Available online at www.sciencedirect.com
ScienceDirect
Discrete Applied Mathematics 157 (2009) 1218–1228
DISCRETE APPLIED MATHEMATICS
www.elsevier.com/locate/dam

Ranking hypotheses to minimize the search cost in probabilistic inference models

Peter Damaschke

Department of Computer Science and Engineering, Chalmers University, 41296 Göteborg, Sweden

Received 14 November 2006; received in revised form 3 July 2007; accepted 3 December 2007
Available online 14 February 2008

Abstract

Suppose that we are given n mutually exclusive hypotheses, m mutually exclusive possible observations, the conditional probabilities for each of these observations under each hypothesis, and a method to probe each hypothesis whether it is the true one. We consider the problem of efficient searching for the true (target) hypothesis given a particular observation. Our objective is to minimize the expected search cost for a large number of instances, and for the worst-case distribution of targets. More precisely, we wish to rank the hypotheses so that probing them in the chosen order is optimal in this sense. Costs grow monotonic with the number of probes. While it is straightforward to formulate this problem as a linear program, we can solve it in polynomial time only after a certain reformulation: We introduce mn^2 the so-called rank variables and arrive at another linear program whose solution can be translated afterwards into an optimal mixed strategy of low description complexity: For each observation, at most n rankings, i.e., permutations of hypotheses, appear with positive probabilities. Dimensionality arguments yield further combinatorial bounds. Possible applications of the optimization goal are discussed.
© 2008 Elsevier B.V. All rights reserved.

Keywords: Probabilistic inference; Searching; Ranking; Mixed strategy; Linear programming; Polytopes

1. Problem statement

We study an optimization problem in probabilistic inference. Assume that among n mutually exclusive hypotheses exactly one is true. We make exactly one observation out of m possible observations which are mutually exclusive as well. For every hypothesis h and every observation D we know $P(D|h)$, defined as the conditional probability to observe D if hypothesis h is true. Note that we mean by D a complete description of what we observe, hence the conditional probabilities satisfy $\sum_D P(D|h) = 1$ for every h. The $P(D|h)$ are known from the background knowledge about causal relations, or estimated from statistical material. In contrast to these conditional probabilities, the distribution of targets may be arbitrary and unpredictable.

The problem is, first in a vague formulation, as follows. Given an observation D, a Searcher wants to find the correct hypothesis h, also called the *target hypothesis* or simply the *target*, in a cheap and efficient way. For the moment we assume that each hypothesis can be tested for being the target by some reliable experiment. (We discuss this matter further in Section 2.) Consider two natural settings:

E-mail address: ptr@cs.chalmers.se.

0166-218X/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2007.12.006

- Let $g < n$ be a fixed number. For any D, we are allowed to choose g hypotheses, and we would like to have the target in our selected set. One may think of g as an acceptable number of attempts to identify the target, whereas additional tests are undesirable or come with an extra cost.
- A smoother cost assumption is that all verification experiments have unit cost, and their number is not limited other than by n. Then, for any D, we have to rank the hypotheses and test them in the chosen order. We want to find the target early, i.e., minimize the number of tests. In other words, we want to rank the target low.

The target is unknown (otherwise there is nothing to search for), but the $P(D|h)$ and the observed D may hint to it. Note that the selection or ranking need not be deterministic, that is, Searcher may apply a randomized strategy. We are interested in strategies that perform well on average, on a large number of problem instances but for the given fixed table of conditional probabilities. That is, Searcher is presented many independent cases and wants to minimize his total search costs in the long run. (There is not much to say about the actual performance on any single instance, but for a large number of instances, expected costs turn into average actual costs.) Since we assume nothing about the distribution of targets h, it is natural to optimize the expected search cost in the worst case, i.e., maximized over all h. This corresponds also to the usual game-theoretic framework: An adversary is presenting the instances, and a strategy is sought that gives the best guarantee for the objective (here: for the expected costs). We will discuss motivations and applications of this optimization goal in Section 2. Note that the probability space, for each h, is defined by the randomness both in the observations and in the strategy, and expectation refers to this probability space. This will become explicit in the formal description below. For calculations it is more convenient to denote the observations by indices $k = 1, \ldots, m$ and hypotheses by $j = 1, \ldots, n$. We write p_{kj} for $P(D_k|h_j)$.

SEARCH COST MINIMIZATION

Let us be given a sequence of costs $c_1 \leq \cdots \leq c_n$, and for each pair k, j ($1 \leq k \leq m$ and $1 \leq j \leq n$) the conditional probability p_{kj} for observation k if j is the target hypothesis. The task is, for each particular observation k, to rank the n hypotheses so that the expected value of c_r (expected search cost) is minimized, where r is the rank of the target. More specifically, since the target j is not known, we wish to *minimize the expected search cost in the worst case*, i.e., maximized over all j. Expectation refers to the random rank of j in the strategy responding to the random observation k caused by j.

Before we give a more formal specification of the problem, we show that cost sequences actually capture the two aforementioned cases. In fact, the "g-selection" problem is now specified by $c_1 = \cdots = c_g = 0$ and $c_{g+1} = \cdots = c_n = 1$, which expresses that the first g tests are for free, and further tests cost some one-time fee. Thus we are interested in maximizing the probability to catch the target. The second case with sequential testing is, obviously, equivalent to $c_r = r$. General monotone sequences c_r can model more complex cost structures (see also Section 2). Anyhow, our algorithm for SEARCH COST MINIMIZATION will work for any monotone costs c_r in the same way, thus we study the problem straightaway in this generality.

Now we turn to the formalization. Since our Searcher can decide on a ranking by a randomized strategy, a Searcher's strategy is specified by a set of probabilities $x_{\pi k} \geq 0$ to choose permutation π of hypotheses in the event of observation k, for all π and k. Of course, strategies must satisfy

$$\forall k: \quad \sum_{\pi} x_{\pi k} = 1. \tag{1}$$

The expected cost in case that j is the target amounts to $\sum_k \sum_\pi c_{r(j,\pi)} x_{\pi k} p_{kj}$, where $r(j, \pi)$ denotes the rank of hypothesis j in permutation π. Our objective is therefore

$$\min \max_j \sum_k \sum_\pi c_{r(j,\pi)} x_{\pi k} p_{kj}. \tag{2}$$

This is an optimization problem with linear constraints and a set of linear objective functions the maximum of which shall be minimized. A standard trick transforms any such problem into a linear program (LP): Introduce a new variable u and new constraints expressing that the jth objective is less than or equal to u, and minimize u. However, this straightforward reduction to an LP does not solve SEARCH COST MINIMIZATION in polynomial time, as we get $mn!$ variables $x_{\pi k}$. (Note that the $r(j, \pi)$ are not problem variables, r is just a "meta"-function describing the indices in general form.)

Download English Version:

https://daneshyari.com/en/article/420630

Download Persian Version:

https://daneshyari.com/article/420630

Daneshyari.com