Computational Complexity of Queries Based on Itemsets

Nikolaj Tatti

HIIT Basic Research Unit, Laboratory of Computer and Information Science,
Helsinki University of Technology, Finland

Abstract

We investigate determining the exact bounds of the frequencies of conjunctions based on frequent sets. Our scenario is an important special case of some general probabilistic logic problems that are known to be intractable. We show that despite the limitations our problems are also intractable, namely, we show that checking whether the maximal consistent frequency of a query is larger than a given threshold is **NP**-complete and that evaluating the Maximum Entropy estimate of a query is **PP**-hard. We also prove that checking consistency is **NP**-complete.

Key words: Computational Complexity, Data Mining, Itemset

1 Introduction

Assume that we have two events, say a and b. Assume further that their probabilities are $P(a) = 0.6$ and $P(b) = 0.5$. What can we say about the probability of $a \wedge b$? We know that the probability must lie within $I = [0.1, 0.5]$. This interval is tight: For each $f \in I$ there is a distribution having f as a probability of $a \wedge b$. Also note that the Maximum Entropy estimate in this case is $0.6 \times 0.5 = 0.3$.

A more complicated example would be the following: Assume three events a_1, a_2, and a_3. Assume that we know $P(a_1)$, $P(a_2)$, $P(a_3)$, $P(a_1 \wedge a_2)$ and $P(a_1 \wedge a_3)$. What can we say about $P(a_1 \wedge a_2 \wedge a_3)$?

Let us make these examples more general: A conjunctive query is a boolean formula having the form $a_1 \wedge a_2 \wedge \ldots \wedge a_L$. Assume that we are given a set $\mathcal{F}$ of conjunctive queries along with their probabilities. Assume also that we are given a conjunctive query B not belonging to $\mathcal{F}$. What can we tell about the probability of this query? We know that the possible probabilities of the query

B correspond to some interval. In the paper we show that checking whether the right side of this interval is larger than some threshold is **NP**-complete. We also show that estimating the probability of the query B using Maximum Entropy is **PP**-hard.

In the paper we adopt the terminology used in data mining of 0–1 data: Conjunctive queries are represented by sets of items called itemsets and the probabilities of conjunctive queries are called itemset frequencies.

Our problems are special cases of much more general problems (see Section 6 for detailed comparison). These general problems are well-studied and they are all (at least) **NP**-hard. The difference is that in our work we concentrate on studying antimonotonic families of itemsets. We should point out that antimonotonic families are important since they tend to arise frequently in practice, for example, in mining of frequent itemsets [1,2]. A similar technique is used in [7] to prove that inference of Belief Networks is **NP**-hard. The result of [7] is essentially Theorem 6 (in this paper) though it is in a different context. The general boolean query scenario is reduced to Linear Programming in [10]. A method worth mentioning is introduced in [15] where the authors estimate the frequencies using Maximum Entropy.

2 Preliminaries

In this section we give basic definitions used in mining of 0–1 data.

By a *binary data set* we mean a collection of binary vectors of length K sampled from some distribution. We define a *sample space* $\Omega = \{0,1\}^K$ to be the collection of all possible binary vectors of length K. From now on Ω will always denote the sample space, K will denote the dimension of binary vectors. Any distribution given in this paper will be defined on Ω.

It is custom to assign an attribute to each dimension of Ω. Thus, when we speak of a_i we mean the ith dimension. The set of all attributes is $A = \{a_1, \ldots, a_K\}$. An *itemset* is a subset of A. Let $B = \{a_{i_1}, \ldots, a_{i_L}\}$ be an itemset. We often use a condensed notation $B = a_{i_1} \cdots a_{i_L}$. A family of itemsets is called *antimonotonic* if all the subsets of any member are also included.

Let p be a distribution defined on Ω. We use the following notation: Let $B = a_{i_1} \cdots a_{i_L}$ be an itemset and let t be a binary vector of length L. Then we shorten the notation $p(a_{i_1} = t_1, \ldots, a_{i_L} = t_L)$ by $p(B = t)$. By $p(B = 1)$ we mean $p(B = t)$, where t contains only ones. The probability $p(B = 1)$ is called the *frequency* of B.

Assume a family $\{B_1, \ldots, B_N\}$ of itemsets and a vector θ of length N. We say that a distribution p *satisfies* the frequencies if $\theta_i = p(B_i = 1)$ for $i = 1, \ldots, N$. We say that these frequencies are *consistent* if there is a distribution satisfying them.

