J Intell Inf Syst (2010) 34:1–19
DOI 10.1007/s10844-008-0077-0

A new classification of datasets for frequent itemsets

Frédéric Flouvat · Fabien De Marchi · Jean-Marc Petit

Received: 30 August 2007 / Revised: 22 December 2008 / Accepted: 22 December 2008 /
Published online: 21 January 2009

Abstract The discovery of frequent patterns is a famous problem in data mining. While plenty of algorithms have been proposed during the last decade, only a few contributions have tried to understand the influence of datasets on the algorithms behavior. Being able to explain why certain algorithms are likely to perform very well or very poorly on some datasets is still an open question. In this setting, we describe a thorough experimental study of datasets with respect to frequent itemsets. We study the *distribution* of frequent itemsets with respect to itemsets size together with the distribution of three concise representations: frequent closed, frequent free and frequent essential itemsets. For each of them, we also study the distribution of their positive and negative borders whenever possible. The main outcome of these experiments is a new classification of datasets invariant w.r.t. minsup variations and robust to explain efficiency of several implementations.

Keywords Pattern mining · Classification of datasets · Experimental study

1 Introduction

The discovery of frequent patterns is a famous problem in data mining, introduced in Agrawal et al. (1993) as a first step for mining association rules. While plenty

F. Flouvat (✉)
University of New Caledonia, PPME, BP R4, 98851, Nouméa, New Caledonia
e-mail: frederic.flouvat@univ-nc.nc

F. De Marchi
Université de Lyon, Université Lyon 1, LIRIS, UMR5205 CNRS, 69621, Lyon, France
e-mail: fabien.demarchi@liris.cnrs.fr

J.-M. Petit
Université de Lyon, INSA-Lyon, LIRIS, UMR5205 CNRS, 69621, Lyon, France
e-mail: jean-marc.petit@insa-lyon.fr

of algorithms have been proposed during the last decade among which we quote Agrawal and Srikant (1994), Burdick et al. (2001), Gouda and Zaki (2001), Han et al. (2000), Uno et al. (2004), only a few contributions have tried to understand the influence of dataset characteristics on the algorithms behavior, such as Bayardo and Zaki (2003), Gouda and Zaki (2001), Palmerini et al. (2004). These studies focus on the number of transactions, average length of transactions, or frequent itemsets distribution, i.e. statistics from frequent itemsets and maximal frequent itemsets. Nevertheless algorithms could have quite different behaviors for (apparently) similar datasets. Benchmarks comparing algorithms performances have been done on real and synthetic datasets in Bayardo et al. (2004), Bayardo and Zaki (2003). Algorithm implementations and datasets are freely available from FIMI website (see Goethals 2003) for mining frequent, frequent closed or frequent maximal itemsets. However, being able to explain why certain algorithms are likely to perform very well or very poorly on some datasets is still an open question.

More generally, studying datasets can provide useful hints for devising adaptive algorithms, i.e. algorithms which adapt themselves to data characteristics in order to increase their time or memory efficiency, such as Flouvat et al. (2004), Orlando et al. (2003). Adaptive behavior of algorithms is not new in the setting of frequent itemsets mining, for example Borgelt (2003), Burdick et al. (2001) use heuristics to decide when tries-like data structure, representing datasets and/or itemset collections, have to be rebuilt. The promising results obtained by these algorithms show the interest of applying specific strategies according to dataset features.

Another key point is that some problems have specific invariant characteristics, whatever the studied datasets. Their impact on algorithms could give useful information about the difficulty to solve these problems while giving hints on the more appropriate strategies to cope with these difficulties.

Contribution In this setting, we describe a thorough experimental study of datasets with respect to frequent itemsets. We study the *distribution* of frequent itemsets with respect to itemsets size together with the distribution of three concise representations: frequent closed, frequent free and frequent essential itemsets. For each of them, we also study the distribution of their positive and negative borders whenever possible. The positive (resp. negative) border corresponds to the maximal frequent (resp. minimal unfrequent) itemsets w.r.t. set inclusion. From this analysis, we exhibit a new classification of datasets and some invariants allowing to better predict the behavior of well known algorithms.

To the best of our knowledge, this work is the first one to address the understanding of datasets for frequent itemsets and other concise representations by using their negative borders.

Paper organization Related works are discussed in Section 2. In Section 3, we introduce some preliminaries on frequent itemsets and usual representations of frequent itemsets. Experimental study of datasets is given in Section 4, including experimental protocol, results and analysis. The Section 5 presents the main result of this work: a new classification of datasets for frequent itemsets related to algorithms performances. The Section 6 shows how this study can be applied to other data mining problems. Finally, we conclude and give some perspectives for this work.

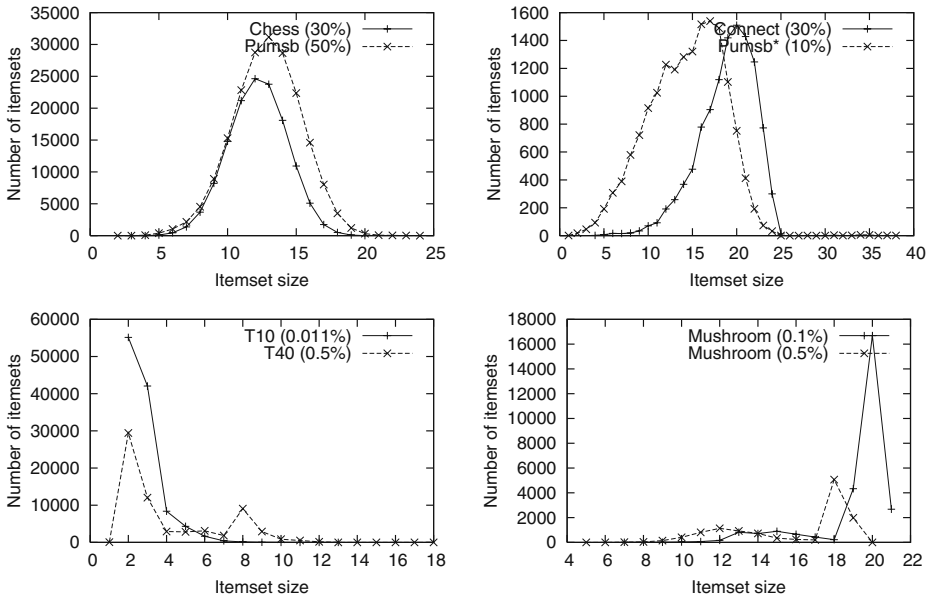

Fig. 1 Distribution of the maximal frequent itemsets from Gouda and Zaki (2001)

2 Related works

Classical characteristics of datasets were studied in Gouda and Zaki (2001), and more particularly a density criteria. Up to our knowledge no formal definition of density does exist. According to Gouda and Zaki (2001), a dataset is *dense* when it produces many long frequent itemsets even for high values of minimum support threshold. The authors studied seven datasets, each of them capturing a fairly large range of typical uses. The result of these experimentations is a classification of datasets in four categories according to the density. The density is estimated by using the characteristics of maximal frequent itemsets, and more precisely their distribution. The Fig. 1 represents the distributions of the maximal frequent itemsets for the datasets and minimum support threshold studied in Gouda and Zaki (2001).

