Support Approximations Using Bonferroni-type Inequalities

Szymon Jaroszewicz, Dan A. Simovici

University of Massachusetts at Boston,

Department of Computer Science,

Boston, Massachusetts 02125, USA

Inspiration

H. Manilla et al. [1996,2001]: Use frequent itemsets to get support of arbitrary queries, e.g.:

$$\mathrm{supp}(\bar{A}\bar{B}) = 1 - \mathrm{supp}(A) - \mathrm{supp}(B) + \mathrm{supp}(AB)$$

(inclusion-exclusion principle)

Questions:

- How to obtain such a formula for an arbitrary query?
- Guarantee of accuracy if some supports are unknown?

Bonferroni inequalities

$$\mathrm{supp}(\bar{A}_1 \wedge \bar{A}_2 \wedge \ldots \wedge \bar{A}_n) \leq \sum_{k=0}^{2m} S_k$$

$$\mathrm{supp}(\bar{A}_1 \wedge \bar{A}_2 \wedge \ldots \wedge \bar{A}_n) \geq \sum_{k=0}^{2m+1} S_k$$

where

$$S_k = \sum_{i_1 \leq \ldots \leq i_k} \mathrm{supp}(A_{i_1} A_{i_2} \ldots A_{i_k})$$

Previous results

DM&DM Workshop, SIAM Conference on Datamining 2002:

1. Derive Bonferroni inequalities for exclusive-or:

$$\begin{aligned}
\mathrm{supp}(p_1 \oplus \ldots \oplus p_m) &\leq \sum_{k=1}^{2t+1} (-2)^{k-1} \sum_{i_1 \leq \ldots \leq i_k} \mathrm{supp}(p_{i_1} \wedge \ldots \wedge p_{i_k}), \\
\mathrm{supp}(p_1 \oplus \ldots \oplus p_m) &\geq \sum_{k=1}^{2t} (-2)^{k-1} \sum_{i_1 \leq \ldots \leq i_k} \mathrm{supp}(p_{i_1} \wedge \ldots \wedge p_{i_k}),
\end{aligned}$$

where $p_1, \ldots, p_k$ are arbitrary Boolean expressions.

Previous results

2. Express support of arbitrary queries using supports of itemsets.
3. Give bounds (not always tight) for support of arbitrary queries.
4. Prove that support of some queries cannot be approximated (based on supports of itemsets)

   Support of *parity function* $A_1 \oplus A_2 \oplus \ldots \oplus A_n$ can't be approximated even if support of a single itemset is unknown.

Parity function can't be approx-imated

$H = A_1 A_2 \ldots A_n$
$Q = A_1 \oplus A_2 \oplus \ldots \oplus A_n$

Let $\rho_{even} = \{t \in \text{Dom}(H) : n_1(t) \text{ is even}\}$
Let $\rho_{odd} = \{t \in \text{Dom}(H) : n_1(t) \text{ is odd}\}$
Then for every $I \subset H$

$$\text{supp}_{\rho_{even}}(I) = \text{supp}_{\rho_{odd}}(I)$$

but $\text{supp}_{\rho_{even}}(Q) = 0\%$ and $\text{supp}_{\rho_{odd}}(Q) = 100\%$
$\text{supp}(H)$ unknown $\Rightarrow$ $\text{supp}(Q)$ can be anywhere between 0% and 100%.

This paper

1. Parity function — not a typical query.
   Can we get useful bounds for more common queries?
2. Bonferroni inequalities require supports of all itemsets of size k — not typical for DM

Trivial bounds for unknown itemset supports

$$\mathrm{supp}(I) \geq 0$$

$$\mathrm{supp}(I) \leq \textsf{minsupp}$$

$$\mathrm{supp}(I) \leq \min\{\mathrm{supp}(J) : J \subset I\}$$

Bonferroni-type bounds for unknown itemset supports

The following inequalities hold for any natural number t:

$$\sum_{k=0}^{2t+1}(-1)^{k+1}\binom{m-k-1}{2t+1-k}S_k \le \operatorname{supp}(A_1A_2\ldots A_m) \le \sum_{k=0}^{2t}(-1)^{k}\binom{m-k-1}{2t-k}S_k$$

Example

We have

$$\mathrm{supp}(ABC) \geq \mathrm{supp}(A) + \mathrm{supp}(B) + \mathrm{supp}(C) - 2$$

and

$$\begin{aligned}\mathrm{supp}(ABC) \leq\ & 1 - \mathrm{supp}(A) - \mathrm{supp}(B) - \mathrm{supp}(C) \\ & + \mathrm{supp}(AB) + \mathrm{supp}(AC) + \mathrm{supp}(BC)\end{aligned}$$