3 Maximal Frequency Query is NP-complete

Assume that we want to find the frequency for an itemset B based on some known family $\mathcal{F}$ of itemsets. We know that generally the frequency for B is not unique: There may be distributions that produce different frequencies for B but have the same frequencies of $\mathcal{F}$. The set of all the consistent frequencies of B is an interval [4]. In this section we focus on finding one side of this interval:

Problem 1 (MaxQuery) *Assume that we are given an antimonotonic family $\mathcal{F}$ having N members along with rational and consistent frequencies θ. Find the maximal frequency for a given itemset B that can be produced by a distribution satisfying the frequencies θ.*

In other words, we ask ourselves that, if we know the frequencies θ, then what is the largest consistent frequency for B. Note that the maximal frequency always exists since the frequencies θ are required to be consistent. Our goal in this section is to show that in general this problem is intractable. First let us give an example where the solution can be easily obtained.

Example 1 *Assume that a family $\mathcal{F}$ contains only the itemsets of size one. Then the frequency θ_{a_i} is the mean of the attribute a_i. The maximal frequency for an itemset $B = b_1 b_2 \cdots b_M$ is $\min \{\theta_{b_i} \mid i = 1, \ldots, M\}$.*

We know that MAXQUERY can be solved by using Linear Programming [4] though the resulting program contains an exponential number of variables. This reduction along with some results from Linear Programming theory [14] has important consequences: There is a distribution, say q, producing the maximal frequency for B and having at most $N + 1$ non-zero entries. Also, q has rational entries, and if L is the number of bits needed to specify the denominator of an element of the frequency vector θ, then the number of bits needed to specify the denominator of an entry of q is $\log_2 \left((N+1)^3 2^{NL} \right) \in O(NL)$. We call such a distribution *canonical*.

Since **NP** is defined for yes/no problems we need the decision version of MAX-QUERY:

Problem 2 (MaxQueryDec) *Assume that we are given an antimonotonic family $\mathcal{F}$ having N members along with rational and consistent frequencies θ.*

Given an itemset B and a rational threshold b is there a distribution satisfying the frequencies θ such that the frequency of B is larger than b?

The relation between MAXQUERY and MAXQUERYDEC is the following: Assume that we can solve MAXQUERY in polynomial time, then we can clearly solve MAXQUERYDEC in polynomial time. Assume now that we can solve MAXQUERYDEC in polynomial time. Let f be the solution of MAXQUERY. We can find f using MAXQUERYDEC and dichotomous search. We know that f is a rational number between 0 and 1 and that the denominator of f can be expressed using $O(NL)$ bits. Thus the number of required search steps is $O(NL)$.

Theorem 2 MAXQUERYDEC *is in* **NP**.

PROOF. Let q be a canonical distribution for MAXQUERY. We can represent this distribution in polynomial space, and hence we can use it as a certificate. To check the certificate we need to check that q is a real distribution, that it satisfies the frequencies and that its frequency for B is larger than the threshold b.

Our next step is to reduce 3SAT to MAXQUERYDEC. In order to do that we need the following lemma:

Lemma 3 *Assume that two distributions p and q satisfy the frequencies θ of an antimonotonic family $\mathcal{F}$ of itemsets. Let $C \in \mathcal{F}$. Then $p(C = t) = q(C = t)$ for any binary vector t.*

PROOF. Fix $C = \{c_1, \ldots, c_N\}$ and t. Let $U = \{c_i \in C \mid t_i = 1\}$ and let $W = C - U$. Denote the elements of W by w_i. Let $p(U = 1, \bigvee_i w_i = 1)$ be the probability of U being 1 and at least one of w_i being 1. We see that

$$p(C = t) = p(U = 1, W = 0) = p(U = 1) - p(U = 1, \bigvee w_i = 1). \quad (1)$$

Let $\mathcal{H} = \{H \subseteq W \mid H \neq \emptyset\}$ be the collection of non-empty subsets of W. We can express the last term of Eq. 1 by using the inclusion-exclusion principle

$$p(U = 1, \bigvee w_i = 1) = \sum_{H \in \mathcal{H}} (-1)^{|H|+1} p(U = 1, H = 1). \quad (2)$$

By combining Eqs. 1 and 2 we have expressed $p(C = t)$ as a linear combination of terms having the form $p(B = 1)$ where $B \subseteq C$. Antimonotonicity implies that all these frequencies are included in θ. This makes $p(C = t)$ unique and the lemma follows.

Theorem 4 3SAT *is polynomial-time reducible to* MaxQueryDec.

