Upper bound for the approximation ratio of a class of hypercube segmentation algorithms*

Jouni K. Seppänen†

November 10, 2004

Abstract

The Hypercube Segmentation problem was recently introduced by Kleinberg et al. [J. ACM 51 (2004) 263–280], along with several algorithms that select each segment's prototype vector from the segment. The algorithms were shown to have an approximation ratio of at least $2(\sqrt{2}-1) \approx 0.828$. We show that a lemma used in this proof is tight, and that the asymptotic approximation ratio of no algorithm of this type can exceed $5/6 \approx 0.833$.

1 Introduction

The Hypercube Segmentation task [KPR04] is to find for a set S of d-dimensional binary vectors a partitioning $S = S_1 \cup S_2 \cup \cdots \cup S_k$ and corresponding prototype vectors $P_1, P_2, \ldots, P_k$, so as to maximize the sum

$$\sum_{j=1}^{k} \sum_{t \in S_j} t \odot P_j, \tag{1}$$

where the Hamming overlap operator $\odot$ counts the number of positions where two vectors have the same value. Kleinberg et al. [KPR04] give three approximation algorithms for the task, all of which have the property that the prototype vectors are chosen from the input set S. The approximation ratio of one algorithm is shown to be at least $2(\sqrt{2}-1)$; the other two are sampling versions of the same algorithm, and have approximation ratios approaching that of the first one. The proofs rely on the following result:

Lemma 1. *For any set S,*

$$\frac{1}{|S|} \sum_{v \in S} \sum_{t \in S} t \odot v \geq 2(\sqrt{2}-1) \sum_{t \in S} t \odot P^*, \tag{2}$$

where P^* is the optimal prototype for S, i.e., agreeing in each position with the majority of S.

Kleinberg et al. give an example showing that the constant $2(\sqrt{2}-1) \approx 0.828$ in Lemma 1 cannot be increased beyond $5/6 \approx 0.833$. In this paper we give two different generalizations of this example: we show in Section 2 that the constant of the lemma cannot in fact be increased

*Accepted for publication in Information Processing Letters, http://www.elsevier.com/locate/issn/00200190, doi:10.1016/j.ipl.2004.10.006

†HIIT Basic Research Unit, Laboratory of Computer and Information Science, P.O. Box 5400, FI-02015 Helsinki University of Technology, Finland

at all, and give in Section 3 an upper bound of 5/6 for all k-segmentation algorithms that select the prototypes from the data.

While a polynomial approximation scheme exists [AS99], the cited algorithms are still of interest, since the sampling versions have a running time linear in the size of the input and may thus be applicable for data mining tasks, where data sets are often very large.

2 Tightness of the lemma

Since the sum (1) is always an integer, it is clearly impossible to obtain equality in (2). However, we can prove that no larger bound is possible by constructing examples that approach the optimal bound in the limit $d \to \infty$.

Theorem 2. *For any constant $C > 2(\sqrt{2}-1)$, there exists a set S such that for every vector $v \in S$,*

$$\sum_{t \in S} t \odot v < C \sum_{t \in S} t \odot P^*, \tag{3}$$

where P^* is the optimal prototype for S.

Proof. Choose an integer $d \geq 3$, and let c be the integer nearest to $d/\sqrt{2}$. Consider the set S of all $m = \binom{d}{c}$ dimension-d binary vectors that have exactly c ones. The fraction of vectors that have a one in any given position is $c/d > 1/2$, so the optimal prototype P^* has a one in every position. The right-hand side of (3) is thus Ccm.

For a fixed prototype $v \in S$, consider a vector t selected uniformly at random from S. For a randomly selected position, the probability that both vectors have a one is $(c/d)^2$, and the probability that both have a zero is $((d-c)/d)^2$. Thus the expected contribution of this position towards $v \odot t$ is $(d^2 - 2cd + 2c^2)/d^2$, and

$$E(v \odot t) = \frac{d^2 - 2cd + 2c^2}{d}.$$

By summing over all $t \in S$ we find that the left-hand side of (3) divided by cm is

$$\frac{m \cdot E(v \odot t)}{cm} = \frac{d^2 - 2cd + 2c^2}{cd} = \frac{d}{c} - 2 + \frac{2c}{d}. \tag{4}$$

By increasing d we can obtain values of d/c that are arbitrarily good approximations of $\sqrt{2}$, so we can bring the value of (4) arbitrarily close to $\sqrt{2} - 2 + 2/\sqrt{2} = 2(\sqrt{2} - 1)$. □

In fact, S need not include all $\binom{d}{c}$ possible vectors. A smaller construction yielding the same result is obtained by taking an arbitrary vector v having c ones and all cyclic shifts of v.

3 Upper bound for the algorithm

The result of the previous section implies that the analysis by Kleinberg et al. of their algorithm is tight in the case $k = 1$. In the more interesting case of larger k, the example does not seem to be easily generalizable: the similarity between two vectors in the constructed set S can be as low as $2c - d$, which for sufficiently large d is less than half of d. Modifying the construction so that the minimum within-segment similarity is $d/2$ helps guarantee that the induced segmentation coincides with the optimal one. Thus the following example works with arbitrarily many segments but gives a slightly larger bound of 5/6. The result applies to all algorithms that select the prototype vectors from the input data; we shall call such prototypes and algorithms *constrained*.

