A note on consecutive ones in a binary matrix

Lemnouar Noui

Abstract. A binary matrix A is said to have the "Consecutive Ones Property" (C1P) if its columns can be permuted so that in each row, the ones appear in one run (i.e., all ones are adjacent). The Consecutive Ones Submatrix (COS) problem is, given a binary matrix A and a positive integer m_0, to find m_0 columns of A that form a submatrix with the C1P property. The matrix reordering problem is to find a matrix A' obtained by permuting the columns of A that minimizes $C_r(A)$ the number of sequences of consecutive ones in A. In this paper, by using two quadratic forms, we calculate the number $C_r(A)$. We apply the obtained results to the orthogonal matrices and Hamming matrices, in addition, the two above problems can be solved for these matrices.

M.S.C. 2000: 65F50, 05A05, 11H50, 11H71.
Key words: consecutive ones, data compression, quadratic form, Hamming distance.

§1. Introduction

Let M be an $n \times m$ matrix over a field K. A *run* in a row of M is a sequence of non-zero entries. The *cost runs* of M is the sum $C_r(M)$ of the number of runs in each of its rows. In the notations of this paper, a binary matrix is a matrix with coefficients in $\mathbb{R}$, but whose entries are only zeros or ones. A binary matrix M has the Consecutive Ones Property (C1P) if there is a permutation of its columns that leaves the 1's consecutive in each row.

The question of whether M has C1P was solved by Booth and Lueker [4] using a linear time algorithm. The problem of verifying the property C1P has applications in computational biology [3], recognizing interval graphs [5] and file organization [6, 9]. It is a special case of the two following problems.

P1. Matrix Reordering.

Given a binary matrix M find a matrix M' obtained by permuting the columns of M that minimizes $C_r(M')$.

This problem is NP-hard [1]. It has applications in data compression which is the process used to reduce the physical size of a block of information.

P2. Consecutive Ones Submatrix (COS). Given a binary matrix M and a positive integer m_0, are there m_0 columns of M that form a submatrix with C1P property?

Applied Sciences, Vol.9, 2007, pp. 141-147.
© Balkan Society of Geometers, Geometry Balkan Press 2007.

Such a question has potential applications in physical mapping with hybridization data [2]. According to [7, 8] this problem is NP-complete.

In his paper, by using two quadratic forms, we calculate the number $C_r(A)$ of runs in a binary matrix A and we give a solution of the two above problems for the orthogonal matrices and Hamming matrices.

§2. Main results

2.1 Computation of the number $C_r(A)$

. Let $E = M_{k,n}(K)$ be the vector space of $k \times n$ matrices over a field K and define the mapping $\phi : E \times E \to K$ by $\phi(A, B) = \mathrm{Tr}\,({}^tBA)$. We denote by P the image of the identity matrix I_n from a permutation of its columns.

Let us note by C_i the ith column and R_j the jth row of the matrix A and $x \cdot y = \langle x, y\rangle$ the standard inner-product of two elements x and y of K^n, then

Proposition 1. *a) ϕ is a symmetric nondegenerate bilinear form.*

b) If q is the quadratic form associated with ϕ, then $\forall A \in E$ $q(AP) = q(A)$, that is, S_n the symmetric group of degree n is contained in the orthogonal group of q.

c) If $q' : E \to K$ is the mapping defined by $q'(A) = C_1C_2 + \cdots + C_{n-1}C_n$, then q' is a quadratic form (not necessarily nondegenerate).

Proof. $a)$ If A, A', B, B' are matrices in E, then

$$\phi(A + A', B) = \mathrm{Tr}\left({}^tB(A + A'\right) = \mathrm{Tr}\left({}^tBA\right) + \mathrm{Tr}\left({}^tBA'\right) = \phi(A, B) + \phi(A', B).$$

Also, it is easy to see that $\phi(A, B + B') = \phi(A, B) + \phi(A, B')$.

For every $\lambda \in K$, we have $\phi(\lambda A, B) = \mathrm{Tr}\,({}^tB(\lambda A)) = \lambda\phi(A, B) = \phi(A, \lambda B)$. Thus ϕ is a bilinear form and symmetric because $\mathrm{Tr}\,({}^tBA) = \mathrm{Tr}\,({}^t({}^tBA)) = \mathrm{Tr}\,({}^tAB)$.

To prove that ϕ is nondegenerate, we assume that $\phi(A, B) = 0$ for every matrix A in E. Let us denote by $E_{i,j}$ the matrix whose entries equal zero except $e_{i,j}$ which equals 1. Let $A = E_{1,1}$ and $B = (b_{i,j})$, so that $\mathrm{Tr}\,({}^tBA) = b_{1,1} = 0$, in the general case, for $A = E_{i,j}$, we obtain $b_{i,j} = 0$, therefore $B = 0$.