PROOF. Let R be an instance of 3SAT having L variables and M clauses. We set the dimension of the sample space to be $K = L + M$. The first L items correspond to the variables of R and the last M items correspond to the clauses. We use the following notation: Let t be a truth assignment and let C_i be a clause, then $C_i(t)$ is a function resulting 1, if C_i is satisfied by t, and 0 otherwise. We denote the first L items by v_i and the last M items by c_i. We also set $V = \{v_1, \ldots, v_L\}$ and $W = \{c_1, \ldots, c_M\}$.

We will now define an antimonotonic family $\mathcal{F}$ of itemsets. Let C_i be some clause and let c_i be its corresponding item. Assume that the items corresponding to the variables in C_i are v_1, v_2, and v_3. We add an itemset $v_1 v_2 v_3 c_i$ to the family $\mathcal{F}$ along with its subsets. We repeat this procedure to each clause in R. The resulting family $\mathcal{F}$ contains $16M$ members at maximum.

The following step is to define the frequencies θ. In order to do this we define a distribution p over the attributes to be

$$
p(V = t, W = u) = \begin{cases} 2^{-L} & \text{if for all } i \text{ we have } u_i = C_i(t) \\ 0 & \text{otherwise.} \end{cases}
$$

That is, the first L items are distributed uniformly and the values of the last M items are set to correspond to the truth values of the clauses.

We define the frequencies $\theta_i = p(F_i = 1)$, where $F_i \in \mathcal{F}$. We note that the frequencies are rational and consistent. There is a closed formula for evaluating these frequencies. For example, assume that we have a clause $C_1 \equiv (v_1 \vee v_2 \vee v_3)$. The frequency of the itemset $v_1 v_2 v_3 c_1$ is then

$$
\sum_{t,u} p(V = t, W = u) = \sum_{t, u_i = C_i(t)} p(V = t, W = u) = 2^{L-3} 2^{-L} = \frac{1}{8},
$$

where in the first summation t ranges over truth assignments such that $t_1 = t_2 = t_3 = 1$ and u ranges over binary vectors of length M such that $u_1 = 1$. In the second summation t ranges similarly as in the first summation and u is now set to correspond to the clauses. The frequencies for the other members of $\mathcal{F}$ can be deduced in a similar way. Thus we can obtain the frequencies θ in polynomial time.

Let f be the maximal frequency for the itemset W. We claim that the formula R is satisfiable if and only if $f > 0$.

Assume that R is satisfiable by a truth assignment, then we have

$$f = p(W = 1) \geq p(V = t, W = 1) = 2^{-L} > 0.$$

Assume now that there is a distribution q satisfying the frequencies and producing a positive frequency for W. Let t be a truth assignment not satisfying the formula, that is, there is a clause, say $C_1 = (v_1 \vee v_2 \vee v_3)$, that is not satisfied. Define $G = v_1 v_2 v_3$ and $u = [t_1, t_2, t_3]$. Lemma 3 implies that $q(V = t, W = 1) \leq q(G = u, c_1 = 1) = p(G = u, c_1 = 1) = 0$. By reversing this property we get the following: If t is such that

$$q(V = t, W = 1) > 0 \qquad (3)$$

holds, then t must satisfy R.

By the assumption $q(W = 1) > 0$ so there exists a truth assignment t such that Eq. 3 holds. Thus R is satisfiable. The reduction is complete if we set the query $B = W$ and the threshold $b = 0$.

Example 5 *Consider the formula* $(v_1 \vee v_2) \wedge (\neg v_2 \vee v_3)$. *We have two clauses,* C_1 *and* C_2, *and three variables,* v_1, v_2, *and* v_3. *The itemset family along with its frequencies (given in parenthesises) is*

$$\mathcal{F} = \left\{ \begin{array}{l} \emptyset\,(1)\,, v_1\left(\frac{1}{2}\right), v_2\left(\frac{1}{2}\right), v_3\left(\frac{1}{2}\right), v_1 v_2\left(\frac{1}{4}\right), v_2 v_3\left(\frac{1}{4}\right), \\ c_1\left(\frac{3}{4}\right), v_1 c_1\left(\frac{1}{2}\right), v_2 c_1\left(\frac{1}{2}\right), v_1 v_2 c_1\left(\frac{1}{4}\right), \\ c_2\left(\frac{3}{4}\right), v_2 c_2\left(\frac{1}{4}\right), v_3 c_2\left(\frac{1}{2}\right), v_2 v_3 c_2\left(\frac{1}{4}\right) \end{array} \right\}.$$

