DIMACS Technical Report 2000-37
October, 2000

An inequality limiting the number of maximal frequent sets

by

Endre Boros[1]
RUTCOR, Rutgers University
640 Bartholomew Road
Piscataway NJ 08854
(boros@rutcor.rutgers.edu)

Vladimir Gurvich
Russian Academy of Sciences
Moscow, Russia
(gurvich@rutcor.rutgers.edu)

Leonid Khachiyan[1]
Department of Computer Science
Rutgers University
New Brunswick, New Jersey 08903
(leonid@cs.rutgers.edu)

Kazuhisa Makino
Department of Systems and Human Science
Graduate School of Engineering Science
Osaka University
Toyonaka, Osaka, 560-8531, Japan
(makino@sys.es.osaka-u.ac.jp)

[1]Permanent Members of DIMACS
[2]This research was supported in part by the Office of Naval Research (Grant N00014-92-J-1375). The first author is also grateful for the partial support by the National Science Foundation (Grant DMS 98-06389). Visits of the second author to Rutgers University were also supported by DIMACS (Grant NSF STC 91-19999). The research of the third author was also supported by the National Science Foundation (Grant CCR-9618796).

ABSTRACT

Given an arbitrary $m \times n$ binary matrix A and a threshold $t \in \{1, \ldots, m\}$, a subset C of the columns of A is called *t-frequent* if there are at least t rows in A having a 1 in each of these columns, and otherwise C is called *t-infrequent*. Denoting by α the number of maximal t-frequent, and by β the number of minimal t-infrequent column sets in the given matrix A, we prove that the inequality $\alpha \leq (m - t + 1)\beta$ holds. This inequality is sharp, and allows for an incremental quasi-polynomial algorithm for generating all minimal t-infrequent sets, while the analogous problem for maximal t-frequent sets is NP-hard. Our proof is based on an inequality from extremal set theory, which may be of independent interest.

Key Words: data mining, frequent sets, infrequent sets, independent sets, hitting sets, transversals, dualization.

1 Introduction

Let us consider an $m \times n$ binary matrix $A : \mathcal{R} \times \mathcal{C} \to \{0, 1\}$, and an integral threshold value $t \in \{1, \dots, m\}$. To each subset of columns $C \subseteq \mathcal{C}$, let us associate the subset $R(C) \subseteq \mathcal{R}$ of all those rows $r \in \mathcal{R}$, for which $A(r, c) = 1$ in every columns $c \in C$. The cardinality $|R(C)|$ is called the *support* of the set C. Let us call a subset $C \subseteq \mathcal{C}$ of the columns *frequent* (or more precisely, *t-frequent*) if its support is at least the given integral threshold t, i.e. if $|R(C)| \geq t$, and let us denote by $\mathcal{F}_t$ the family of all t-frequent subsets of the columns of the given binary matrix A. Let us further call a subset $C \subseteq \mathcal{C}$ *infrequent* (or *t-infrequent*) if its support does not exceed the given threshold t, i.e. if $|R(C)| < t$. Clearly, subsets of frequent sets are also frequent, and supersets of infrequent sets are also infrequent. Let us denote by $\mathcal{M}_t \subseteq \mathcal{F}_t$ the family of all maximal t-frequent sets (i.e. those which are t-frequent, but no superset of them is t-frequent), and by $\mathcal{I}_t$ the family of all minimal t-infrequent sets (i.e. those which are infrequent but all proper subsets of them are t-frequent.)

The generation of frequent sets of a given binary matrix A is an important task of knowledge discovery and data mining, e.g. it is used for mining association rules [1, 2, 15, 19, 20], correlations [8], sequential patterns [3], episodes [21], emerging patterns [10], and appears in many other applications. Most practical procedures to generate $\mathcal{F}_t$ are based on the anti-monotone *Apriori* heuristic (see [2]) and build frequent sets in a bottom-up way, running in time proportional to the number of frequent sets. It was also demonstrated recently in [9] that these methods are inadequate in practice when there are (many) frequent sets of large size (see also [4, 12, 17]), due the fact that $|\mathcal{F}_t|$ can be exponentially larger than $|\mathcal{M}_t|$.

