New Strategies for Evaluating the Performance of Typical Testor Algorithms

Eduardo Alba, Diego Guilcapi, and Julio Ibarra

Universidad San Francisco de Quito, Colegio de Ciencias e Ingeniería,
Diego de Robles y Vía Interoceánica, Quito, Ecuador
{ealba,jibarra}@usfq.edu.ec, diego.guilcapi@estud.usfq.edu.ec
http://www.usfq.edu.ec

Abstract. Typical testors have been used in feature selection and supervised classification problems in the logical combinatorial pattern recognition. Several algorithms have been used to find the set of all typical testors of a basic matrix. This algorithms are based on different heuristics. There is no doubt these algorithms find the set of all typical testors. However, the time spent on this search strategies, differs between them. Due to size of this set, the search time is a critical factor. There is not a standard procedure to evaluate the time performance of typical testors algorithms. In this paper we introduce a strategy to solve this problem through a new set of test matrices. These test matrices have the property that the set's cardinality of all typical testors is known in advance.

Keywords: Logical combinatorial PR, feature selection, testor theory, typical testors algorithms, test matrices.

1 Introduction

When is used the logical combinatorial approach in the solution of supervised pattern recognition problems, typical testors play a very important role [1],[2]. In a basic approximation, a testor is a collection of features that discriminates all descriptions of objects belonging to different classes, and is minimal in the partial order determined by set inclusion. Through the stages of the problem solution, typical testors can be applied to satisfy different goals. For example, to construct a hierarchical order of features according to their relevance [3],[4],[5],[6] and/or to determine the support sets in the partial precedence algorithms [2]. In recent works, typical testors have been used in text mining [7], [8],[9].

In the logical combinatorial approach, the data of a supervised pattern recognition problem can be reduced to a matrix [2]. The typical testors are searched among all possible subsets of column labels of this matrix. The computation of the set of all typical testors may take a lot of time. Several deterministic heuristics have been used to find this set [10], [16]. Also evolutionary algorithms have been introduced to deal with matrices of high dimension [12], [13].

It is necessary to have a set of instances for testing the performances of the different proposed algorithms to calculate the typical testors. On the other hand,

L. Alvarez et al. (Eds.): CIARP 2012, LNCS 7441, pp. 813–820, 2012.

having these test matrices the typical testor problem becomes an useful benchmark problem to evaluate stochastic optimization algorithms. In [15] were given the first steps for solving this topic. Test matrices should be built so that the set of all typical testors can be determined a priori. Moreover, these matrices should allow controlling aspects such as: dimension, cardinality of each typical testor and the number of typical testors.

The paper is organized as follows: first, we formally introduce the concept of typical testor for boolean matrices. Then, we present theoretical results for the construction of test matrices. Finally, we illustrate how to generate some classes of test matrices.

2 The Concept of Testor and Typical Testor

Let U be a collection of objects. These objects are described by a set of n features and are grouped into l classes. By comparing feature to feature, each pair of objects belonging to different classes, we obtain a matrix $M = [m_{ij}]_{l \times n}$ where $m_{ij} \in \{0, 1\}$ and l is the number of pairs. $m_{ij} = 1\,(0)$ means that the objects of pair denoted by i are different (similar) in the feature j. Let $I = \{i_1, \ldots, i_j\}$ be the set of the rows of M and $J = \{j_1, \ldots, j_n\}$ the set of labels of its columns (features). Let $T \subseteq J$, $M_{/T}$ is the matrix obtained from M eliminating all columns not belonging to the set T.