The Table 1 shows the corresponding classification proposed.

Table 1 Datasets classification based on the maximal frequent itemsets from Gouda and Zaki (2001)

Type	Type 1	Type 2	Type 3	Type 4
Distribution of maximal frequent itemsets	Symetric	Gradual increase with a sharp drop	Exponentially decaying distribution	Explosion of long maximal itemsets
Size of maximal frequent itemsets	Relatively short	Long	Very short	Long
Examples of datasets	*Chess*(30%) *Pumsb*(50%)	*Connect*(30%) *Pumsb**(10%)	$T10I4D100K$ (0.011%) $T40I10D100K$ (0.5%)	*Mushroom* (0.1%) Mushroom (0.5%)

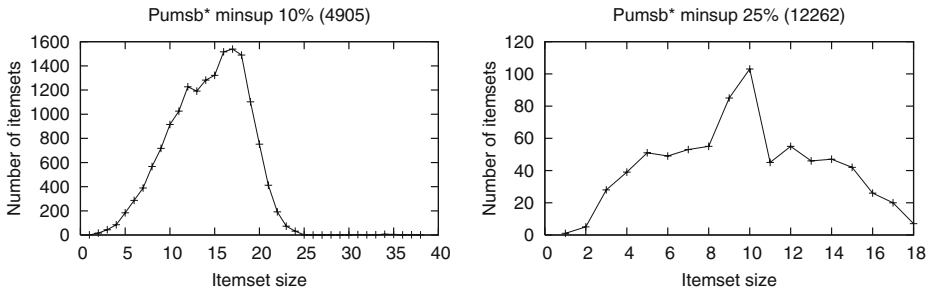

Fig. 2 Distribution of maximal frequent itemsets for Pumsb* (minsup 10% and 25%)

We highlight two limitations of this classification. First, its variability with respect to minimum support threshold values. For example, a dataset could belong to the first category for a given threshold value, and to the second category for another threshold value. As a concrete example, this case arises with *Pumsb*∗ dataset for minimum support threshold values equal to 10% and 25%. As shown by the Fig. 2, *Pusmb*∗ for a minimum support threshold of 10% corresponds to the type 2 of the classification (Table 1), whereas for a minimum support threshold of 25% it corresponds to the type 1 (Table 1). Other examples are given in Flouvat (2008).

Secondly, there is no clear relationship between the proposed classification and algorithms performances. Even worse, a surprising result was obtained in the last FIMI workshop (see Bayardo et al. 2004): algorithms seem to be more efficient on some very dense datasets than on some other sparser datasets. For example in Fig. 3, algorithms seems more efficient on *Mushroom* than on *Chess*, whereas *Mushroom* has much longer maximal frequent itemsets (Fig. 1).

The Fig. 4 represents this difference between *Mushroom* and *Chess* by comparing the average execution time for *Apriori* and *Eclat* (left figure), and for the other algorithms (right figure).

Other works such as Ramesh et al. (2003, 2005), Palmerini et al. (2004) showed that observations from datasets study could be very useful in many fields, from performance prediction, minimum support threshold range determination, sampling, generation of synthetics datasets to strategy decisions.

In Ramesh et al. (2003, 2005), the positive border distribution (i.e. the number of maximal elements in each level) is considered as a key parameter to characterize and

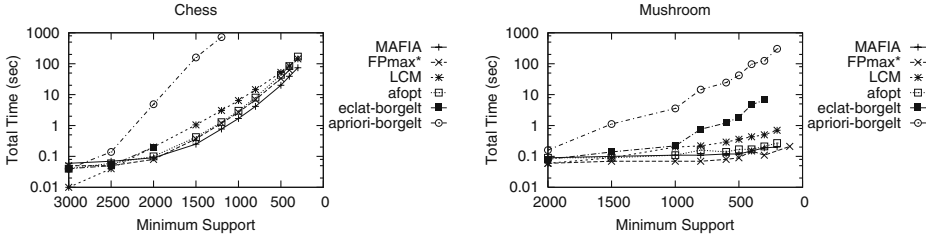

Fig. 3 Algorithms performances for maximal frequent itemsets discovery for *Chess* and *Mushroom* (Bayardo et al. 2004)

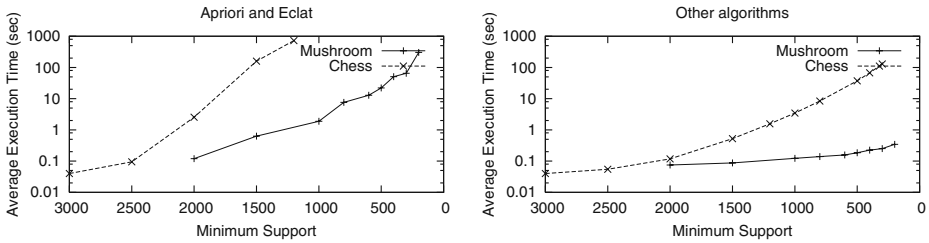

Fig. 4 Algorithms average execution time for maximal frequent itemsets discovery for *Chess* and *Mushroom*

generate transaction databases. It is proved that any distribution is "feasible", and thus susceptible to be met in practice. Moreover, a constructive theorem is proposed to compute a synthetic transaction database given a positive border distribution as input. Nevertheless, the negative border is never considered and as a result, such synthetic databases do not match the "complexity" of real-world datasets.

In Palmerini et al. (2004), the authors proposed a statistical property of transactional datasets to characterize dataset density. They considered the dataset as a transaction source and measure an entropy signal, i.e. the transactions produced by such a source. The authors showed how this characteristic could be used in many fields, such as predicting the number of frequent itemsets for a given minimum support threshold. Thus, it is possible to estimate the algorithms performances for very low supports thresholds. Nevertheless, it does not always explain algorithms performances anymore. This may be due to the fact that frequent itemsets only are used to calculate the entropy measure.

3 Preliminaries

3.1 Frequent itemsets

Let R be a set of symbols called *items*, and r a transaction database of subsets of R. The elements of r are called transactions. An *itemset* X is a set of some items of R. The support of X is the number of transactions in r that contain all items of X. An itemset is frequent if its support in r exceeds a minimum support threshold value, called *minsup*. Given a minimal support threshold and a transaction database, the goal is to find all frequent itemsets FI.

We recall the notion of borders of a set using notations given in Mannila and Toivonen (1997). The positive border of frequent itemsets $\mathcal{B}d^+(FI)$ is the set of all maximal frequent itemsets w.r.t. set inclusion. The negative border of frequent itemsets $\mathcal{B}d^-(FI)$ is the set of all minimal unfrequent itemsets w.r.t. set inclusion.

$$\mathcal{B}d^+(FI) = \{X \in FI \mid \forall Y \supset X, Y \notin FI\}$$

$$\mathcal{B}d^-(FI) = \left\{X \in 2^{\mathbf{R}} \setminus FI \mid \forall Y \subset X, Y \in FI\right\}$$

Each border represents all frequent itemsets, i.e. using one of the borders it is possible to determine if any itemset of the search space is frequent or not without accessing data.