These results show that it is perhaps more important to find the *boundary* of the frequent sets, i.e. the families of maximal frequent and minimal infrequent sets $\mathcal{M}_t \cup \mathcal{I}_t$ (proposed e.g. in [22]), and use those as a condensed representation of the data set, as suggested in [19]. Furthermore, no algorithm using membership queries "$X \in \mathcal{F}_t$?" can generate all (maximal) frequent sets in fewer than $|\mathcal{M}_t \cup \mathcal{I}_t|$ steps (see e.g. [15]). There were several other examples presented in [19] to show the usefulness of maximal frequent sets and minimal infrequent sets, e.g. providing error bounds for the confidence of an arbitrary Boolean rule, in terms of minimal infrequent sets.

In this short paper we prove the following inequality, improving on a weaker quadratic relation shown in [6].

Theorem 1 *If $\mathcal{I}_t \neq \emptyset$ then*

$$|\mathcal{M}_t| \leq (m - t + 1)|\mathcal{I}_t|. \tag{1}$$

Note that the requirement that $\mathcal{I}_t$ be non-empty is necessary because for $|\mathcal{I}_t| = 0$ we would have $\mathcal{M}_t = \{\mathcal{C}\}$ and hence $|\mathcal{M}_t| = 1$ in contradiction with (1). The condition $\mathcal{I}_t \neq \emptyset$ thus excludes the degenerate case when the entire columns set of A is t-frequent.

Before proceeding further, let us mention some algorithmic implications of (1). It follows from the results of [5, 15, 16] that the incremental complexity of generating $\mathcal{M}_t \cup \mathcal{I}_t$ is equivalent with that of the transversal hypergraph problem (for definitions and related results see e.g. [11]). The latter problem is known to be solvable in incremental quasi-polynomial time [13], implying thus the same for the joint generation of maximal frequent and minimal infrequent sets. Specifically, it follows from [13] that for each $k \leq |\mathcal{M}_t \cup \mathcal{I}_t|$, we can generate k sets in $\mathcal{M}_t \cup \mathcal{I}_t$ in $poly(n,m) + k^{o(\log k)}$ time. The above inequality (1) clearly implies that if we can generate $\mathcal{M}_t \cup \mathcal{I}_t$ in time $T(|\mathcal{M}_t \cup \mathcal{I}_t|)$, then the entire set $\mathcal{I}_t$ can be generated in time $T((m-t+2)|\mathcal{I}_t|)$. We thus conclude that the family of minimal infrequent sets $\mathcal{I}_t$ can be generated in output quasi-polynomial time, i.e. in time bounded by a quasi-polynomial in $|\mathcal{I}_t|$. This can be further improved to show that the incremental complexity of generating $\mathcal{I}_t$ is also equivalent with that of the transversal hypergraph problem (see [6, 7] for more detail). Hence

Corollary 1 *For each $k \leq |\mathcal{I}_t|$, we can compute k minimal t-infrequent sets of the columns of the matrix A in $poly(n,m)$ + $k^{o(\log k)}$ time.*

In contrast to this we include here a negative result about the generation of maximal frequent sets strengthening a statement of [18].

Theorem 2 *Given an $m \times n$ matrix A, a threshold t, and a subfamily $\mathcal{S} \subseteq \mathcal{M}_t$, it is NP-hard to decide if $\mathcal{S} \neq \mathcal{M}_t$, even if $|\mathcal{S}| = O(n)$ and $|\mathcal{M}_t|$ is exponentially large in n whenever $\mathcal{S} \neq \mathcal{M}_t$.*

Finally, let us remark that the inequality (1) is best possible, as the following examples show. Let A be an $m \times (m-t+1)$ matrix, in which every entry is 1, except the diagonal entries in the first $m-t+1$ rows, which are 0. Then any $m-t$ element subset of the columns is a maximal t-frequent set, while the set $\mathcal{C}$ of all columns is the only minimal t-infrequent set. Thus we have equality in (1) for such matrices.

