Partition Codes

Arkadii D'yachkov[1] Vyacheslav Rykov[2] David Torney[3] Sergey Yekhanin[4]

Abstract — **We introduce the distance concept between two q-ary n-sequences, $2 \le q < n$, called partition distance. This distance is a metric in the space of partitions of a finite n-set, where each partition contains $\le q$ disjoint subsets of the n-set. For the metric, we study codes called q-partition codes which can be applied to statistical analysis of psychological or sociological tests using questionnaires. A construction of q-partition codes based on the first order Reed - Muller codes is presented. A random coding bound is obtained.**

I. Introduction

Let $n > q \ge 2$ be integers, $A_q \triangleq \{0, \dots, q-1\}$ be q-ary alphabet, $\mathcal{M}_q = \{\mu\}$, $\mu = \mu(x)$, be the set containing all $q!$ permutations of A_q.

Let $[n] = \{1, 2, \dots, n\}$. Let $\mathbf{x} = (x_1, \dots, x_n) \in A_q^n$, be an arbitrary fixed q-ary n-sequence. $\mathbf{x}$ identifies a *q-partition* $\{E_0, E_1, \dots, E_{q-1}\}$ of the set

$$[n] = E_0 + E_1 + \cdots + E_{q-1},$$

where $E_a = \{i : x_i = a\}, \quad a \in A_q$.

For $\mu \in \mathcal{M}_q$, let $\mathbf{x}^\mu = (\mu(x_1) \dots, \mu(x_n))$. We call $\mathbf{x}^\mu$ a μ-complement of $\mathbf{x}$. Clearly, all μ-complements of $\mathbf{x}$, $\mu \in \mathcal{M}_q$, identify the same *q-partition*. We denote this q-partition by $\widetilde{\mathbf{x}} = \{E_0, E_1, \dots, E_{q-1}\}$. Let $\mathbf{x}, \mathbf{y} \in A_q^n$. $H(\mathbf{x}, \mathbf{y})$ denotes the Hamming distance between $\mathbf{x}$ and $\mathbf{y}$.

Definition:

$$\mathcal{P}_q(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) = \min_{\mu \in \mathcal{M}_q} H(\mathbf{x}, \mathbf{y}^\mu) = \min_{\mu \in \mathcal{M}_q} H(\mathbf{x}^\mu, \mathbf{y}).$$

is called the *partition distance* between q-ary n-sequences $\mathbf{x}$ and $\mathbf{y}$ or between q-partitions $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{y}}$ of the set $[n]$.

Proposition 1: $\mathcal{P}_q(\mathbf{x}, \mathbf{y})$ satisfies the triangle inequality

$$\mathcal{P}_q(\mathbf{x}, \mathbf{y}) \le \mathcal{P}_q(\mathbf{x}, \mathbf{z}) + \mathcal{P}_q(\mathbf{z}, \mathbf{y}).$$

Hence, $\mathcal{P}_q(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ is a metric in the space of q-partitions.

A different definition of partition distance satisfying the triangle inequality was considered in [2].

Proposition 2: For arbitrary q-partitions $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{y}}$ of the set $[n]$ $\mathcal{P}_q(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) \le \lfloor (q-1)n/q \rfloor$.

[1] A. D'yachkov is with the Department of Probability Theory, Faculty of Mechanics and Mathematics, Moscow State University, Moscow, 119992, Russia. E-mail: dyachkov@mech.math.msu.su. He was supported by RFBR Grant 01-01-00495 and INTAS-00-738.

[2] V. Rykov is with the Department of Mathematics, University of Nebraska at Omaha, 6001 Dodge St., Omaha, NE 68182-0243. E-mail: vrykov@mail.unomaha.edu

[3] D. Torney is with the Los Alamos National Laboratory, MS K70, Los Alamos, New Mexico 87545. E-mail: dct@lanl.gov

[4] S. Yekhanin is with the Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139, USA. E-mail: yekhanin@mit.edu. He was supported in part by NTT Award MIT 2001-04 and NSF grant CCR-0219218.

II. Partition Codes

Let $X = \{\widetilde{\mathbf{x}}_1, \dots, \widetilde{\mathbf{x}}_N\}$ be a family of q-partitions of the set $[n]$. We refer to X as a q-partition code. We say that code X has distance $D(X) = \min_{j \ne j'} \mathcal{P}_q(\widetilde{\mathbf{x}}_j, \widetilde{\mathbf{x}}_{j'})$.

Proposition 3: Let q be a prime power. For every $m = 2, 3, \dots$, there exists a q-partition code X for the set $[n]$ with parameters: $n = q^m$, code size $N = \frac{q^m - 1}{q-1} + 1$ and distance $D(X) = (q-1)q^{m-1} = n \cdot \frac{q-1}{q}$.

The construction is based on the first order Reed-Muller code [3]. Note that the codes presented above have maximal possible distance in the q-partition metric.

Let D, $1 \le D \le n(1 - 1/q)$, be an integer. Let $N_{\mathcal{P}}(q, n, D)$ denote the maximal size of q-partition code X for the set $[n]$ having distance $D(X) \ge D$. Let δ, $0 < \delta < 1 - 1/q$, be fixed, then

$$R_{\mathcal{P}}(q, \delta) = \overline{\lim_{n \to \infty}} \frac{\log_q N_{\mathcal{P}}(q, n, \delta n)}{n}$$

is called the *rate* of q-partition codes.

Proposition 4. For any d, $0 < d < 1 - 1/q$,

$$R_{\mathcal{P}}(q, d) \ge 1 - d \log_q (q-1) - h_q(d),$$

where $h_q(d) = -d \log_q d - (1-d) \log_q (1-d)$.

Note that the bound above coincides with the classical GV bound for distance codes in Hamming metric.

III. Applications

Given integers $2 \le q < n$ consider an arbitrary q-partition code $X = \{\widetilde{\mathbf{x}}(j),\ j = 1, 2 \dots, N\}$ for the set $[n]$. The following psychological testing can be called "nonparametric" (n, q)-*questionnaire*.

- An individual (patient) is asked to split n objects up into q' groups, $1 \le q' \le q$, putting two seemingly similar objects in the same group.
- On the base of preliminary testing of individuals with K known diagnoses one chooses from X a subset of q-partitions $S = \{\widetilde{\mathbf{x}}(j_1), \widetilde{\mathbf{x}}(j_2), \dots \widetilde{\mathbf{x}}(j_K)\}$, $K \le N$, corresponding to these K putative subtypes of a complex disease. The reasonable choice of S can be based on a statistical analysis of q-partition samples.
- If the patient creates an q-partition $\widetilde{\mathbf{y}}$, then we calculate q-partitions $\widetilde{\mathbf{x}}(j_k)$, for which the minimum of partition distance $\mathcal{P}_q(\widetilde{\mathbf{x}}(j_k), \widetilde{\mathbf{y}})$, $k = 1, 2, \dots, K$, is achieved. They yield tentative diagnoses for the disease subtypes.

References

[1] D. Gusfield, "Partition-distance: A problem and class of perfect graphs arising in clustering," *Information Processing Letters*, **82** (2002), pp. 159-164.

[2] B.G. Mirkin, L.B. Tcherny, "On measurement of proximity between various partitions of finite set," *Avtomatika i Telemekhanika*, 1970, No. 5, pp. 120-127 (in Russian).

[3] F.J. MacWilliams, N.J.A. Sloane, *The Theory of Error Correcting Codes*, Amsterdam, The Netherlands: North Holland, 1977.