This notion of borders can be generalized in the following way: A set $S \subseteq 2^{\mathbf{R}}$ is *closed downwards* if, for all $X \in S$, all subsets of X are also in S. S can be represented by its *positive border* $\mathcal{B}d^+(S)$ or its *negative border* $\mathcal{B}d^-(S)$ defined by: $\mathcal{B}d^+(S) = \{X \in S \mid \forall Y \supset X, Y \notin S\}$ and $\mathcal{B}d^-(S) = \left\{X \in 2^{\mathbf{R}} \setminus S \mid \forall Y \subset X, Y \in S\right\}$.

Let p be an anti-monotone predicate on $(2^{\mathbf{R}}, \subseteq)$, i.e. $\forall X, Y \in 2^{\mathbf{R}}, X \subseteq Y$, if $p(Y)$ is true, then $p(X)$ is true. If S is the subset of $2^{\mathbf{R}}$ satisfying p, then S is said to be closed downwards.

In order to introduce our experimental study, we recall three classical representations of frequent itemsets.

3.2 Usual representation of frequent itemsets

Several concise (or condensed) representations of frequent itemsets have been studied see for example Calders and Goethals (2003), Mannila and Toivonen (1996). Their goal is twofold: improving efficiency of frequent itemsets mining whenever possible, and compacting the storage of frequent itemsets for future usages.

Formally, a condensed representation must be equivalent to frequent itemsets: one can retrieve each frequent itemset *together with its frequency* without accessing data (see Calders and Goethals 2003). Such a representation is known as *closed sets* (Pasquier et al. 1999). Two other representations are considered in this paper: frequent free itemsets (Bastide et al. 2000; Boulicaut et al. 2003) and frequent essential itemsets (Casali et al. 2005). We believe that this choice of concise representations covers a fairly large range of typical cases. Notice that these sets do not convey enough information to be condensed representation of frequent sets. They need one of the borders to become a condensed representations (Calders and Goethals 2003).

We briefly describe these representations in the rest of this section.

Frequent Closed sets Given an itemset X, *the closure* of X is the set of all items that appear in all transactions where X appears. Formally, given a transaction database r:

$$Cl(X) = \bigcap \{t \in r | X \subseteq t\}$$

If $Cl(X) = X$ then X is said to be closed.

Frequent free itemsets An itemset X is said to be free if there is no exact rule of the form $X_1 \rightarrow X_2$ where X_1 and X_2 are distinct subsets of X. Free sets can be efficiently detected through the following property:

$$X \text{ is free } \iff \forall x \in X, sup(X) < sup(X - x)$$

Frequent essential itemsets The notion of essential itemsets has been defined in Casali et al. (2005). It is based on the notion of disjunctive rule defined in Bykowski and Rigotti (2001), Kryszkiewicz and Gajek (2002). A *disjunctive rule* is of the form $X \rightarrow A_1 \vee A_2 ... \vee A_n$. Such a rule is satisfied if, every transaction that contains X contains at least one of the elements $A_1, ..., A_n$.

An itemset X is said to be essential if there is no *disjunctive rule* of the form $A_1 \rightarrow A_2 \vee \ldots \vee A_k$, where $(A_i)_{i=1..k}$ are distinct elements in X. As for free sets, they can be efficiently tested exploiting the following property:

$$X \text{ is essential } \Longleftrightarrow \forall x \in X, sup_{dij}(X) > sup_{dij}(X - x)$$

where $sup_{dij}(X) = |\{t \in r \mid t \cap X \neq \emptyset\}|$

The three predicates "being a frequent free itemset" and "being a frequent essential itemset" are anti-monotone w.r.t. set inclusion. In the following, we study the distributions of these three collections w.r.t. itemsets size.

Other concise representations based on the notion of disjunctive rules have been defined, the reader is referred to the general framework proposed in Calders and Goethals (2003) for more details.

4 Thorough experimental study of datasets

4.1 Experimental protocol

For frequent itemsets, a benchmark of fourteen datasets is commonly used (see Goethals 2003). Most of them are real-life datasets, only two being synthetic ones, created using the generator from the IBM Almaden Quest research group. For all these datasets, we have studied the distributions of frequent itemsets, frequent closed, frequent free and frequent essential itemsets for many representative minimum support thresholds w.r.t. itemset size. Moreover, we have studied the positive and *negative* borders distributions of frequent, frequent free and frequent essential itemsets.[1]

To perform these tests, we used algorithms available at the FIMI website (see Goethals 2003). The discovery of frequent itemsets and frequent closed itemsets has been done using $FPClose$ and $FP-growth*$ algorithms from Grahne and Zhu (2003). ABS from Flouvat et al. (2004) has been updated to find frequent free and frequent essential itemsets. All these experimentations are available at Flouvat (2008).

4.2 Experimental results

In order to perform a fair comparison with Gouda and Zaki (2001), results given in this paper focus on the same datasets, i.e. *Chess*, *Pumsb*, *Connect*, *Pumsb*$*$, *Mushroom* and $T10I4D100K$. Some characteristics of these datasets are in Table 2.

Notations used in the sequel are reported in Table 3.

Given a dataset and a minimum support threshold value, the Table 4 describes a typical example of our experimental results. The reader is referred to Flouvat (2008)

[1] The set of closed itemsets is not closed downwards, and thus the notion of borders does not apply.

Table 2 Characteristics of datasets studied in Gouda and Zaki (2001)

Dataset	Number of items	Average transactions size	Number of transactions
Chess	75	37	3196
Connect	129	43	67557
Mushroom	119	23	8124
Pumsb*	2088	50.5	49046
Pumsb	2113	74	49046
T10I4D100K	1000	10	100000

for full results from which the analysis made in this paper has been performed. A wider range of minimum support threshold values and other datasets are also described in Flouvat (2008).

4.3 Analysis

We discuss our experimental results with respect to two main axes: borders distribution of frequent itemsets, and borders distribution of frequent free and essential itemsets.

4.3.1 Borders distribution of frequent itemsets

Consider the positive and negative borders of frequent itemsets from five datasets as given in Fig. 5. In all experiments, the negative border appears to be "lower" than its corresponding positive border. From a theoretical point of view, the negative border may have elements one level after the positive border. This case never occurs in our experiments.

The following will study more in details the distribution of the borders w.r.t. each other.

For *Chess*, *Pumsb* and $T10I4D100K$ (Fig. 5), the borders distributions are close to each other, i.e. the mean of the negative border curve is only a few levels before the mean of the positive border curve.

For datasets *Connect*, $Pumsb*$ and *Mushroom* (Fig. 5), a larger distance between the borders exists.

The dataset $T10I4D100K$ is different from the others since its borders are made of small itemsets.

The interest of this "distance" criteria is that it represents the part of the search space "between" the two borders. Recall that once a border has been find all the remaining frequent itemsets, i.e. those "between" the two borders, can be deduced by anti-monotonicity without accessing data (but not their supports).