It is also worth mentioning that (1) stays accurate, up to a factor of $\log m$, even if $m \gg n$ and $|\mathcal{I}_t|$ is arbitrarily large. To see this, let us consider a binary matrix A with $m = 2^k$ rows and $n = 2k$ columns ($k \geq 1$, integer), such that each row contains exactly one 0 and one 1 in each pair of the adjacent columns $\{1,2\}$, $\{3,4\}$, …, $\{2k-1, 2k\}$, and in all 2^k possible ways in the $m = 2^k$ rows. It is not difficult to see that for $t = 1$ there are 2^k maximal 1-frequent sets (every row of the matrix is the characteristic vector of a maximal 1-frequent set), and that there are only k minimal 1-infrequent sets, namely $\{2i-1, 2i\}$ for $i = 1, ..., k$. Thus, for such examples we have $|\mathcal{M}_t| = (m/\log m)|\mathcal{I}_t|$. The same examples also show that $|\mathcal{M}_t|$ cannot be bounded by a polynomial function in $|\mathcal{I}_t|$ and n, the number of columns of A. Needless to say that in general, $|\mathcal{I}_t|$ cannot be bounded by a polynomial in $|\mathcal{M}_t|$, n and m.

2 Proofs

For the proof of Theorem 1 we shall need the following combinatorial lemma.

Lemma 1 *Given a base set V of size $|V| = m$ and a threshold $t \in \{1, \ldots, m\}$, let $\mathcal{S} = \{S_1, \ldots, S_\alpha\}$ and $\mathcal{T} = \{T_1, \ldots, T_\beta\}$ be two families of subsets of V such that*

(i) $|S| \geq t$ for all $S \in \mathcal{S}$, while $|T| < t$ for all $T \in \mathcal{T}$, and

(ii) for each of the $\alpha(\alpha-1)/2$ pairs $S', S'' \in \mathcal{S}$ there exists a $T \in \mathcal{T}$, such that $S' \cap S'' \subseteq T$.

Then $\alpha \leq (m - t + 1)\beta$, whenever $\alpha \geq 2$.

Let us remark first that if $\alpha = 1$ then the family $\mathcal{T}$ might be empty, which would violate the inequality $\alpha \leq (m-t+1)\beta$. Let us also mention that by (ii) $\beta \geq 1$ must hold whenever $\alpha \geq 2$. In addition, conditions (i) and (ii) together imply that $\mathcal{S}$ is a Sperner family, i.e. $S_i \not\subseteq S_j$ whenever $i \neq j$ (since otherwise $S_i = S_i \cap S_j \subseteq T_k$ would follow by (ii) for some $T_k \in \mathcal{T}$, contradicting condition (i).) Without loss of generality we can assume that $\mathcal{T}$ is also Sperner, for otherwise we can replace $\mathcal{T}$ by the family of all maximal sets of $\mathcal{T}$.

Proof of Lemma 1. We shall prove the Lemma by induction on t. If $t = 1$ then $\mathcal{T} = \{\emptyset\}$ by condition (i). In view of (ii), this implies that the sets of $\mathcal{S}$ are pairwise disjoint, and hence $\alpha \leq m = (m - t + 1)\beta$.

In a general step, let us define subfamilies $\mathcal{S}_v = \{S \setminus \{v\} \mid S \in \mathcal{S},\ v \in S\}$ and $\mathcal{T}_v = \{T \setminus \{v\} \mid S \in \mathcal{T},\ v \in T\}$ for each $v \in V$. Let us further introduce the notations $\alpha_v = |\mathcal{S}_v|$, and $\beta_v = |\mathcal{T}_v|$.

