha) = r(X;\sigma,D,\alpha) - r(Y;\sigma,D,\alpha) \quad .$$

Since the measure is a finite sum of probabilities that are, according to Eq. 1, polynomials of α, the function f is a polynomial. This implies that f can have only finite number of 0s. Consequently there is a neighborhood $N = [\beta, 1]$ such that either $f(\alpha) \geq 0$ for any $\alpha \in N$, or $f(\alpha) \leq 0$ for $\alpha \in N$. Since there are only finite number of itemsets, we can take the maximum of all βs to prove the theorem. □

Proposition 18 allows us to define an order for itemsets based on the measure for $\alpha \approx 1$.

DEFINITION 19 Given a predicate σ, and a dataset D, we say that $X \preceq_\sigma Y$, where X and Y are itemsets, if there is $\beta < 1$ such that $r(X;\sigma,D,\alpha) \leq r(Y;\sigma,D,\alpha)$ for any α such that $\beta \leq \alpha \leq 1$. Moreover, if $r(X;\sigma,D,\alpha) < r(Y;\sigma,D,\alpha)$ for some $\alpha \geq \beta$, then we write $X \prec_\sigma Y$.

Note that Proposition 18 implies that for any X and Y, either $X \preceq_\sigma Y$ or $Y \preceq_\sigma X$. That is, we can use this relation to order itemsets.

B. Free and totally shattered itemsets

In this section we will demonstrate that we can compute the order for free and totally shattered itemsets without finding an appropriate α. We will do this by analyzing the coefficients of the measure viewed as a polynomial of α.

The key step is the following lemma that can be proven by elementary real analysis.

LEMMA 20 *Let $f(x) = \sum_{i=0}^{N} a_i x^i$ be a non-zero polynomial. Let k be the first index such that $a_k \neq 0$ If $a_k > 0$, then there is a $\beta > 0$ such that $0 \leq x \leq \beta$ implies $f(x) \geq 0$. Similarly, if $a_k < 0$, then there is a $\beta > 0$ such that $0 \leq x \leq \beta$ implies $f(x) \leq 0$.*

We cannot use Lemma 20 directly with Proposition 13 and Proposition 14 because both polynomials contain an exponential number of terms. However, the polynomials are regular enough so that we can compute the order without expanding the polynomials. In order to that we need the following definition for ordering sequences.

DEFINITION 21 Given two non-decreasing sequences $s = s_1, \ldots, s_K$ and $t = t_1, \ldots, t_N$, we write $s \prec t$ if either there is $s_n < t_n$ and $s_i = t_i$ for all $i < n$ or s is a proper prefix sequence of t, that is, $s_i = t_i$ for $i \leq K < N$. We write $s \preceq t$, if $s = t$ or $s \prec t$.

The following proposition will allow us to order itemsets without expanding the polynomials in Propositions 13–14.

PROPOSITION 22 *Assume two polynomials*

$$f(\alpha) = \prod_{i=1}^{K}(1-(1-\alpha)^{s_i}) \text{ and } g(\alpha) = \prod_{i=1}^{N}(1-(1-\alpha)^{t_i}),$$

where $s = s_1, \ldots, s_K$ and $t = t_1, \ldots, t_N$ are non-decreasing sequences of integers, $s_i, t_i \geq 0$. If $t \preceq s$, then there is a $\beta < 1$ such that $\beta \leq \alpha \leq 1$ implies $f(\alpha) \geq g(\alpha)$.

Proof: The case $s = t$ is trivial. Hence we assume that $s \neq t$. If $s_1 = 0$ or $t_1 = 0$, then $f(\alpha) = 0$ or $g(\alpha) = 0$, and the result follows, hence we will assume that $s_i, t_i > 0$.