Table 3 Notations

FI	Frequent itemsets
FCI	Frequent closed itemsets
FFI	Frequent free itemsets
FEI	Frequent essential itemsets

Table 4 *Chess* dataset, *minsup* = 30%

Itemset size	FI	FCI	Bd-(FI)	Bd+(FI)	FFI	FEI	Bd-(FFI)	Bd+(FFI)	Bd-(FEI)	Bd+(FEI)
1	50	27	25		50	50	25	1	25	1
2	896	338	329		828	828	397	2	397	149
3	9049	2568	928	1	7628	4240	988	34	4376	853
4	59589	13221	3371	9	44096	6283	3440	268	8519	3178
5	273069	49002	10118	89	170161	1635	10178	1343	1764	1186
6	907800	137564	21405	439	456826	116	21416	4876	36	109
7	2255159	303661	33711	1369	875938	1	33720	11963		1
8	4276852	540861	39910	3686	1216501		39910	22521		
9	6291848	787143	33890	8200	1231162		33890	31137		
10	7263312	940504	21894	14804	903996		21894	32243		
11	6626801	923310	10160	21183	474618		10160	25491		
12	4790827	740773	3507	24638	172688		3507	15326		
13	2738089	481499	791	23766	41186		791	6403		
14	1227702	250715	114	18088	5787		114	1951		
15	425896	102977	8	10934	360		8	314		
16	111726	32875		5085	3			3		
17	21328	7908		1734						
18	2757	1370		496						
19	206	145		97						
20	6	6		6						
Total	37282962	5316467	180161	134624	5601828	13153	180438	153876	15117	5477

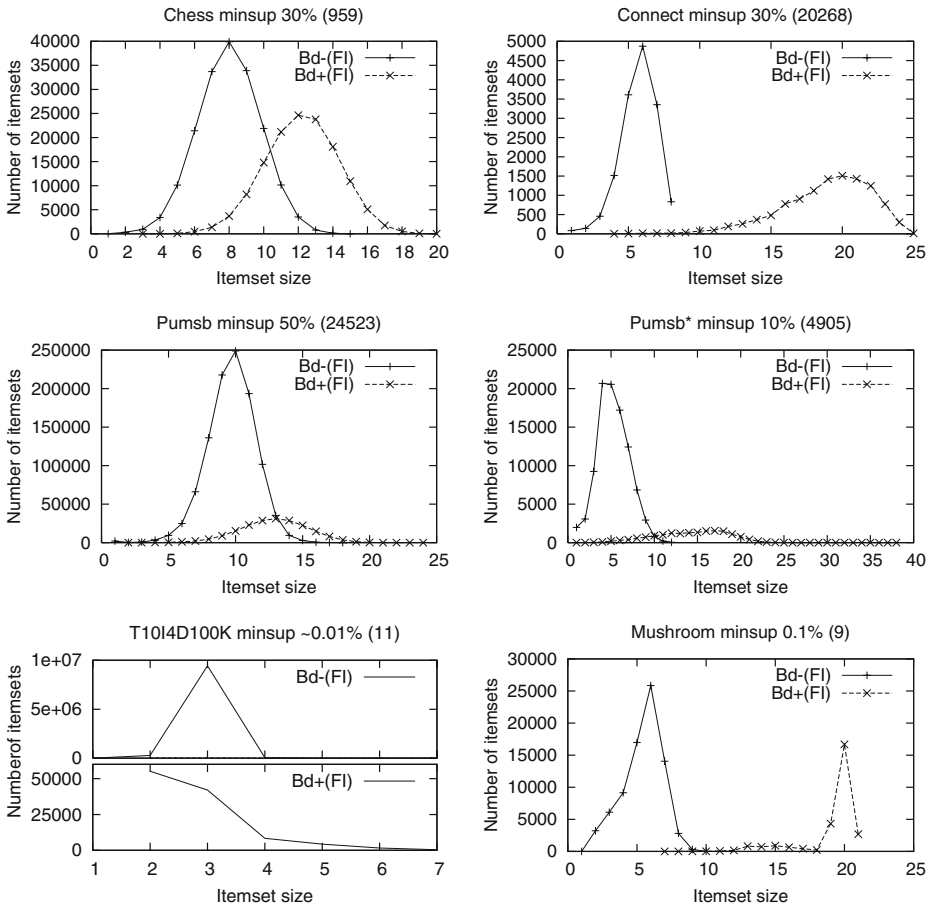

Fig. 5 Borders of frequent itemsets

As we will see, the relative position of the negative border w.r.t. the positive border will be of special interest, in particular to predict the "hardness" of a dataset for algorithms.

4.3.2 Borders of concise representations

Now, we consider the positive and negative borders of *frequent free itemsets* and *frequent essential itemsets* on *Chess* and *Connect* given in Figs. 6 and 7.

From these two figures, the distance between the mean of the negative and positive borders appears to be small for each concise representation. The same behavior has been observed in all our experiments (see Flouvat 2008), suggesting that such kind of distributions is specific to these predicates.

5 A new classification of datasets for frequent itemsets

Observations described in previous section lead us to devise a new classification for datasets w.r.t. borders distribution.

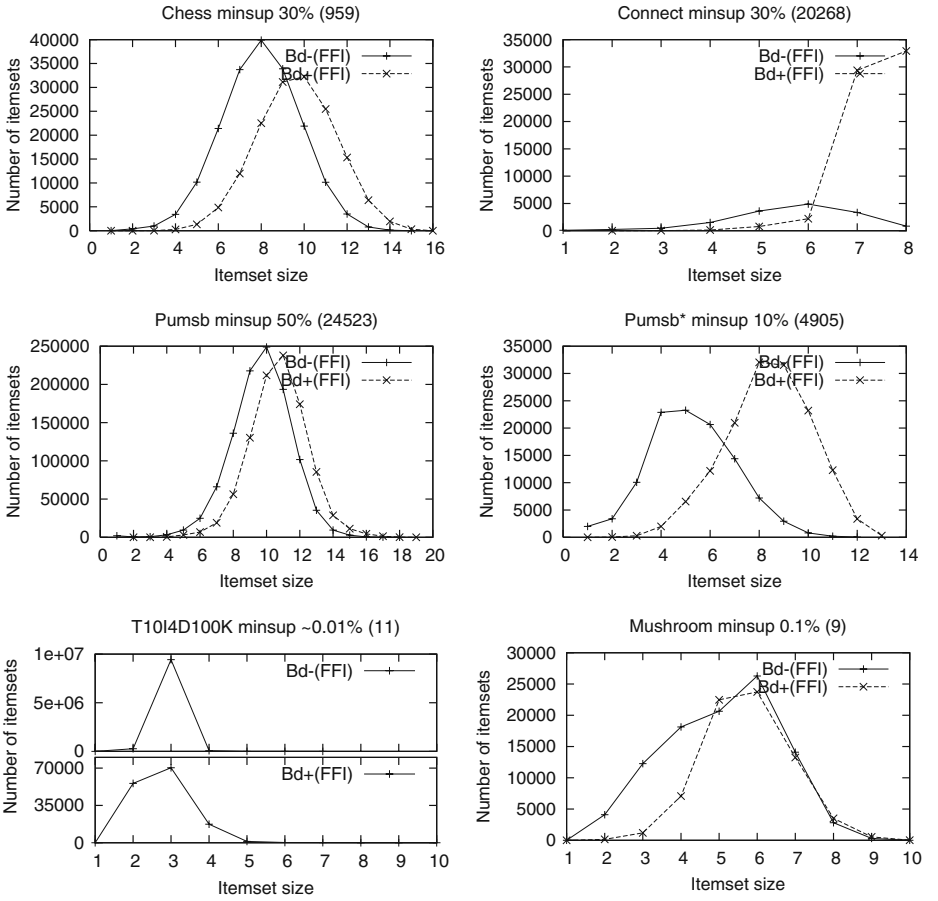

Fig. 6 Borders of frequent free itemsets

5.1 The new classification

This new classification differs from the classification given in Gouda and Zaki (2001) since it takes into account both the negative border and the positive border of frequent itemsets.