For vertices $v \in V$ for which $\alpha_v \geq 2$ (and thus $\beta_v \geq 1$) the families $\mathcal{S}_v$ and $\mathcal{T}_v$ satisfy all the assumptions of the Lemma with $m' = m - 1$ and $t' = t - 1$, and hence

$$\alpha_v \leq (m' - t' + 1)\beta_v = (m - t + 1)\beta_v \tag{2}$$

follows by the inductive hypothesis. Let us then consider the partition $V = V_1 \cup V_2$, where $V_1 = \{v \in V \mid \alpha_v \leq 1\}$, and $V_2 = \{v \in V \mid \alpha_v \geq 2\}$. Summing up the inequalities (2) for all $v \in V_2$, we obtain

$$\sum_{v \in V_2} \alpha_v \leq (m - t + 1) \sum_{v \in V_2} \beta_v. \tag{3}$$

On the left hand side, using condition (i) and the definition of α_v we obtain

$$\alpha t - |V_1| \leq \sum_{S \in \mathcal{S}} |S| - |V_1| \leq \sum_{S \in \mathcal{S}} (|S| - |S \cap V_1|) = \sum_{S \in \mathcal{S}} |S \cap V_2| = \sum_{v \in V_2} \alpha_v, \tag{4}$$

where the first inequality follows by $|S| \geq t$ for $S \in \mathcal{S}$, while the second one is implied by $|V_1| \geq \sum_{S \in \mathcal{S}} |S \cap V_1|$, which follows from the definition of V_1.

On the right hand side of (2) we can write

$$\sum_{v \in V_2} \beta_v = \sum_{T \in \mathcal{T}} |T \cap V_2| \leq \sum_{T \in \mathcal{T}} |T| \leq \beta(t-1), \tag{5}$$

where the first equality follows by the definition of β_v and $\mathcal{T}_v$, and the last inequality follows by the conditions $|T| < t$ for $T \in \mathcal{T}$.

Putting together (3),(4) and (5) we obtain $t\alpha - |V_1| \leq (m-t+1)(t-1)\beta$, or equivalently that

$$\alpha \leq \frac{|V_1|}{t} + \frac{t-1}{t}(m-t+1)\beta. \tag{6}$$

If $|V_1| \leq m - t + 1$, then

$$\frac{|V_1|}{t} + \frac{t-1}{t}(m-t+1)\beta \leq (m-t+1)\beta,$$

and hence $\alpha \leq (m-t+1)\beta$ by (6). On the other hand, if $|V_1| > m - t + 1$, then for each set $S \in \mathcal{S}$ we have $|S \cap V_1| \geq |S| - |V_2| \geq t - |V_2| > 1$. Now by the definition of the set V_1 we obtain $\alpha \leq |V_1|/(t - |V_2|) = (m - |V_2|)/(t - |V_2|) \leq m - t + 1 \leq (m-t+1)\beta$. □

Proof of Theorem 1. Assume without loss of generality that $|\mathcal{M}_t| \geq 2$, for otherwise (1) readily follows from the assumption of the theorem that $|\mathcal{I}_t| \geq 1$. Let us recall that to any subset $C \subseteq \mathcal{C}$ of the columns we have associated the subset $R(C)$ of those rows $r \in \mathcal{R}$ for which $A(r,c) = 1$ for every column $c \in C$. Thus, by definition we have $R(C) = \bigcap_{y \in C} R(\{y\})$, implying

$$R(C' \cup C'') = R(C') \cap R(C'') \quad \text{for all} \quad C', C'' \subseteq \mathcal{C}. \tag{7}$$

In its turn, (7) implies that the mapping $C \mapsto R(C)$ is anti-monotone, i.e. $R(C') \supseteq R(C'')$ whenever $C' \subseteq C''$. Furthermore, $|R(F)| \geq t$ for every maximal t-frequent set $F \in \mathcal{M}_t$, while $|R(U)| < t$ for every minimal t-infrequent set $U \in \mathcal{I}_t$. It is also easy to see that the restriction of the above mapping on $\mathcal{M}_t$ is injective, i.e. $R(F') \neq R(F'')$ for any two distinct maximal t-frequents sets of columns $F', F'' \in \mathcal{M}_t$. If $F', F'' \in \mathcal{M}_t$ then their union $F' \cup F''$ is not t-frequent, and hence there exists a minimal t-infrequent set $U \in \mathcal{I}_t$, for which $R(F') \cap R(F'') = R(F' \cup F'') \subseteq R(U)$. Thus, the families $\mathcal{S} = \{R(F) \mid F \in \mathcal{M}_t\}$ and $\mathcal{T} = \{R(U) \mid U \in \mathcal{I}_t\}$ satisfy the conditions of Lemma 1 with $V = \mathcal{R}$, which implies the inequality $|\mathcal{S}| \leq (m-t+1)|\mathcal{T}|$. Since the mapping $C \mapsto R(C)$ is a one-to one correspondence between $\mathcal{M}_t$ and $\mathcal{S}$, we have $|\mathcal{S}| = |\mathcal{M}_t|$. Now (1) follows from the trivial inequality $|\mathcal{T}| \leq |\mathcal{I}_t|$. □