Let $\{a_i\}$ and $\{b_i\}$ be coefficients such that

$$f(1-x) = \sum_i a_i x^i \text{ and } g(1-x) = \sum_i b_i x^i \quad .$$

Let $\mathcal{I}_n$ be the collection of all subsequences of s that sum to n. Similarly, let $\mathcal{J}_n$ be the collection of all subsequences of t that sum to n. Then, it follows that

$$a_n = \sum_{I \in \mathcal{I}_n} (-1)^{|I|} \text{ and } b_n = \sum_{J \in \mathcal{J}_n} (-1)^{|J|} \quad .$$

Assume that $t \prec s$. If s is a prefix sequence of t, then

$$g(\alpha) = f(\alpha) \prod_{i=K+1}^{N} (1-(1-\alpha)^{t_i}) \leq f(\alpha),$$

which proves the proposition. Let n be as given in Definition 21. For every $i < s_n$, the subsequences in $\mathcal{I}_i$ and $\mathcal{J}_i$ contain entries from s and t with indices smaller than n. Since s and t are identical up to n, then it follows that $\mathcal{I}_i = \mathcal{J}_i$ and consequently $a_i = b_i$. Let $I \in \mathcal{I}_{s_n}$. Assume that $|I| > 1$. Since, we assume that $s_i > 0$, I is a subsequence of $s_1, \ldots, s_{n-1}$. This means that we will find the same subsequence in $\mathcal{J}_n$. Let A be the number of singleton sets in $\mathcal{I}_{s_n}$ and let B be the number of singleton sets in $\mathcal{J}_{s_n}$. These singleton sets correspond to the entries in s and t having the same value as s_n. By definition, $B > A$. We have now $a_n - b_n = B - A > 0$. Lemma 20 now implies that $f(1-x) \geq g(1-x)$, when x is close to 0. Write $\alpha = 1 - x$ to complete the proof. □

The polynomials in Propositions 13–14 have the form used in Proposition 22. Consequently, we can use the proposition to order itemsets. In order to do that we need the following definitions.

DEFINITION 23 Given a dataset D and an itemset X, we define a *free margin vector* $mv(X; D, \sigma_f)$ to be the sequence of $|X|$ integers $sp(X = v; D)$, where v is a binary vector having $|X| - 1$ ones, *ordered* in the increasing order.

Similarly, we define a *totally shattered margin vector* $mv(X; D, \sigma_s)$ to be a sequence of $2^{|X|}$ integers $sp(X = v; D)$ *ordered* in the increasing order.

COROLLARY 24 *Given itemsets X and Y and a dataset D, $X \preceq_{\sigma_f} Y$ if and only if $mv(X; D, \sigma_f) \preceq mv(Y; D, \sigma_f)$.*

COROLLARY 25 *Given itemsets X and Y and a dataset D, $X \preceq_{\sigma_s} Y$ if and only if $mv(X; D, \sigma_f) \preceq mv(Y; D, \sigma_s)$.*

EXAMPLE 26 *In our running example, $sp(ab = [1, 0]) = 1$ and $sp(ab = [0, 1]) = 2$, hence the free margin vector is equal to $mv(ab; \sigma_f) = [1, 2]$. Similarly, we have $sp(ae = [1, 0]) = 1$ and $sp(ae = [0, 1]) = 3$, hence the free margin vector is equal to $mv(ae; \sigma_f) = [1, 3]$. Hence, we conclude that $ab \prec_{\sigma_f} ae$.*

C. Closed itemsets

In this section we will introduce a technique for estimating the ranking for closed itemsets. As the measure for closed itemsets has a different form than for free or totally shattered itemsets we are forced to seek for alternative approaches.

Let us consider Proposition 15. Let a_k be the coefficient for the kth term of the polynomial for $r(X; \sigma_c, \alpha)$ given in Proposition 15. If we can compute these numbers efficiently, we can use Lemma 20 to find the ranking.

We will do this by first expressing a_k using closed itemsets. In order to do that let $cl(X)$ be the closure of an itemset X. Let us define

$$e(Y, X) = \sum_{\substack{Z \supseteq X, \\ cl(Z) = Y}} (-1)^{|Z|+|X|}$$

to be the alternating sum over all itemsets containing X and having Y as their closure. Since all the itemsets having the same closure will have the same support we can write the coefficients a_k using $e(Y, X)$,

$$a_k = \sum_{\substack{Y \supseteq X, \\ sp(X) - sp(Y) = k}} (-1)^{|Y|+|X|} = \sum_{\substack{Y \supseteq X, Y = cl(Y) \\ sp(X) - sp(Y) = k}} e(Y, X) \quad . \qquad (3)$$

To compute $e(Y, X)$, first note that $e(X, X) = 1$. If $Y \neq X$, then using the following identity

$$\sum_{\substack{Y \supseteq Y' \supseteq X \\ Y' = cl(Y')}} e(Y', X) = \sum_{Z \supseteq X} (-1)^{|Z|+|X|} = 0$$

we arrive to

$$e(Y, X) = -\sum_{\substack{Y \supsetneq Y' \supseteq X \\ Y' = cl(Y')}} e(Y', X) \quad . \tag{4}$$

Thus, we can compute $e(Y, X)$ from $e(Y', X)$, where Y' is a closed subset of Y. This is convenient, because when computing $e(Y, X)$, say for a_k, we have already computed all the subsets of Y for previous coefficients.