These different types of datasets have been identified by taking advantage of the "distance" between positive and negative borders distributions of frequent itemsets. As a consequence, we introduce a new classification of datasets made of three types:

- Type I datasets are datasets having long itemsets in the positive border and a negative border closed to the positive border, i.e. the mean of the negative border curve is not far from the mean of the positive border curve. In other words, most of the itemsets in the two borders have approximately the same size. *Chess* and *Pumsb* fall into this category.
- Type II datasets are datasets having long itemsets in the positive border and a large distance between the two borders distributions. In other words, the itemsets

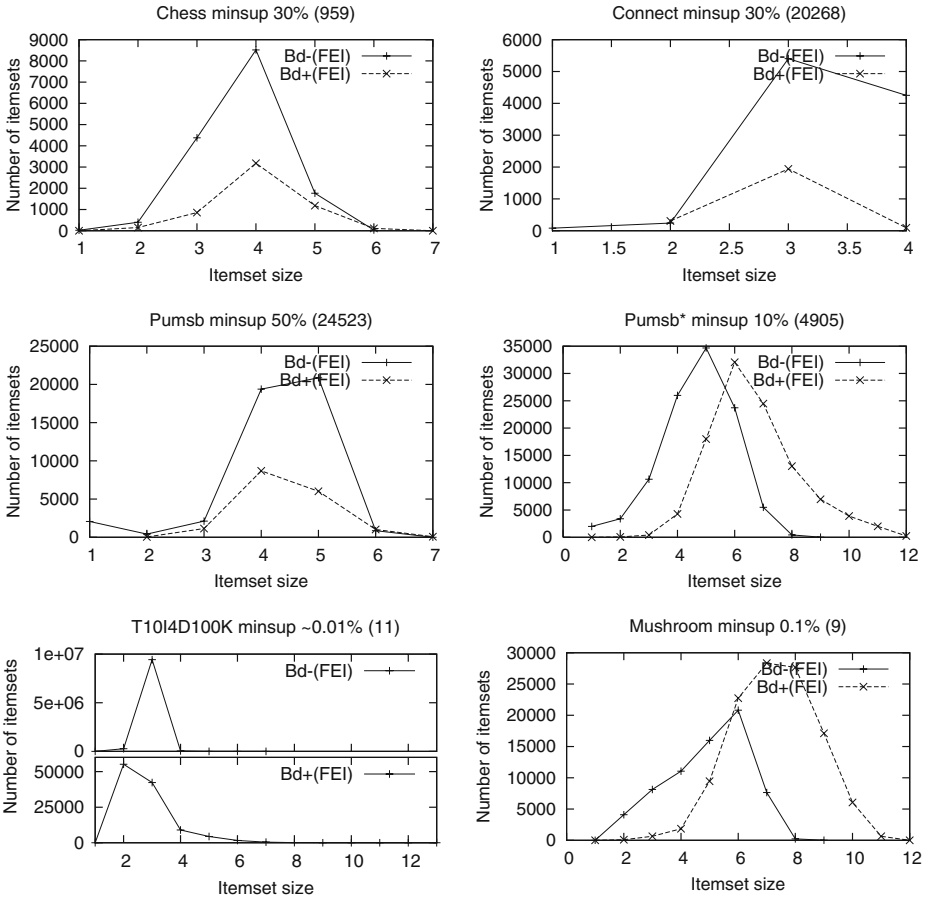

Fig. 7 Borders of frequent essential itemsets

in the negative border are much smaller than those of the positive border. *Connect*, *Pumsb*∗ and *Mushroom* fall into this category.

– Type III is a very special case of type I: the two distributions are very close, but they are concentrated in very low levels. This type allows to catch the notion of sparseness (for example $T10I4D100K$).

The next section will show the two main interests of this classification: the stability w.r.t. variation of minimum support thresholds and the better correspondence with algorithms performances.

5.2 Properties of the new classification

5.2.1 Stability of the classification

In this section, we study the variation of minimum support threshold values on borders distribution of frequent itemsets. A surprising observation is that the borders

distributions and their relative position are *stable* w.r.t. variation of minimum support threshold values. For example in Fig. 8, we consider *Chess*, *Connect*, *Pumsb* and *Pumsb*∗ (a dataset per row) for various minimum support threshold values.

In other words, this observation suggests a kind of *global structure* for frequent itemsets borders distribution invariant to variation of minimum support threshold values.