Proof of Theorem 2. We reduce our problem from the following well-known NP-complete problem: Given a graph $G = (V, E)$ and an integer threshold t, determine if G contains an independent vertex set of size at least t. Let us first substitute every vertex $v \in V$ of G by two new vertices v' and v'' connected by an edge, i.e., consider the graph $G' = (V', E')$, where $V' = \{v', v'' \mid v \in V\}$ and $E' = \{\{(v', v'') \mid v \in V\} \cup \{(v', u'), (v', u''), (v'', v'), (v'', u'') \mid (u, v) \in E\}$. Clearly, $G' = (V', E')$ has an independent set of size t if and only if G has one, moreover, if G' has one, then it has at least 2^t.

Let us now associate a matrix A to G' as follows. Let $\mathcal{C} = V'$ be the set of columns of the matrix A. To every edge $(v, w) \in E'$ we assign $t - 2$ identical rows in A containing 0 in the columns v and w, and 1 in all other columns. Furthermore, to every vertex $v \in V'$ we assign one row containing 0 in the column v and 1 in all other columns. Thus, A has $m = (t-2)|E'| + |V'| = t|V| + 4(t-2)|E|$ rows, and $n = |V'| = 2|V|$ columns.

Clearly, for every edge $e = (v, w) \in E'$ the set $C_e = \mathcal{C} \setminus \{v, w\}$ is a maximal t-frequent set of A. Let $\mathcal{S} = \{C_e \mid e \in E'\}$. We claim that $\mathcal{S} \neq \mathcal{M}_t$ for this matrix, if and only if there exists an independent set I of size $|I| \geq t$ in the graph G'.

To see this claim, let assume first that $I \subseteq V'$ is an independent set of G', $|I| \geq t$. Then $R(V' \setminus I)$ contains all rows corresponding to vertices $v \in I$, and hence $|R(V' \setminus I)| \geq |I| \geq t$. Since I does not contain an edge of the graph, the set $C = V' \setminus I$ is not a subset of the member of $\mathcal{S}$, and thus it is contained by a maximal t-frequent set of the matrix A, which does not belong to $\mathcal{S}$.

For the other direction, let us assume now that $C \subseteq \mathcal{C} = V'$ is a maximal t-frequent set of A, not contained by members of $\mathcal{S}$. This latter implies that $I = V' \setminus C$ does not contain an edge of G', i.e. that I is an independent set of G'. This also implies that $R(C)$ cannot contain any of the rows corresponding to an edge of G', and hence $|R(C)| = |V' \setminus C| = |I|$. Thus, $|I| \geq t$ follows by our assumption that C is a t-frequent set, i.e. I is an independent set of size at least t.

Let us recall finally that the maximum independent set problem remains NP-hard, even if the input is restricted to cubic planar graphs (see e.g. [14]), i.e. we can assume $|E| = O(|V|)$. Therefore, we have $|\mathcal{S}| = |E'| = |V| + 4|E| = O(|V|)$, and either we have $\mathcal{S} = \mathcal{M}_t$, or $|\mathcal{M}_t| \geq |\mathcal{S}| + 2^t$. Since we can assume without loss of generality that $t = \Theta(|V|)$, the theorem follows. □

References

[1] R. Agrawal, T. Imielinski and A. Swami. Mining associations between sets of items in massive databases. In: *Proceedings of the 1993 ACM-SIGMOD International Conference on Management of Data*, pp. 207-216.