EXAMPLE 27 *Consider itemset e in our running example. There are two closed supersets of e, namely bde and $abcde$, having the supports 4 and 2, respectively. Using the update equations, we see that $e(e, e) = 1$, $e(bde, e) = -1$, and $e(abcde, e) = 0$. As $sp(e) = 5$, we see that the non-zero coefficients a_i are $a_0 = 1$ and $a_1 = -1$.*

The problem with this approach is that we can still have an exponential number of closed itemsets. Hence, we chose to estimate the ranking by only using *frequent* closed itemsets and estimate the remaining itemsets to have a support of 0.

This estimation is achieved by removing all closed non-frequent itemsets from the sums of Eqs. 3 and 4 and adding an itemset containing all the items and having the support 0. The code for this estimation is given in Algorithm 1.

Algorithm 1: Algorithm for estimating coefficients of the polynomial given in Proposition 15.

input : X an itemset, $\mathcal{C}$, frequent closed itemsets
output: $\{a_k\}$, coefficients of the polynomial

```
if A ∉ C then add A to C with sp(A) = 0;
C ← {Y ∈ C | X ⊆ Y} ;
L ← sets in C ordered by the subset relation;
e(X, X) ← 1;
for Y ∈ L do
    e(Y, X) ← −Σ_{Z∈C, Z⊊Y} e(Z, X);
    k ← sp(X) − sp(Y);
    a_k ← a_k + e(Y, X);
```

Algorithm 1 takes $O(|\mathcal{C}|^2)$ time. In practice, this is much faster because an average itemset does not have that many supersets.

Now that we have a way of estimating a_k from frequent closed itemsets, we can, given two itemsets X and Y, search the smallest k for which the coefficients differ in order to apply Lemma 20. Note that if the index of the differing coefficient, say k, is such that $sp(X) - k$ is larger or equal to the support threshold, then a_k is correctly computed by our estimation, and our approximation yields a correct ranking.

D. Non-derivable itemsets

In this section we will discuss how to estimate the ranking non-derivable itemsets. The ranking for non-derivable is particularly difficult because we cannot use Proposition 22 to avoid expanding the polynomial given in Proposition 17. We cannot expand the polynomial since it has $O(2^{2^{|X|}})$ terms. Moreover, we cannot use the estimation trick done with closed itemsets because the problem is the exponential number of combinations of subsets of $|X|$. Hence, we resort to a simple heuristic.

First note that we can rewrite the measure as

$$o(X, \alpha, V) + o(X, \alpha, W) - o(X, \alpha, V)o(X, \alpha, W),$$

where V and W are as defined in Proposition 17. This formulation implies that any term of form $(1-\alpha)^{\sum_v sp(X=v)}$, where v sums either over a subset of V or a subset of W, is canceled out. On the other hand, the terms having the form $(1-\alpha)^{sp(X=v)+sp(X=w)}$, where $v \in V$ and $w \in W$, will be among the smallest ones. Hence, we propose the following margin vector to use as a heuristic.

DEFINITION 28 Given a dataset D and an itemset X, we define a *non-derivable margin vector* $mv(X; D, \sigma_n)$ to be a sequence of $4^{|X|-1}$ integers $sp(X = v; D) + sp(X = w; D)$, where v is a binary vector having odd number of ones and w is a binary vector having even number of ones, *ordered* in the increasing order.

We will rank itemsets by comparing their margin vectors. We should stress that this is heuristic since, unlike with free and totally shattered itemsets, we have no guarantee that terms containing more than two supports will cancel out and not disrupt the ranking. Nevertheless, this ranking makes sense in the light of Proposition 3: an itemset is derivable if and only if one the entries in the margin vector is 0. If the entries in the margin vector are large, then the itemset is 'far away' of being derivable.