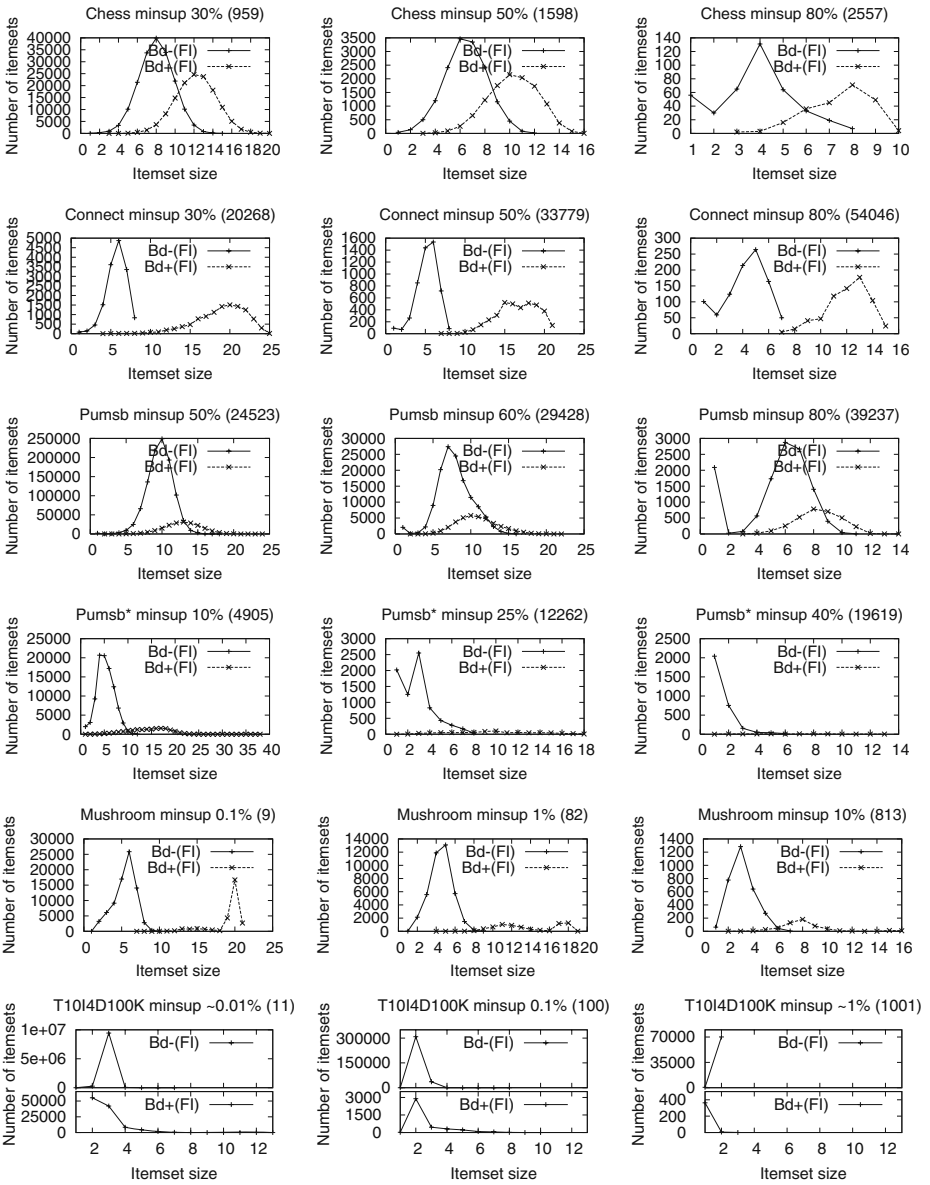

Fig. 8 Borders of frequent itemsets with a different minimum support threshold

Note that this "stability" has been observed for all our experiments (see Flouvat 2008). Moreover, it appears that the distributions of the others itemsets studied (closed frequent itemsets, free frequent, frequent essential and their borders) are also relatively stable w.r.t. each others.

Consequently, the proposed classification is very stable w.r.t. variation of minimum support thresholds, while being simpler than the one presented in Gouda and Zaki (2001).

5.2.2 Impact on algorithms performances

We focus on the discovery of *maximal frequent itemsets*, and we study the performances of implementations available at the FIMI website (see Goethals 2003). Let us consider results given in Fig. 9 showing execution times of major implementations on the datasets studied in Gouda and Zaki (2001). Note that the y-axis represents the execution time on a logarithmic scale, and the x-axis represents the minimum support threshold values in decreasing order.

On *Chess* dataset (Fig. 9, upper-left corner), execution times increase exponentially for every implementation, whereas for *Connect* (Fig. 9,upper-right corner) they appear to be almost linear for *Mafia* (Burdick et al. 2003), $fp-zhu$ (Grahne and Zhu 2003), *LCM* (Uno et al. 2004) and *afopt* (Liu et al. 2003).

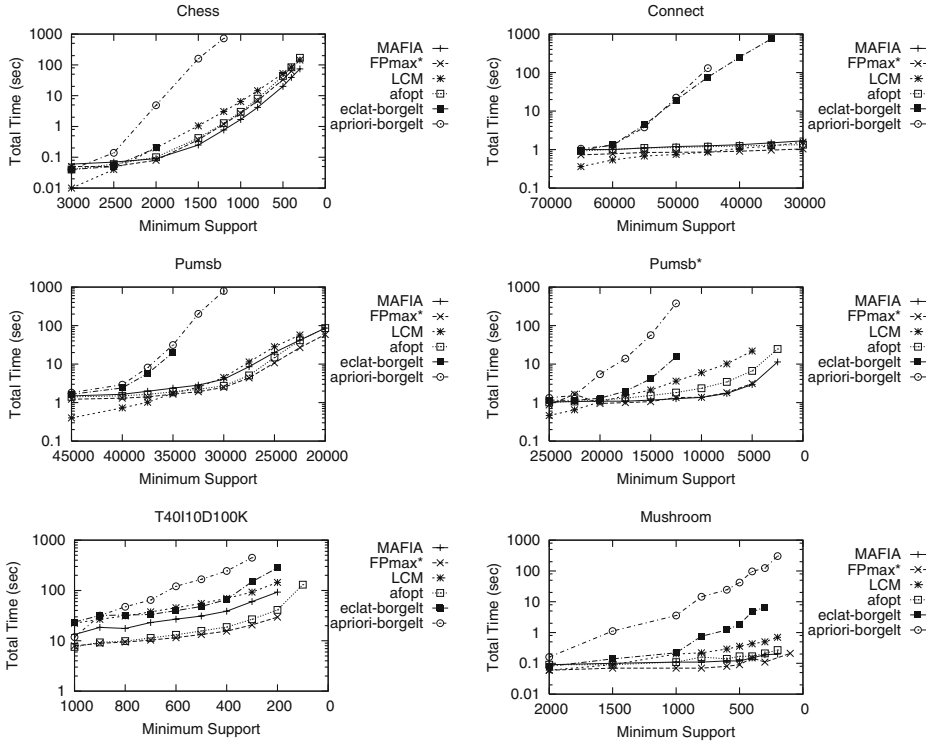

Fig. 9 Algorithms performances

For classical algorithms such as *Apriori* (Agrawal and Srikant 1994) and *Eclat* (Zaki et al. 1997), the execution time is also better for *Connect* than for *Chess*, but still exponential w.r.t. decreasing minimum support threshold.

Recall that *Connect* has more and longer transactions, and more items than *Chess*. Moreover, *Connect* has longer maximal frequent itemsets. Therefore, *Connect* should be "harder" to mine than *Chess*.

The same kind of behavior can be noticed for datasets such as *Chess* and *Mushroom*, or *Pumsb* and *Pumsb*∗ (Fig. 9). For example, the datasets *Pumsb* and *Pumsb*∗ are very similar w.r.t. the transactions and number of items, but their borders distribution is very different (Fig. 5). Algorithms for *Pumsb*∗ are still very effective for very low minimum support threshold, whereas for *Pumsb*, algorithms do not perform very well for relatively high minimum support threshold values.

One could think that this difference is due to a more important number of frequent itemsets for datasets such as *Chess* or *Pumsb*, since in the worst case algorithms have to explore all these itemsets (in addition to some unfrequent itemsets). But as shown by the Table 5, the number of frequent itemsets for *Connect* (30%) and *Mushroom* (2.5%) is more important than *Chess* (34.4%), whereas the algorithms are more efficient for these datasets and minimum support thresholds. This remark is also true for *Pumsb*∗(10%) and *Pumsb* (50%).

Therefore, we deduce from Fig. 9 that the position of the borders in the search space and the "distance" between them influence implementations performances. Algorithms are indeed more efficient when the negative border is mainly composed of small itemsets, i.e. for the datasets of type II and III w.r.t. the new classification. However, implementations of "classical" algorithms such as *Apriori* and *Eclat* still have difficulties when the positive border is composed of long itemsets, i.e. for type II datasets. The efficiency of the others algorithms depends on the "distance" between the two borders: the more important is the distance, the more implementations are likely to be efficient.