Conjunctions with negated attributes

$$\sum_{k=0}^{2t+1}(-1)^k \sum_{r<i_1<\ldots<i_k\leq m} \mathrm{supp}(A_1\ldots A_r A_{i_1}\ldots A_{i_k})$$

$$\leq \mathrm{supp}(A_1\ldots A_r \bar{A}_{r+1}\ldots \bar{A}_m) \leq$$

$$\sum_{k=0}^{2t}(-1)^k \sum_{r<i_1<\ldots<i_k\leq m} \mathrm{supp}(A_1\ldots A_r A_{i_1}\ldots A_{i_k})$$

Example

We have

$$\operatorname{supp}(A\bar{B}\bar{C}) \geq \operatorname{supp}(A) - \operatorname{supp}(AB) - \operatorname{supp}(AC)$$

and

$$\begin{aligned} \operatorname{supp}(A\bar{B}\bar{C}) \;\; \leq \;\; & \operatorname{supp}(A) - \operatorname{supp}(AB) - \operatorname{supp}(AC) \\ & + \operatorname{supp}(ABC) \end{aligned}$$

Recursive application of Bonferroni inequalities

After finding frequent itemsets, not all itemsets of given size k have known supports $\Rightarrow$ Bonferroni-type inequalities cannot be applied directly.

Our approach: If for a given k some supports are unknown we apply the inequalities ‘recursively’.

Example

Table over ABC

A	B	C	Frequency
0	0	0	0
0	0	1	0
0	1	0	0.10
0	1	1	0.25
1	0	0	0.10
1	0	1	0.25
1	1	0	0.05
1	1	1	0.25

Frequent itemsets

(minsupp $= 0.35$)

Itemset	support
A	0.65
B	0.65
C	0.75
AC	0.50
BC	0.50

Example ctd.

$$\begin{aligned}\mathrm{supp}(ABC) &\leq 1 - \mathrm{supp}(A) - \mathrm{supp}(B) - \mathrm{supp}(C) \\ &+ \mathrm{supp}(AB) + \mathrm{supp}(AC) + \mathrm{supp}(BC)\end{aligned}$$

Estimate $\mathrm{supp}(AB)$ recursively.
Trivial bound: $\mathrm{supp}(AB) \leq$ minsupp $= 0.35$
giving

$$\mathrm{supp}(ABC) \leq 0.30$$

better than any trivial bound.

Experimental results

Method is useful for dense datasets and itemsets with high support

- UCI `mushroom` dataset
- Elderly people census data

Experimental results

Mine frequent itemsets of size ≤ 2
Try to estimate if larger itemsets are frequent

size	minsupp :	18%	25%	37%	43%	49%	61%
3	Frequent	1761	893	308	152	70	23
	Estimated	19.59%	27.32%	41.23%	56.58%	77.14%	82.61%
4	Frequent	4379	1769	368	147	48	16
	Estimated	6.81%	11.42%	23.10%	36.05%	64.58%	62.50%

Conclusion: large percentage of itemsets with high supports can be discovered based on supports of their subsets (without any data access).

Percentage of itemsets with non-trivial bounds

census dataset, 9% minimum support
Mine frequent itemsets of size ≤ 2, estimate larger ones.

itemset size	3	4	5	6
average upper bound	0.235	0.223	0.225	0.231
average lower bound	0.063	0.018	0.002	0
itemsets w. nontrivial bounds	48.55%	16.79%	3.40%	0.14%

Conclusion: Bonferroni-type inequalities provide nontrivial results

Itemsets with negations

census data with 1.8% Minimum Support
supports of frequent itemsets of size ≤ 2 known
2 negated items in each itemset

itemset size	3	4	5	6
avg interval width	0.0405	0.082	0.067	0.039
average upper bound	0.171	0.121	0.069	0.039
average lower bound	0.131	0.0387	0.001	2.73e-05

Conclusion: Fairly tight bounds can be obtained

Conclusions

- Bonferroni type inequalities provide useful bounds in dense datasets, for itemsets with large support
- For sparse databases / itemsets with low support Bonferroni-type inequalities seem not to produce useful results.

Future research

- Bounds for other types of queries
- Other methods for the case when not all itemsets of given size are known
- Investigate more sophisticated variants of Bonferroni-type inequalities