EXAMPLE 29 *Consider itemset ac in our running example. We have $sp(ac = [0, 0]) = 3$, $sp(ac) = 2$, $sp(ac = [1, 0]) = 1$, and $sp(ac = [0, 1]) = 0$. Thus the margin vector is equal to $[0 + 2, 0 + 3, 1 + 2, 1 + 3] = [2, 3, 3, 4]$.*

V. EXPERIMENTS

In this section we present our experiments.[1]

A. Datasets

We used datasets from three repositories. The 10 FIMI [25] datasets include large transaction datasets derived from traffic data, census data, and retail data. Two datasets are synthetically generated to simulate market basket data. The datasets from the UCI Machine Learning Repository [26] represent classification problems from a wide

[1]The implementation of our algorithms is given at http://adrem.ua.ac.be/implementations/

variety of domains. We used the itemset representations of 29 datasets from the LUCS repository [27]. Finally we used 18 text datasets shipped with the Cluto clustering toolkit [28] but converted to itemsets using a binary representation of words in documents discarding the term frequencies.

B. Reducing the number of patterns

The goal of the first experiment is to show that this new constraint for itemsets can significantly reduce the number of itemsets reported in the results by removing itemsets that are spurious in the sense that they are unlikely to be observed on many subsamples. Throughout this section we will use σ to indicate the threshold for the support.

A first question is how the parameters should be chosen. The smaller we set α, the stricter the filtering will be. α should not be very close 1, because otherwise brittle itemsets that could lose their predicate by removing only a few transactions still have a high likelihood of being found. This implies that robustness values are packed close to 1 when α is large, and this might lead to problems due to floating point arithmetics. So a small α is important to emphasize the quantitative difference between itemsets of various robustness, however, too small α will skew the distribution towards 0 too much, which can lead to computational issues. The larger the minimum robustness, the stricter the filtering will be. The robustness threshold is more application dependent but it should not be close to zero, otherwise no reduction will be observed.

We did a parameter study for the itemset version of the *Zoo* dataset that describes 101 animals with 42 boolean attributes. The number of itemsets reported is shown in Figure 1. One can see how smaller α and larger robustness thresholds reduce the numbers of free itemsets by almost 2 orders or magnitude. The transition is very smooth, showing that the parameters can be chosen without unexpected effects. The results for non-derivable and totally shattered itemsets and other datasets were very similar.

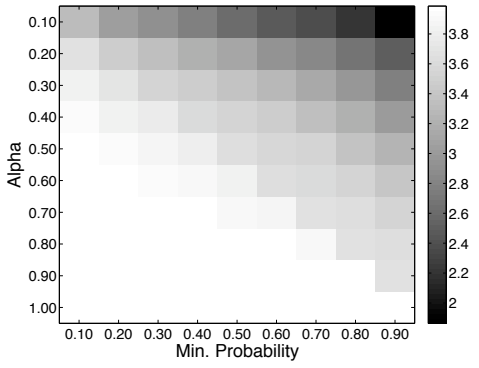


Figure 1. Log of the number of free itemsets on *Zoo* ($\sigma = 0.01$) dataset using thresholds for the subsample size (α) and the minimum robustness.

Based on this study we chose $\alpha = 0.5$ and a minimum robustness of 0.1 to drill deeper into the robustness of the reported itemsets. Excluding singleton itemsets we plotted histograms with the empirical distribution of robustness values associated with the reported itemsets. Figure 2(a) shows that for the *Zoo* dataset there are many free itemsets with very different robustness showing a rich structure that can be exploited to rank and reduce the number of itemsets. Similar results were observed for many of the UCI datasets. Figure 2(a) shows a representative example for the text datasets. While the distribution is much more skewed, a large robustness threshold would also reduce the number of itemsets by about 50%. Finally, Figure 2(c) shows an example for a large transactional dataset with 88k transactions. Using $\alpha = 0.5$ generated a distribution where all values were close to one so we needed to set $\alpha = 0.01$ to better show the quantitative difference of the itemsets.