To summarize, the following cases may arise:

- either the borders are composed of small itemsets (i.e. the borders are in the "low" levels of the search space), algorithms have no difficulties until very low supports; such dataset belongs to the type III of our classification;
- or the positive border have long itemsets:
  - either a large distance between the two borders does exist. Every implementations perform well except classical algorithms such as *Apriori* and *Eclat*; such dataset belongs to the type II of our classification;
  - or there is a small distance between the two borders, all the algorithms have difficulties; such dataset belongs to the type I of our classification.

Table 5 Number of frequent itemsets

Chess minsup 34.4% (1100)	16 763 342
Connect minsup 30% (20268)	1 331 673 367
Pumsb minsup 50% (24523)	$\approx 1,65 \times 10^8$
Pumsb∗ minsup 10% (4905)	$\approx 5,5 \times 10^{11}$
Mushroom minsup 2.5% (200)	18 094 857

In addition to classical characteristics of datasets, the "distance" between the mean of the negative and positive border distributions makes possible a better evaluation of the difficulty of a dataset.

Consequently, the main interests of the new classification are:

– a better correspondence between algorithms performances and the classification. In other words, this classification is a first attempt in order to evaluate the "hardness" of a dataset.
– a stability w.r.t. the variation of minimum support thresholds.

The Table 6 summarizes this new classification and shows the type of each dataset used for FIMI.

Finally, we also intent to use these results for other data mining problems, i.e. those problems said to be "representable as sets" (defined in Mannila and Toivonen 1997). In the next section, we will more particularly show how this study could also be applied to the discovery of inclusion dependencies.

6 Towards predicate classification

In the setting of this paper, we focus our analysis on datasets with respect to frequent itemsets. In our experiments, we studied three anti-monotone predicates, one for frequent itemsets, another one for frequent free itemsets and the last one for frequent essential itemsets. These three predicates exhibit very different behaviors on the same dataset (see Fig. 5 to 7 on *Connect* and *Chess* for different minimum support threshold values). Moreover, note that for frequent free and frequent essential itemsets, the new classification suggests that almost all datasets belong to type I or III.

Quite clearly, this work could be generalized to other data mining problems, i.e. those which are *representable as sets* (Mannila and Toivonen 1997). We argue that the study of both positive and negative borders for a given anti-monotone predicate may allow us to come up with some general results.

From the previous sections, we deduced that studying the gap between the negative and positive borders may be very insightful to explain the behavior of algorithms and may also give some hints to guess the existence of properties associated with anti-monotone predicates. In spite of the huge amount of work done for frequent itemset

Table 6 New classification of frequent itemsets datasets

Type	Type I	Type II	Type III
Algorithms performances (from FIMI experimentations)	Fast	Slow	Fast for most of the supports thresholds Slow for very small supports thresholds (<1%)
"Distance" between the borders	Small	Large	Small
Itemsets size	Long	Long	Small
Examples of datasets	*Chess*, *Pumsb*, *Accidents*, *Webdocs*	*Connect*, *Pumsb*∗, *Mushroom*	$T10I4D100K$, $BMS-Pos$, $BMS-WebView1$, *retail*, $BMS-WebView2$, *Kosarak*, $T40I10D100K$

Fig. 10 An interaction between FD and IND

$$\left.\begin{array}{c} R[XY] \subseteq S[UV] \\ R[XZ] \subseteq S[UW] \\ S : U \rightarrow V \end{array}\right| \Rightarrow \quad R[XYZ] \subseteq S[UVW]$$

mining, we are not aware of such kind of contributions. Nevertheless, we introduce in the sequel another data mining problem known to be representable as sets where such properties have been clearly identified (see Marchi and Petit 2003).

Application to inclusion dependency mining Inclusion dependencies (IND) are fundamental semantic constraints for relational databases (see Mannila and Räihä 1994). Let r and s be two relations over schemas R and S, and X and Y be sequences of attributes into R and S respectively. The IND $R[X] \subseteq S[Y]$ is true in (r, s) if all the values of X in r are also values of Y in s. This notion generalizes foreign keys constraints, very popular in practice.

The underlying data mining problem can be stated as follows: "Given a database, find all inclusion dependencies satisfied in this database" (see Kantola et al. 1992; Mannila and Toivonen 1997; Koeller and Rundensteiner 2003; Marchi and Petit 2003 for related works). From Mannila and Toivonen (1997), the set of IND candidates can be organized in a levelwise manner; a given level, say k, corresponds to INDs whose arity is equal to k. Moreover, a partial order for INDs can be defined as follows: if i and j are two INDs, $j \preceq i$ if j can be obtained by performing the same projection on the two sides of i. For example, $R[AB] \subseteq S[EF] \preceq R[ABC] \subseteq S[EFG]$. In this setting, the predicate "being satisfied in a database" is anti-monotone with respect to $\preceq$ (see Mannila and Toivonen 1997 for the proof).

Consider now the well known inference rule for inclusion dependencies together with functional dependencies (see Casanova et al. 1984) given in Fig. 10. Intuitively, consider an inclusion dependency $i = R[XAB] \subseteq S[YEF]$ where X and Y are attribute sequences and A, B, E and F are single attributes. Suppose that every IND j such that $j \preceq i$ is satisfied, and let $j_1 = R[XA] \subseteq S[YE]$ and $j_2 = R[XB] \subseteq S[YF]$ be two of them. The more $|Y|$ is large, the more Y is likely to determine E or F. In other words, i is likely to be satisfied (from inference rule of Fig. 10).

From this result, one may logically expect that large INDs should never appear in the negative border, even if large INDs exist. It implies a potentially large gap between the two borders distribution, like for type II datasets for frequent itemsets.

All our experiments corroborate this hypothesis; We tested three synthetic databases built using the *chase* procedure presented in Beeri and Vardi (1984). We enforced large INDs in their positive border, until size 18. For all databases, INDs in the negative border were all of size lower than 3.

This particular behavior of the positive border of INDs underlines the interest of frequent itemsets algorithms for this problem, and justifies an algorithm based on the negative border discovery (Marchi and Petit 2003).

7 Conclusion and perspectives

In this paper, we have studied datasets for problems related to frequent itemset mining. We have shown that the distribution of the negative and positive borders

have an important impact on datasets classification and algorithms performances. For frequent itemsets mining, a new classification of datasets has been proposed. This work is a first step toward a better understanding of the behavior of algorithms with respect to the search space to be discovered.

This work has two main perspectives. The former is to find out theoretical foundation of the stability obtained for the distributions in most of our experiments. The latter is the design of adaptive algorithms with respect to dataset characteristics, i.e. changing dynamically their strategy during runtime.

References

Agrawal, R., Imielinski, T., & Swami, A. N. (1993). Mining association rules between sets of items in large databases. In P. Buneman & S. Jajodia (Ed.), *SIGMOD Conference* (pp. 207–216). New York: ACM.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In J. B. Bocca, M. Jarke, & C. Zaniolo (Ed.), *VLDB* (pp. 487–499). San Francisco: Morgan Kaufmann.

Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., & Lakhal, L. (2000). Mining frequent patterns with counting inference. *SIGKDD Explorations, 2*(2), 66–75.