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A. I. Verkamo, Fast discovery of association rules, In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy eds., *Advances in Knowledge Discovery and Data Mining*, 307-328, AAAI Press, Menlo Park, California, 1996.

[3] R. Agrawal and R. Srikant. Mining sequential patterns. In: *Proceedings of the 11th International Conference on Data Engineering, 1995*, pp.3-14.

[4] R.J. Bayardo, Efficiently mining long patterns from databases. In: *Proceedings of the 1998 ACM-SIGMOD International Conference on Management of Data*, pp. 85-93.

[5] J. C. Bioch and T. Ibaraki, Complexity of identification and dualization of positive Boolean functions, *Information and Computation* 123 (1995) 50-63.

[6] E. Boros, V. Gurvich, L. Khachiyan and K.Makino, Generating partial and multiple transversals of a hypergraph. In: *Proceedings of the 27th International Colloquium on Automata, Languages and Programming (ICALP)*, (U. Montanari, J.D.P. Rolim and E. Welzl, eds.) Lecture Notes in Computer Science **1853** pp. 588-599, (Springer Verlag, Berlin, Heidelberg, New York, 2000).

[7] E. Boros, V. Gurvich, L. Khachiyan and K.Makino, Generating Weighted Transversals of a Hypergraph, DIMACS Technical Report 00-17, Rutgers University, 2000. (http://dimacs.rutgers.edu/TechnicalReports/2000.html)

[8] S. Brin, R. Motwani, and C. Silverstein. Beyond market basket: Generaliing association rules to correlations. In: *Proceedings of the 1997 ACM-SIGMOD Conference on Management of Data*, pp. 265-276.

[9] S. Brin, R. Motwani, J. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In: *Proceedings of the 1997 ACM-SIGMOD Conference on Management of Data*, pp. 255-264.

[10] G. Dong and J. Li. Efficient mining of emerging patterns. In: *Proceeding of the 1999 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 43-52.

[11] T. Eiter and G. Gottlob, Identifying the minimal transversals of a hypergraph and related problems, *SIAM Journal on Computing*, 24 (1995) 1278-1304.

[12] J. Han, J. Pei, and Y. Yin, Mining frequent patterns without candidate generation, In: *Proceedings of the 2000 ACM-SIGMOD Conference on Management of Data*, pp. 1-12.

[13] M. L. Fredman and L. Khachiyan, On the complexity of dualization of monotone disjunctive normal forms. *J. Algorithms*, 21 (1996) 618-628.

[14] M. R. Garey and D. S. Johnson, *Computers and Intractability*, Freeman, New York, 1979.

[15] D. Gunopulos, R. Khardon, H. Mannila, and H. Toivonen, Data mining, hypergraph transversals and machine learning. In: *Proceedings of the 16th ACM-SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, (1997) pp. 12-15.

[16] V. Gurvich and L. Khachiyan, On generating the irredundant conjunctive and disjunctive normal forms of monotone Boolean functions, *Discrete Applied Mathematics*, 1996-97, issue 1-3, (1999) 363-373.

[17] D. Lin and Z.M. Kedem. Pincer-search: a new algorithm for discovering the maximum frequent set. In: *Proceedings of the Sixth European Conference on Extending Database Technology*, to appear.

[18] K. Makino and T. Ibaraki, Inner-core and outer-core functions of partially defined Boolean functions, *Discrete Applied Mathematics*, 1996-97, issue 1-3 (1999), 307-326.

[19] H. Mannila and H. Toivonen, Multiple uses of frequent sets and condensed representations. In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, (1996) pp. 189-194.

[20] H. Mannila and H. Toivonen, Levelwise search and borders of theories in knowledge discovery. Series of Publications C C-1997-8, University of Helsinki, Department of Computer Science (1997).

[21] H. Mannila, H. Toivonen, and A. I. Verkamo. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1 (1997), 259-289.

[22] R. H. Sloan, K. Takata, G. Turan, On frequent sets of Boolean matrices, *Annals of Mathematics and Artificial Intelligence* 24 (1998) 1-4.