This shows that the more transactions a dataset contains, the more skewed the distribution for a fixed α will be. For experiments with all datasets we set $\alpha = \max(0.1, \min(0.5, 1000/|D|))$, that is we use samples of 1000 transactions but for small datasets we use 50% and for very large datasets we use 10%. Using this parameter we computed all robust itemsets with a robustness ≥ 0.1 and computed the median robustness of the reported itemsets to summarize the distributions. Figure 3 plots the median robustness against the order of magnitude that the itemsets can be reduced when using a robustness threshold of 0.9. For many datasets a significant reduction is observed. For some datasets with a median close to or equal to 1 the reduction is small, indicating that most itemsets found are quite robust in this data.

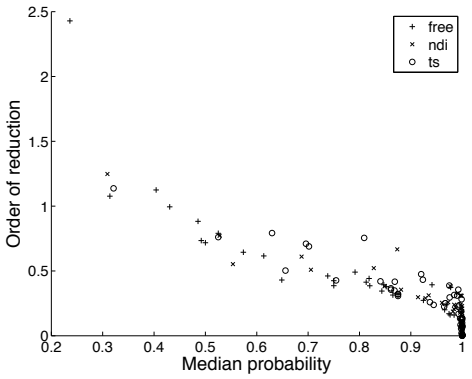
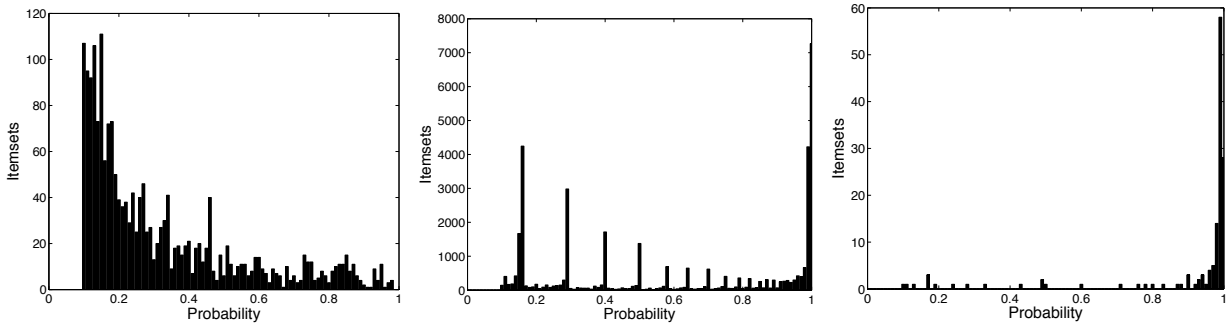


Figure 3. Median of the itemset robustness vs. order of magnitude in numerosity reduction ($\log_{10}$ scale) using robustness threshold 0.9.

C. Ranking without α

Our next experiment was to compare parameter-free rankings described in Section IV against the rankings based on robustness. We expect that rankings are similar for high α values and increase when we lower α. For comparison we used Kendall's τ distance, that is the number of discordant pairs, normalized such that the distance ranges between



(a) *Zoo* ($\alpha = 0.5, \sigma = 0.01$) (b) *LA12* ($\alpha = 0.5, \sigma = 0.05$) (c) *Retail* ($\alpha = 0.01, \sigma = 0.0080$)

Figure 2. Distribution of robustness for free itemsets and minimum robustness 0.1.

0 and 1. Values close to 0 means that rankings are in agreement. Since the values are typically very small we represent the values on a log-scale. A typical example is given in Table II for *Mushroom* and *Zoo* datasets.

In general, the distance values are small, suggesting that the rankings are similar. The values increase as we lower α which is expected since the parameter-free approach is based on large α values. Rankings for non-derivable itemsets tend to be higher, which is also expected, since the ranking for non-derivable itemsets is a heuristic.

Table II
$\log_{10}$(KENDALL'S TAU DISTANCE) FOR *Mushroom* AND *Zoo* DATASETS.