The maximal frequency of $c_1 c_2$ *for this setup (solved by linear programming) is* $\frac{1}{2}$. *Clearly, the formula is satisfiable.*

4   MaxEnt Frequency Query is PP-hard

In the previous section we showed that searching for the maximal frequencies is a very hard problem. The maximal frequencies, however, are not so useful if our goal is to estimate boolean queries from a given set of itemsets. A much more useful approach is to use Maximum Entropy approach. Given a distribution p defined on Ω, the *entropy* of p is $\mathcal{E}(p) = -\sum_{\omega \in \Omega} p(\omega) \log(p(\omega))$. It is custom to define $0 \log(0) = 0$ so that $\mathcal{E}(p)$ is always defined.

Problem 3 (EntrQuery) *Assume that we are given an antimonotonic family* $\mathcal{F}$ *having* N *members along with rational and consistent frequencies* θ. *Find*

a frequency for a given itemset B produced by the distribution p satisfying the frequencies θ and maximising the entropy $\mathcal{E}(p)$.

It has been empirically shown that ENTRQUERY results in a good approximation [15].

Again we need a decision version of the problem:

Problem 4 (EntrQueryDec) *Assume that we are given an antimonotonic family $\mathcal{F}$ having N members along with rational and consistent frequencies θ. Let f be a frequency for a given itemset B produced by a distribution satisfying the frequencies θ and maximising entropy. Is f larger than a given rational threshold b?*

The following theorem shows that ENTRQUERYDEC is **NP**-hard.

Theorem 6 3SAT *is polynomial-time reducible to* ENTRQUERYDEC.

PROOF. Let R be an instance of 3SAT. Let $\mathcal{F}$, θ, V and B be the same as in the proof of Theorem 4. Let $\mathbb{P}$ be the set of distributions satisfying the frequencies θ. Let $q \in \mathbb{P}$. A marginal distribution q_V is obtained from q by keeping only the items included in V. The distribution q has the following property: The items corresponding to the clauses are completely determined by the items corresponding to the variables. This implies that the entropy of $\mathcal{E}(q) = \mathcal{E}(q_V)$ [11, Theorem 4.2].

Let $\hat{q} \in \mathbb{P}$ be the distribution maximising the entropy. Let $p \in \mathbb{P}$ be the distribution defined in the proof of Theorem 4. Note that $\mathcal{E}(\hat{q}_V) = \mathcal{E}(\hat{q}) \geq \mathcal{E}(p) = \mathcal{E}(p_V)$. We know that there is no distribution that has larger entropy than the uniform distribution [11, Theorem 3.1]. Since p_V is uniform, we must have $\mathcal{E}(\hat{q}_V) = \mathcal{E}(p_V)$. Hence $\mathcal{E}(\hat{q}) = \mathcal{E}(p)$. We also know that the distribution maximising entropy is unique [8, Theorem 3.1]. This implies that $\hat{q} = p$. To complete the proof we note that p produces a positive frequency for B if and only if R is satisfiable.

A problem P is in **PP** if there is a machine such that an input x is a yes-instance of P iff more than half of the computation paths end up accepting [13]. The class **PP** is (believed to be) larger than **NP**. We can show that ENTRQUERYDEC is **PP**-hard: In the proof the frequency of B is exactly the number of satisfying assignments divided by 2^{-L}. Hence, if we set the threshold $b = 2^{-L/2}$, the instance will be in ENTRQUERYDEC iff the square root of the number of assignments satisfy the given 3SAT formula. This problem is known to be **PP**-complete [3].

5   Checking Consistency is NP-complete

So far we have assumed that the itemset frequencies given in our problems are consistent. Let us remove this constraint and consider the following problem.

Problem 5 (Consistent) *Assume that we are given an antimonotonic family $\mathcal{F}$ having N members along with rational frequencies θ. Are the frequencies θ consistent?*

The following theorem proves that CONSISTENT is a very hard problem.

Theorem 7 CONSISTENT *is **NP**-complete.*

PROOF. First, we need to show that CONSISTENT is in **NP**. We know from Linear Programming theory that if the frequencies are valid then there is a canonical distribution satisfying the frequencies. This is our certificate and thus CONSISTENT is in **NP**.

We now prove that 3SAT is polynomial-time reducible to CONSISTENT. We use the same construction as in the proof of Theorem 4 with some additions: We add one special attribute, say c_0, to the set of attributes. We add an itemset c_0 to $\mathcal{F}$, and we also add itemsets having the form $c_0 c_i$ to $\mathcal{F}$. The frequencies for the new itemsets are set to be 2^{-L}, where L is the number of variables appearing in the 3SAT instance R.