$b)$ The quadratic form q associated with ϕ is defined by $q(A) = \mathrm{Tr}\,({}^tA \cdot A)$. Then

$$q(A) = C_1^2 + \cdots + C_n^2 = R_1^2 + \cdots + R_k^2$$

and we conclude that ϕ is a positive form if $K = \mathbb{R}$.

Let P be a permutation matrix. Then

$$q(AP) = \mathrm{Tr}\left({}^t(AP)\,AP\right) = \mathrm{Tr}\left({}^tP\left({}^tAAP\right)\right).$$

Note that the matrices ${}^tP$ and ${}^tAAP$ are square and $\mathrm{Tr}[{}^tP\,({}^tAAP)] = \mathrm{Tr}[({}^tAAP)^t\,P]$. Also $P\,{}^tP = I_n$. Hence $q(AP) = q(A)$.

Let $f_P : E \to E$ be the linear mapping defined by $f_P(A) = AP$. It is easy to verify that $q(f_P(A)) = q(A)$, that is, f_P belongs to $\mathcal{O}_q(E)$ the orthogonal group of

q and $S_n \subset \mathcal{O}_q(E)$.
c) Let $\psi : E \times E \to K$ be the mapping defined by

$$\psi(A, B) = \frac{1}{2}\left[A_1 B_2 + \sum_{i=2}^{n-1} A_i (B_{i-1} + B_{i+1}) + A_n B_{n-1}\right],$$

where A_i is the ith column of A and B_j is the jth column of B. It is easy to show that ψ is a symmetric bilinear form. This form is not necessarily nondegenerate. To see this, it is enough to consider the matrix $B = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & -1 \end{pmatrix}$, we have $\psi(A, B) = 0$ for every matrix A in the space $M_{3,3}(K)$.
Obviously q' is the quadratic form associated with ψ. □

Now, we shall compute the number of runs $C_r(A)$ in the matrix A and minimize it. We have

Proposition 2.

a) $q(A)$ *is the number of ones in* A.

b) If q' *is the quadratic form associated with* ψ*, then* $C_r(A) = q(A) - q'(A)$.

c) $C_r(A)$ *is minimized if and only if* $q'(A)$ *is maximized.*

Proof. a) Since C_i^2 equals the number of ones in the ith column, $q(A) = C_1^2 + \cdots + C_n^2$ is the number of ones in A.

b) Let us write ${}^tA \cdot A = (l_{i,j})$, then the number of runs in A is equal to $\mathrm{Tr}\,({}^tAA) - \sum_{j=1}^{n-1} l_{j,j+1}$, but we can express this by using the columns C_i, that is, $C_r(A) = q(A) - q'(A)$ where $q'(A) = C_1C_2 + \cdots + C_{n-1}C_n$.

c) It follows from Proposition 1, that $q(AP) = q(A)$; thus $q(A)$ is invariant by permuting the columns of A, then we deduce the last result. □

Remark 1. *To simplify the computation of the number* $C_r(A)$*, we distinguish the two cases:*

a) If $n \leq k$, we write $q(A) = C_1^2 + \cdots + C_n^2$ and $q'(A) = C_1C_2 + \cdots + C_{n-1}C_n$.

b) If $k \leq n$, we write $q(A) == R_1^2 + \cdots + R_k^2$ and $q'(A) = q'(R_1) + \cdots + q'(R_k)$ where R_m is the mth row of A.

Example 1. *Let the matrices*

$$A = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad \textit{and} \quad B = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

Then for the first matrix, we have $k = 5$ *and* $n = 4$*. Hence*

$$C_r(A) = q(A) - q'(A) = 10 - 2 = 8.$$

For the second matrix, we have $k = 3$, $n = 6$ and

$$q(B) = R_1^2 + R_2^2 + R_3^2 = 9,\ q'(B) = q'(R_1) + q'(R_2) + q'(R_3) = 4.$$

So $C_r(B) = 5$.

If x is a vector in K^n, the Hamming weight $\mathrm{wt}(x)$ is the number of nonzero components in x, the Hamming distance $d(x, y)$ between two vectors x and y is the Hamming weight of $x - y$.

Let $A = (a_{i,j})$ be a matrix, let us note $a'_{i,j} = d(C_i, C_j)$ where C_i is the ith column of A.