	Mushroom ($\sigma = 0.05$)			*Zoo* ($\sigma = 0.01$)		
α	free	ts	ndi	free	ts	ndi
0.1	-2.55	-2.13	-2.21	-0.81	-1.14	-1.07
0.2	-3.17	-2.65	-2.54	-0.99	-1.06	-1.22
0.3	-3.91	-3.08	-2.8	-1.22	-1.27	-1.33
0.4	-4.98	-3.46	-2.92	-1.54	-1.56	-1.36
0.5	-6.72	-3.79	-3.07	-1.95	-1.88	-1.44
0.6	-7.58	-4.06	-3.11	-2.49	-2.24	-1.5
0.7	$-\infty$	-4.98	-3.63	-3.76	-2.7	-1.55
0.8	$-\infty$	-4.99	-4.11	$-\infty$	$-\infty$	-1.6
0.9	$-\infty$	$-\infty$	-5.07	$-\infty$	$-\infty$	-1.64

D. Top-k closed itemsets

Closed itemsets are often used for tasks requiring interpretation of the itemsets, because a maximum elements of an equivalence class they offer the most detailed description. We studied the highest ranked closed itemsets for text datasets that are easily understood without domain knowledge. As an illustrative example, we used the *re0* news dataset from which we mine 2493 closed itemsets with minimum support $\sigma = 0.05$. We ordered these itemsets using the estimation technique given in Section IV-C and list the top 45 itemsets in Table III. The ranking is different from one using support, less frequent (but more robust) itemsets are commonly ranked higher that frequent itemset. For example, 'bank pct rate' occurs before much more frequent itemset 'bank pct'.

Table III
TOP-45 CLOSED ITEMSETS FROM *re0* ($\sigma = 0.05$) DATASET.

1.	pct	792	16.	week	310	31.	canada	117
2.	bank	702	17.	pct earlier	127	32.	pct month	261
3.	trade	485	18.	japan	318	33.	econom	295
4.	billion	552	19.	trade current	126	34.	billion dlr mln	116
5.	market	554	20.	dlr	472	35.	told bank	116
6.	billion dlr	346	21.	bank pct rate	287	36.	told nation	116
7.	offici	342	22.	dollar	336	37.	pct japan	115
8.	mln	420	23.	statem	122	38.	pct adjust	115
9.	nation	323	24.	committe	121	39.	billion current	115
10.	rate	566	25.	nation month	121	40.	european	114
11.	bank market	369	26.	ministri	120	41.	month japan	114
12.	foreign	331	27.	pct rise	269	42.	bank ad market	114
13.	pct figur	132	28.	bank pct	407	43.	action	114
14.	pct rate	418	29.	pct rate feb	119	44.	trade world	114
15.	month	391	30.	lead	118	45.	nation japan	114

VI. DISCUSSION

The experiments have shown that the number of itemsets can be largely reduced on many datasets when requiring a certain robustness. The fact that the results vary by dataset are another indication of the well known fact that itemset data with different structures (dense vs. sparse, many items vs. many transactions) behave very differently in mining tasks.

We believe that robust itemsets can be beneficial for post-processing techniques such as [29] or [30] that use itemsets as their input and remove redundancy in the pattern set. Robust itemsets can be used as an alternative input reducing their runtime without sacrificing performance. Also, robust itemsets could be used instead of closed-itemsets as seeds to the AC-Close algorithm for approximate itemset mining [23] improving its efficiency that was criticized in [19].

The ranking of itemsets by robustness presents a new interestingness measure that can be used to choose the top-k itemsets for interpretation or other data mining tasks. The intuition of robustness should be easy to understand for analysts but which ranking is better for specific data mining tasks remains to be studied.

VII. Summary

We have shown how robustness under subsampling for common classes of itemsets can be computed efficiently without actually sampling the data. The experimental results show that the number of reported itemsets can be largely reduced on many datasets, in other words spurious itemsets that would not have been found in many subsets of the data are removed. The approach can further be used to rank itemsets for top-k mining by robustness. Future work will investigate the effect of using robust itemsets on data mining tasks such as clustering, classification, and rule generation using itemsets.

Acknowledgments

Nikolaj Tatti is supported by Post-Doctoral Fellowships of the Research Foundation—Flanders (FWO).

References

[1] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases," in *SIGMOD*, 1993, pp. 207–216.

[2] J. Hipp, U. Güntzer, and G. Nakhaeizadeh, "Algorithms for association rule mining - a general survey and comparison." *SIGKDD Explorations*, vol. 2, no. 1, pp. 58–64, 2000.

[3] K. Wang, C. Xu, and B. Liu, "Clustering transactions using large items," in *CIKM*, 1999, pp. 483–490.

[4] H. Cheng, X. Yan, J. Han, and C. Hsu, "Discriminative frequent pattern analysis for effective classification," in *ICDE*, 2007, pp. 716–725.