Assume that R is satisfiable by a truth assignment t. We define a distribution q by extending the distribution p to c_0. The extension is done such that c_0 is 1 iff $V = t$. Clearly, q satisfies the frequencies.

To prove the other direction, assume that there exists a distribution, say q, that satisfies the frequencies. To prove that R is satisfiable we must prove that $q(W = 1) > 0$. Select two attributes, say c_1 and c_2. Note that $q(c_0 = 1, c_1 = 0) = 0$ and $q(c_0 = 1, c_2 = 0) = 0$. This implies that $q(c_0 = 1) = q(c_0 = 1, c_1 = 1, c_2 = 1)$. We can prove in an iterative fashion that

$$q(W = 1) \geq q(c_0 = 1, W = 1) = q(c_0 = 1) = 2^{-L}.$$

This proves the result.

6   Connections to Related Work

An **NP**-complete problem called FREQSAT introduced in [5,6] is a generalisation of CONSISTENT — in FREQSAT we are allowed to have non-antimonotonic

families and inequality constraints. We can transform MAXQUERYDEC into FREQSAT by changing the query into an inequality constraint. We should also point out that the proof of **NP**-hardness of FREQSAT given in [5] is (although not explicitly mentioned) actually a valid proof for CONSISTENT.

An even more general scenario is introduced in [12] in which we are allowed to have conditional first-order logic sentences as constraints/queries. This scenario can be emulated by itemsets [6]. Also, a famous problem called PSAT in which we are given a CNF-formula, a frequency for each clause, and we are asked whether there is a distribution satisfying the frequencies is known to be **NP**-complete [9].

7 Conclusions

In this paper we studied certain boolean query problems. Our problems were specialised (but frequently occurring and thus important) problems of much general scenarios and we showed that despite the limitations our problems remained intractable. The crux of the paper lies within the construction in the proof of Theorem 4.

There are some open problems: For example, what is the exact complexity of MAXQUERY? Is it **FNP**-complete or $\mathbf{FP^{NP}}$-complete? Also, what is the complexity of the opposite problem MINQUERY? In addition, it is worthwhile to study the conditions under which the boolean query problems can be solved efficiently.

References

[1] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., 26–28 1993.

[2] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and Aino Inkeri Verkamo. Fast discovery of association rules. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI Press/The MIT Press, 1996.

[3] Delbert D. Bailey, Victor Dalmau, and Phokion G. Kolaitis. Phase transitions of PP-complete satisfiability problems. In *IJCAI*, pages 183–192, 2001.

[4] Artur Bykowski, Jouni K. Seppänen, and Jaakko Hollmén. Model-independent bounding of the supports of Boolean formulae in binary data. In Pier Luca Lanzi and Rosa Meo, editors, *Database technologies for data mining*. Springer Verlag, 2003.

[5] Toon Calders. *Axiomatization and Deduction Rules for the Frequency of Itemsets*. PhD thesis, University of Antwerp, Belgium, 2003.

[6] Toon Calders. Computational complexity of itemset frequency satisfiability. In *Proceedings of the 23nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database System*, 2004.

[7] Gregory Cooper. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, 42(2–3):393–405, Mar. 1990.

[8] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, Feb. 1975.

[9] George Georgakopoulos, Dimitris Kavvadias, and Christos H. Papadimitriou. Probabilistic satisfiability. *Journal of Complexity*, 4(1):1–11, March 1988.

[10] Theodore Hailperin. Best possible inequalities for the probability of a logical function of events. *The American Mathematical Monthly*, 72(4):343–359, Apr. 1965.

[11] Solomon Kullback. *Information Theory and Statistics*. Dover Publications, Inc., 1968.

[12] Thomas Lukasiewicz. Probabilistic logic programming with conditional constraints. *ACM Transactions on Computational Logic (TOCL)*, 2(3):289–339, July 2001.

[13] Christos Papadimitriou. *Compuitional Complexity*. Addison-Wesley, 1995.

[14] Christos Papadimitriou and Kenneth Steiglitz. *Combinatorial Optimization Algorithms and Complexity*. Dover, 2nd edition, 1998.

[15] Dmitry Pavlov, Heikki Mannila, and Padhraic Smyth. Beyond independence: Probabilistic models for query approximation on binary transaction data. *IEEE Transactions on Knowledge and Data Engineering*, 15(6):1409–1421, 2003.