Definition 1. *A set $T = \{j_{k_1}, \ldots, j_{k_l}\} \subseteq J$ is a testor of M if there is not any zero row in $M_{/T}$.*

Definition 2. *The feature $j_{k_r} \in T$ is typical with respect to (wrt) T and M if $\exists q$, $q \in \{1, \ldots, l\}$ such that $a_{i_q j_{k_r}} = 1$ and for $s > 1$ $a_{i_q j_{k_p}} = 0$, $\forall p$, $p \in \{1, \ldots, s\}\, p \neq r$.*

Definition 3. *A set T has the property of typicality wrt a matrix M if all features in T are typical wrt T and M.*

Proposition 1. *A set $T = \{j_{k_1}, \ldots, j_{k_l}\} \subseteq J$ has the property of typicality wrt matrix M if and only if identity matrix can be obtained in $M_{/T}$, by eliminating some rows.*

Definition 4. *A set $T = \{j_{k_1}, \ldots, j_{k_s}\} \subseteq J$ is denominated typical testor of M, if T is a testor and it has the property of typicality wrt M .*

Let $\mathbf{a}$ and $\mathbf{b}$ be two rows from M.

Definition 5. *We say that $\mathbf{a}$ is less than $\mathbf{b}$ $(\mathbf{a} < \mathbf{b})$ if $\forall i\, a_i \leq b_i$ and $\exists j$ such that $a_j \neq b_j$.*

Definition 6. *$\mathbf{a}$ is a basic row from M if there is not any row less than $\mathbf{a}$ in M.*

Definition 7. *The basic matrix of M is the matrix M' that only containing all different basic rows of M.*

The following proposition [10] is a characterization of the basic matrix:

Proposition 2. *M' is a basic matrix if and only for any two rows* $\mathbf{a}, \mathbf{b} \in M'$, *there exist two columns i and j that $a_i = b_j = 1$ and $a_j = b_i = 0$.*

Given a matrix A, let $\Psi^*(A)$ be the set of all typical testors of matrix A.

Proposition 3. $\Psi^*(M) = \Psi^*(M')$.

According with the proposition 3, to obtain the set $\Psi^*(M)$, it is very convenient to find the matrix M', and then, to calculate the set $\Psi^*(M')$. Taking into account that M' has equal or less number of rows than M, the efficiency of the algorithms should be better for M'. In fact, all generated test matrices described in this paper are basic.

3 Matrix Operators and It's Properties

We will define the operator *concatenation* denoted by φ :

$$\varphi : \mathcal{M}_{p\times q} \times \mathcal{M}_{p\times q'} \to \mathcal{M}_{p\times(q+q')},\ q > 0,\ p > 0 \tag{1}$$

Given two matrices $A \in \mathcal{M}_{p\times q}$ and $B \in \mathcal{M}_{p\times q'}$, $\varphi(A, B) = [A\, B]$ i.e. the matrix $C = \varphi(A, B)$ is the matrix formed by two blocks: A and B.

Properties of operator φ:

1. $\varphi(\varphi(A, B), C) = \varphi(A, \varphi(B, C)) = \varphi(A, B, C)$.
2. Let A and B be boolean matrices. If A or B are basic matrices then $\varphi(A, B)$ is a basic matrix.

The first property is trivial starting from the characteristics of the *concatenation* operator. The second property can be demonstrated starting from the characteristics of this operator and the proposition 2.

We will define the operator *combinatory merge* denoted by θ:

$$\theta :\ \mathcal{M}_{p\times q} \times \mathcal{M}_{p'\times q'} \to \mathcal{M}_{pp'\times(q+q')} \tag{2}$$

Given two matrices $A = [a_{ij}]_{p\times q}$ and $B = [b_{ij}]_{p'\times q'}$ the operation θ is defined as follows:

$$\theta(A, B) = \begin{bmatrix} A(1,:) & B(1,:) \\ \dots & \dots \\ A(1,:) & B(p',:) \\ \dots & \dots \\ A(p,:) & B(1,:) \\ \dots & \dots \\ A(p,:) & B(p',:) \end{bmatrix} \tag{3}$$

Properties of operator θ

1. $\theta(\theta(A,B),C) = \theta(A,\theta(B,C)) = \theta(A,B,C)$.
2. Let A and B be boolean matrices. If A and B are basic matrices, then $\theta(A,B)$ is a basic matrix.