Proposition 3. *If A is an $k \times n$ binary matrix, then*

$$C_r(A) = \frac{1}{2}\Big[\sum_{i=1}^{n-1} d(C_i, C_{i+1}) + \mathrm{wt}(C_1) + \mathrm{wt}(C_n)\Big].$$

Proof. Note that $a'_{i,j} = d(C_i, C_j) = (C_i - C_j)^2$, by Proposition 2,

$$\begin{array}{rcl} C_1^2 + \sum_{i=1}^{n-1} d(C_i, C_{i+1}) + C_n^2 & = & C_1^2 + \sum_{i=1}^{n-1} (C_i - C_{i+1})^2 + C_n^2 \\ & = & 2\left(C_1^2 + \cdots + C_n^2 - C_1C_2 - \cdots - C_{n-1}C_n\right) \\ & = & 2N_b(A). \end{array}$$

Obviously $C_i^2 = \mathrm{wt}(C_i)$, consequently the last proposition is proved. □

2.2 COS and Matrix Reordering problems

2.2.1 Orthogonal matrices

An $n \times n$ matrix M over the field K satisfying $M\,{}^tM = I_n$ is called an orthogonal matrix.

Proposition 4. *a) If A is a $n \times n$ binary orthogonal matrix, then*

$$C_r(A) = C_r(AP) = n \ \ \textit{for every permutation matrix } P.$$

b)Every orthogonal binary matrix A has C1P property.

Proof. a) Since A is orthogonal, $q'(A)$ =0, by using Proposition 2, we have $C_r(A) = q(A) = Tr(A\,{}^tA) = n$.
b) The orthogonal matrix A is invertible, then each row of A contains at least one non-zero element. By proposition 4 we have $C_r(A) = n$, so A has C1P property. □

For the first problem the suitable permutation for the orthogonal binary matrix is for example the identity permutation, as answer to the second problem $m_0 = n$ is the largest number of columns that form a submatrix with C1P property.

2.2.2 Hamming matrices

Let $n = 2^k - 1$ with $k \geq 2$. The Hamming matrix is a $k \times \left(2^k - 1\right)$ binary matrix whose columns consist of all nonzero binary vectors of length k:

$$A = \begin{pmatrix} 0 & 0 & 0 & . & 1 \\ 0 & 0 & 0 & . & 1 \\ . & . & . & . & . \\ 0 & 1 & 1 & . & . \\ 1 & 0 & 1 & . & 1 \end{pmatrix}.$$

Let $p_k = \begin{pmatrix} 1 & 2 & \dots & n \\ x_1 & x_2 & \dots & x_n \end{pmatrix}$ and P_k denote respectively a permutation of $n = 2^k - 1$ columns of A and the image of the identity matrix I_{2^k-1} by this permutation.

Proposition 5. *Let A be an $k \times \left(2^k - 1\right)$ Hamming matrix, then*

a) The number of ones in A equals $q(A) = k \times 2^{k-1}$,
$q'(R_m) = 2^{m-1} \times \left(2^{k-m} - 1\right)$ and $C_r(A) = 2^k - 1$.

b) There is a permutation p_k defined recursively as follows:

$$\begin{cases} p_1 : \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ \text{If } p_{k'-1} = \begin{pmatrix} 1 & . & m \\ x_1 & . & x_m \end{pmatrix} \text{ where } m = 2^{k'-1} - 1, \text{ then} \\ p_{k'} = \begin{pmatrix} 1 & . & m & m+1 & . & 2m & 2m+1 \\ x_1 & . & x_m & x_m + 2x_m & . & x_1 + 2x_m & 2x_m \end{pmatrix} \end{cases}$$

such that $C_r(AP_k)$ is minimized, that is, $C_r(AP_k) = 2^{k-1}$.

Proof. a) If A is a Hamming matrix, its mth column is determined by writing the number m in the basic 2. Since every row of A consists of 2^{k-1} ones, it follows that $q(A) = k \times 2^{k-1}$. As $k \leq n = 2^k - 1$, to find $C_r(A)$ we write $C_r(A) = q(A) - q'(A) = k \cdot 2^{k-1} - (q'(R_1) + q'(R_2) + ... + q'(R_k))$ where $q'(R_m) = 2^{m-1} \times \left(2^{k-m} - 1\right)$, thus we obtain $C_r(A) = 2^k - 1$.

b) By proposition 3, in order to minimize $C_r(AP)$, it is enough to minimize $d(C_i, C_{i+1})$, $wt(C_1)$ and $wt(C_n)$, with this intention we shall define a permutation p_k recursively as follows

$$p_1 : \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$p_2 : \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}$$

$$\vdots$$

$$p_{k'-1} = \begin{pmatrix} 1 & ... & m \\ x_1 & ... & x_m \end{pmatrix} \text{ with } m = 2^{k'-1} - 1$$

$$p_{k'} = \begin{pmatrix} 1 & ... & m & m+1 & ... & 2m & 2m+1 \\ x_1 & ... & x_m & x_m + 2x_m & ... & x_1 + 2x_m & 2x_m \end{pmatrix}.$$

Then $C_r(AP_k)$ is minimized because $d(C_i, C_{i+1}) = \text{wt}(C_1) = \text{wt}(C_n) = 1$, so that $C_r(AP_k) = \frac{1}{2}\left[\left(2^k - 1\right) - 1 + 2\right] = 2^{k-1}$. □

Proposition 6. *Let A be an $k \times \left(2^k - 1\right)$ Hamming matrix, then*
a) A has C1P property if and only if $k = 1$ or $k = 2$.

b) For $k' \leq 2k-1$ there exists an $k' \times k$ submatrix with C1P property.