Bayardo Jr., R. J., Goethals, B., & Zaki, M. J. (Eds.) (2004). *FIMI '04, proceedings of the IEEE ICDM workshop on frequent itemset mining implementations, November 1, 2004, CEUR workshop proceedings* (Vol. 126). Brighton, UK: CEUR-WS.org.

Bayardo Jr., R. J., & Zaki, M. J. (Eds.) (2003). *FIMI '03, proceedings of the IEEE ICDM workshop on frequent itemset mining implementations, November 19, 2003, CEUR workshop proceedings* (Vol. 90). Melbourne, Florida, USA: CEUR-WS.org.

Beeri, C., & Vardi, M. Y. (1984). A proof procedure for data dependencies. *Journal of the Association for Computing Machinery, 31*(4), 718–741.

Borgelt, C. (2003). Efficient implementations of Apriori and Eclat. In R. J. Bayardo Jr., & M. J. Zaki (Eds.), *1st workshop of frequent item set mining implementations*. Melbourne, FL, USA: FIMI 2003.

Boulicaut, J.-F., Bykowski, A., & Rigotti, C. (2003). Free-sets: A condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery, 7*(1), 5–22.

Burdick, D., Calimlim, M., Flannick, J., Gehrke, J., & Yiu, T. (2003). Mafia: A performance study of mining maximal frequent itemsets. In R. J. Bayardo Jr. & M. J. Zaki (Eds.), *Journal of Intelligent Information Systems*.

Burdick, D., Calimlim, M., & Gehrke, J. (2001). Mafia: A maximal frequent itemset algorithm for transactional databases. In *ICDE* (pp. 443–452). Los Alamitos: IEEE Computer Society.

Bykowski, A., & Rigottim, C. (2001). A condensed representation to find frequent patterns. In *PODS*. New York: ACM.

Calders, T., & Goethals, B. (2003). Minimal *k*-free representations of frequent sets. In N. Lavrac, D. Gamberger, H. Blockeel, & L. Todorovski (Eds.), *PKDD*. *Lecture Notes in Computer Science* (Vol. 2838, pp. 71–82). New York: Springer.

Casali, A., Cicchetti, R., & Lakhal, L. (2005). Essential patterns: A perfect cover of frequent patterns. In A. Min Tjoa & J. Trujillo (Eds.), *DaWaK*. *Lecture notes in computer science* (Vol. 3589, pp. 428–437). New York: Springer.

Casanova, M. A., Fagin, R., & Papadimitriou, C. H. (1984). Inclusion dependencies and their interaction with functional dependencies. *Journal of Computer and System Sciences, 28*(1), 29–59.

De Marchi, F., & Petit, J.-M. (2003). Zigzag: a new algorithm for mining large inclusion dependencies in database. In *ICDM* (pp. 27–34). Los Alamitos, IEEE Computer Society.

Flouvat, F. (2008). Study of frequent itemsets datasets. http://pages.univ-nc.nc/~flouvat/.

Flouvat, F., De Marchi, F., & Petit, J.-M. (2004). ABS: Adaptive Borders Search of frequent itemsets. In R. J. Bayardo Jr., B. Goethals, & M. J. Zaki (Eds.), *Journal of Intelligent Information Systems*.

Goethals, B. (2003). Frequent itemset mining implementations repository. http://fimi.cs.helsinki.fi/.

Gouda, K., & Zaki, M. J. (2001). Efficiently mining maximal frequent itemsets. In N. Cercone, T. Y. Lin, & X. Wu (Eds.), *ICDM* (pp. 163–170). Los Alamitos: IEEE Computer Society.

Grahne, G., & Zhu, J. (2003). Efficiently using prefix-trees in mining frequent itemsets. In R. J. Bayardo Jr., B. Goethals, & M. J. Zaki (Eds.), *Journal of Intelligent Information Systems*.

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. In W. Chen, J. F. Naughton, & P. A. Bernstein (Eds.), *SIGMOD conference* (pp. 1–12). New York: ACM.

Kantola, M., Mannila, H., Räihä, K.-J., & Siirtola, H. (1992). Discovering functional and inclusion dependencies in relational databases. *International Journal of Intelligent Systems, 7*, 591–607.

Koeller, A., & Rundensteiner, E. A. (2003). Discovery of high-dimensional. In U. Dayal, K. Ramamritham, & T. M. Vijayaraman (Eds.), *ICDE* (pp. 683–685). Los Alamitos: IEEE Computer Society.

Kryszkiewicz, M., & Gajek, M. (2002). Concise representation of frequent patterns based on generalized disjunction-free generators. In M.-S. Cheng, P. S. Yu, & B. Liu (Eds.), *PAKDD*. *Lecture notes in computer science* (Vol. 2336, pp. 159–171). New York: Springer.

Liu, G., Lu, H., Yu, J. X., Wei, W., & Xiao, X. (2003). Afopt: An efficient implementation of pattern growth approach. In R. J. Bayardo Jr., B. Goethals, & M. J. Zaki (Eds.), *Journal of Intelligent Information Systems*.

Mannila, H., & Räihä, K-J. (1994). *The design of relational databases (2nd ed.)*. Reading: Addison-Wesley.

Mannila, H., & Toivonen, H. (1996). Multiple uses of frequent sets and condensed representations (extended abstract). In *KDD* (pp. 189–194).

Mannila, H., & Toivonen, H. (1997). Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery, 1*(3), 241–258.

Orlando, S., Lucchese, C., Palmerini, P., Perego, R., & Silvestri, F. (2003). kdci: a multi-strategy algorithm for mining frequent sets. In R. J. Bayardo Jr., B. Goethals, & M. J. Zaki (Eds.), *Journal of Intelligent Information Systems*.

Palmerini, P., Orlando, S., & Perego, R. (2004). Statistical properties of transactional databases. In H. Haddad, A. Omicini, R. L. Wainwright, & L. M. Liebrock (Eds.), *SAC* (pp. 515–519). New York: ACM.

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999). Discovering frequent closed itemsets for association rules. In C. Beeri & P. Buneman (Eds.), ICDT. *Lecture notes in computer science* (Vol. 1540, pp. 398–416). New York: Springer.

Ramesh, G., Maniatty, W., & Zaki, M.-J. (2003). Feasible itemset distributions in data mining: theory and application. In *PODS* (pp. 284–295). New York: ACM.

Ramesh, G., Zaki, M.-J., & Maniatty, W. (2005). Distribution-based synthetic database generation techniques for itemset mining. In *IDEAS* (pp. 307–316). Los Alamitos: IEEE Computer Society.

Uno, T., Asai, T., Uchida, Y., & Arimura, H. (2004). An efficient algorithm for enumerating closed patterns in transaction databases. In E. Suzuki, S. Arikawa (Eds.), Discovery Science. *Lecture notes in computer science* (Vol. 3245, pp. 16–31). New York: Springer.

Zaki, M.-J., Parthasarathy, S., Ogihara, M., & Li, W. (1997). New algorithms for fast discovery of association rules. In *KDD* (pp. 283–286). *Journal of Intelligent Information Systems*.