[5] F. Moerchen, M. Thies, and A. Ultsch, "Efficient mining of all margin-closed itemsets with applications in temporal knowledge discovery and classification by compression," *KAIS*, 2010.

[6] K. Smets and J. Vreeken, "The odd one out: Identifying and characterising anomalies," in *SDM*, 2011.

[7] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering frequent closed itemsets for association rules," in *ICDT*, 1999, pp. 398–416.

[8] J. Pei, J. Han, and L. V. S. Lakshmanan, "Mining frequent itemsets with convertible constraints," in *ICDE*, 2001, pp. 433–442.

[9] T. Calders, C. Rigotti, and J.-F. Boulicaut, "A survey on condensed representations for frequent sets," in *Constraint-Based Mining and Inductive Databases*, 2006, pp. 64–80.

[10] S. Brin, R. Motwani, and C. Silverstein, "Beyond market baskets: Generalizing association rules to correlations," in *SIGMOD*, 1997, pp. 265–276.

[11] N. Tatti, "Maximum entropy based significance of itemsets," *KAIS*, vol. 17, no. 1, pp. 57–77, 2008.

[12] F. Geerts, B. Goethals, and T. Mielikäinen, "Tiling databases," in *Proc. Discovery Science*, 2004, pp. 278–289.

[13] J.-F. Boulicaut, A. Bykowski, and C. Rigotti, "Free-sets: A condensed representation of boolean data for the approximation of frequency queries," *DMKD*, vol. 7, no. 1, pp. 5–22, 2003.

[14] T. Calders and B. Goethals, "Non-derivable itemset mining," *DMKD*, vol. 14, no. 1, pp. 171–206, 2007.

[15] T. Mielikäinen, "Transaction databases, frequent itemsets, and their condensed representations," in *KDID*, 2005, pp. 139–164.

[16] D. Xin, J. Han, X. Yan, and H. Cheng, "Mining compressed frequent-pattern sets," in *VLDB*, 2005, pp. 709–720.

[17] A. Gallo, T. De Bie, and N. Cristianini, "Mini: Mining informative non-redundant itemsets," in *ECMLPKDD*, 2007, pp. 438–445.

[18] G. I. Webb, "Discovering significant patterns," *Mach. Learn.*, vol. 68, no. 1, pp. 1–33, 2007.

[19] R. Gupta, G. Fang, B. Field, M. Steinbach, and V. Kumar, "Quantitative evaluation of approximate frequent pattern mining algorithms," in *KDD*, 2008, pp. 301–309.

[20] T. Calders, B. Goethals, and M. Mampaey, "Mining itemsets in the presence of missing values," in *SAC*, 2007, pp. 404–408.

[21] T. Uno and H. Arimura, "An efficient polynomial delay algorithm for pseudo frequent itemset mining," in *Discovery Science*. Springer, 2007, pp. 219–230.

[22] C. Luccese, S. Orlando, and G. Perego, R.Casas-Garriga, "Mining top-k patterns from binary datasets in presence of noise," in *ICDM*, 2010.

[23] H. Cheng, P. S. Yu, and J. Han, "AC-Close: Efficiently mining approximate closed itemsets by core pattern recovery," in *ICDM*. IEEE, 2006, pp. 839–844.

[24] J.-F. Boulicaut, A. Bykowski, and C. Rigotti, "Approximation of frequency queries by means of free-sets," in *PKDD*, 2000, pp. 75–85.

[25] B. Goethals and M. Zaki, "FIMI '03, frequent itemset mining implementations," in *ICDM 2003 Workshop, FIMI*, 2003.

[26] A. Asuncion and D. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[27] F. Coenen, "The LUCS-KDD discretised/normalised ARM and CARM data library," 2003. [Online]. Available: http://www.csc.liv.ac.uk/~frans/KDD/Software/LUCS_KDD_DN/

[28] Y. Zhao and G. Karypis, "Evaluation of hierarchical clustering algorithms for document datasets," in *CIKM*, 2002, pp. 515–524.

[29] B. Bringmann and A. Zimmermann, "One in a million: picking the right patterns," *KAIS*, vol. 18, no. 1, pp. 61–81, 2009.

[30] J. Vreeken, M. van Leeuwen, and A. Siebes, "Krimp: mining itemsets that compress," *DMKD*, vol. 23, no. 1, pp. 169–214, 2011.