Proof. a) By proposition 5 $C_r\left(AP_k\right) = 2^{k-1}$, then A has C1P property if and only if $2^{k-1} = k$. The function $f(x) = (x-1)ln2 - lnx$ shows that $2^{k-1} = k \Leftrightarrow k = 1$ or $k = 2$.
b) If we take the $2k-1$ columns $C_{2^i-1}, 1 \leq i \leq k$ and $C_{2^k-2^s}$, $1 \leq s \leq k-1$ of A, we obtain a submatrix A' with C1P property. Now for $k' \leq 2k-1$ it is enough to take the first k' columns of A', so the COS problem can be solved for the Hamming matrices. □

The first problem is solved by proposition 5.

Example 2. *Consider the 3×7 Hamming matrix* $A = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$,
then

$k = 3$, $n = 2^3 - 1$ *and* $C_r\left(A\right) = 2^k - 1 = 7$.

To find the permutation p_3, we write $p_1 : \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $p_2 : \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}$,

$p_3 : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 3 & 2 & 6 & 7 & 5 & 4 \end{pmatrix}$, *consequently, we have*

$AP_3 = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$ *and* $C_r\left(AP_3\right) = 2^{k-1} = 4$.

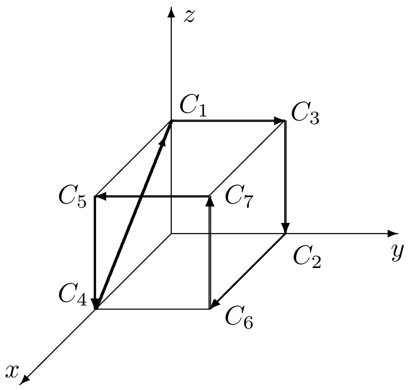
$k' = 5$ *is the largest number of columns of A that form a submatrix*
$A' = \begin{pmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{pmatrix}$ *with C1P property.*

Remark 2. *The traveling Salesman Problem (TSP) [1] requires that we find the shortest path visiting each of a given set of cities and returning to the starting point.*

The permutation p_k gives a solution of the traveling Salesman problem for $n = 2^k - 1$ cities

$$C_1 \begin{pmatrix} 0 \\ 0 \\ . \\ 0 \\ 1 \end{pmatrix} \quad C_2 \begin{pmatrix} 0 \\ 0 \\ . \\ 1 \\ 0 \end{pmatrix} \quad C_3 \begin{pmatrix} 0 \\ 0 \\ . \\ 1 \\ 1 \end{pmatrix} \quad \ldots \quad C_n \begin{pmatrix} 1 \\ 1 \\ . \\ 1 \\ 1 \end{pmatrix}$$

for example if $k = 3$, it is easy to see that the distance is minimal for the path 1326754, *see the figure below.*



References

[1] D. Johnson, S. Krishnan, J.ChhuganI, S. Kumar and S. Venkatasubramanian, *Compressing large Boolean matrices using reodering techniques*, Proc. of the 30th VLDB Conference, Toronto, Canada, 2004, ???.

[2] F. Alizadeh, R.M.Karp, D.K. and G. Zweig, *Physical mapping of chromosomes using unique probes*, Journal of Computational Biology 2 (1995), 159-1.

[3] F. Annexstein, R. Swaminathan, *On testing consecutive-ones property in parallel*, Discrete Applied Mathematics 88 (1998), 7-28.

[4] K.S. Booth, G. S. Lueker, *Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms*, J. Comput. Systems Sci. 13 (3) (1976), 335-379.

[5] J. S. Deogun, K. Gopalakrishnan, *Consecutive retrieval property-revisited*, Information Processing Letters 69 (1999), 15-20.

[6] M. Flammini, G. Gambosi, S. Salomone, *Boolean Routing*, Lecture Notes in Computer Science 725, 1933.

[7] M.R. Garey and D. S. Johnson, *Computers and intractability: A guide to the theory of NP-completeness*, San Francisco: W.H. Freeman, 1979.

[8] S. P. Ghosh, *File organization: the consecutive retrieval property*, Commun. ACM 15 (1972), 802-808.

[9] L. T. Kou, *Polynomial complete consecutive information retrieval problems*, SIAM J. Comput. 6 (1977), 67-75.

Author's address:

Lemnouar Noui
Department of mathematics, Faculty of Science,
University of Batna, Algeria
e-mail: nouilem@yahoo.fr