Theorem 3. *Any constrained algorithm for the k-segmentation problem with $k \geq 2$ has an approximation ratio of at most $5/6$.*

Before proving the theorem, we state a lemma needed in the proof. For a vector v_j, we shall denote by $v_{j,i}$ the i-th coordinate of v_j.

Lemma 4. *For any $k \geq 2$, there is a number b and a set $\{ v_1, \ldots, v_k \}$ of b-dimensional binary vectors such that $v_p \odot v_q = b/2$ for all $1 \leq p < q \leq k$.*

Proof. In the case $k = 2$, we can choose $v_1 = (0, 0)$ and $v_2 = (0, 1)$. For larger k, let $a = \lfloor k/2 \rfloor$ and $m = \binom{k}{a}$, and enumerate all a-element subsets $J_1, \ldots, J_m$ of $\{ 1, \ldots, k \}$. Then let $v_{j,i} = 1$ whenever $j \in J_i$, and $v_{j,i} = 0$ whenever $j \notin J_i$. For any two vectors, there now are $\binom{k-2}{a-2} + \binom{k-2}{a}$ coordinates in which the vectors agree, and $2\binom{k-2}{a-1}$ coordinates in which they disagree. With $k \geq 3$ and our choice of a, the number of disagreements is never smaller than the number of agreements. Indeed, $a - 1 = \lfloor (k-2)/2 \rfloor$, and thus $\binom{k-2}{a-1} \geq \binom{k-2}{a-2}$ and $\binom{k-2}{a-1} \geq \binom{k-2}{a}$. We can thus add some number z of coordinates in which every vector has value 0 until the number of agreements $\binom{k-2}{a-2} + \binom{k-2}{a} + z$ is exactly half of the number of coordinates $b = m + z$. □

Proof of Theorem 3. We shall construct a set S along with a segmentation $S = S_1 \cup \cdots \cup S_k$ and corresponding unconstrained prototypes $P_1, \ldots, P_k$. We shall show that for any constrained prototypes $Q_1, \ldots, Q_k$ the induced segmentation has an approximation ratio of at most $5/6$, compared to the unconstrained prototypes. The dimensionality of the vectors will be $d = 4b$ with the value of b determined by Lemma 4, and each segment S_j will consist of 4 vectors.

We first invoke Lemma 4 to get k vectors $v_1, \ldots, v_k$ in the b-dimensional binary cube such that any two vectors have Hamming overlap $b/2$. From each vector v_j we create a segment S_j of four vectors as follows. We denote by e_p the four-dimensional basis vector that has $e_{p,q} = 1$ if $p = q$, 0 if $p \neq q$, and by $\bar{e}_p$ its complement $1 - e_p$. To construct the p-th vector of S_j, we start from the empty vector, and then for each $i = 1, \ldots, b$ we add to the end of the vector either e_p if $v_{j,i} = 0$, or $\bar{e}_p$ if $v_{j,i} = 1$. For the unconstrained prototype P_j corresponding to segment S_j, we take the median of S_j, which can be obtained by repeating each coordinate of v_j four times.

The value of the segmentation $S_1 \cup \cdots \cup S_k$ with the unconstrained prototypes P_j is $12bk$, since each prototype agrees with each vector in its corresponding segment at exactly $3b$ positions. To complete the proof, we must show for any constrained prototypes $Q_1, \ldots, Q_k$ and a corresponding segmentation $S = T_1 \cup \cdots \cup T_k$ that the sum

$$\sum_{j=1}^{k} \sum_{t \in T_j} t \odot Q_j = \sum_{t \in S} \max_{1 \leq j \leq k} t \odot Q_j \tag{5}$$

is at most $10bk$.

For any two vectors $t, u \in S$, $t \neq u$, we have $t \odot u = 2b$. To see this, we consider two cases. In the first case, both t and u are constructed of copies of the same basis vector e_p and its complement $\bar{e}_p$. They must therefore belong to different segments, say $t \in S_i$ and $u \in S_j$. We then use the facts $v_i \odot v_j = b/2$, $e_p \odot e_p = 4$, and $e_p \odot \bar{e}_p = 0$. In the second case, t and u are constructed of copies of different basis vectors e_p and e_q and their complements, and we can use the fact $e_p \odot e_q = e_p \odot \bar{e}_q = \bar{e}_p \odot e_q = \bar{e}_p \odot \bar{e}_q = 2$. Therefore sum (5) is

$$Q_1 \odot Q_1 + \cdots + Q_k \odot Q_k + (|S| - k) \cdot 2b = k \cdot 4b + 3k \cdot 2b = 10bk.$$

□

4 Acknowledgment

I would like to thank Heikki Mannila for advice.

References

[AS99] Noga Alon and Benny Sudakov. On two segmentation problems. *J. Algorithms*, 33(1):173–184, October 1999.

[KPR04] Jon Kleinberg, Christos Papadimitriou, and Prabhakar Raghavan. Segmentation problems. *J. ACM*, 51(2):263–280, 2004.