Note that for this operation both matrices should be basic.

Let I_n be an $n \times n$ identity matrix.

Properties of matrix I_n:

1. I_n is a basic matrix.
2. The number of all typical testors of matrix I_n is equal to 1, i.e. $|\Psi^*(I_n)| = 1$
3. $\Psi^*(I_n) = \{J_{I_n}\}$ where J_{I_n} is the set of all column labels of matrix I_n.

Once defined these operators we will present the theoretical framework that we use to develop the strategies to construct test matrices.

4 Theoretical Results on the Determination of Ψ^*

Let Q denote the concatenation of N $(N > 1)$ matrices $B \in \mathcal{M}_{p\times q}$. Let $\Psi^*(B) = \{T_1, \ldots, T_\nu\}$.

Theorem 1. $|\Psi^*(Q)| = N^{|T_1|} + \ldots + N^{|T_\nu|}$

Let $A_1, \ldots, A_m$ be m matrices such as $A_i = I_{n_i} \forall i \in \{1, \ldots, m\}$. Let $J_{A_1}, \ldots, J_{A_m}$ denote the sets of all column labels of these matrices and $\left\{J_{A_i} \cap J_{A_j}\right\} = \emptyset \, \forall i, j \in \{1, \ldots, m\}$, $i \neq j$. Let A be a matrix, which is obtained applying the combinatory merge operator to matrices A_i, i. e. $A = \theta(A_1, \ldots, A_m)$.

Theorem 2. $\Psi^*(A) = \{\Psi^*(A_1), \ldots, \Psi^*(A_m)\} = \{J_{A_1}, \ldots, J_{A_m}\}$

Corollary 1. $|\Psi^*(A)| = m$

Let P denote the concatenation of N $(N > 1)$ matrices A.

Corollary 2. $|\Psi^*(P)| = N^{n_1} + \ldots + N^{n_m}$

The proof of these results are provided in [15].

5 Strategies to Generate Test Matrices

In this section, we describe new ways to generate the following test matrices: Matrices of equal size and different number of Typical Testors (TT) and Matrices with different dimensions and the same number of TT, based on the operators φ and θ introduced before.

5.1 Matrices with Equal Size and Different Number of Typical Testors (TT)

Two matrices Q_1 and Q_2, with equal size and different number of T, satisfy two conditions: $dim(Q_1) = dim(Q_2)$ and $|\Psi^*(Q_1)| \neq |\Psi^*(Q_2)|$.

And these matrices are generated as follows:

Suppose $B_1 = I_4$ and $B_2 = \theta(I_2, I_2)$. So $Q_1 = \varphi(\underbrace{B_1, ..., B_1}_{N\ times})$, $Q_2 = \varphi(\underbrace{B_2, ..., B_2}_{N\ times})$

By Theorem 1: $|\Psi^*(Q_1)| = N^4$, $|\Psi^*(Q_2)| = 2 \times N^2$ and $dim(Q_i) = 4 \times 4N$ for $i = 1, 2$.

Note that, because the structure of Q_2, the difference between the number of TT of these matrices increases in a quadratic form. We know $Q_2 = \varphi\ (\underbrace{B_2, ..., B_2}_{N\ times}) = \varphi(\underbrace{\theta(I_2, I_2), ..., \theta(I_2, I_2)}_{N\ times})$. So, the operator θ used to build B_2, helps increase the size of the matrix, but does not change the cardinality of its TT. That means, when we apply the operator φ to B_1 and B_2 respectively ($Q_1 = \varphi(\underbrace{B_1, ..., B_1}_{N\ times})$; $Q_2 = \varphi(\underbrace{B_2, ..., B_2}_{N\ times})$), the number of TT of these matrices will be influenced by the cardinality of the TT of the involved matrices Q_1 and Q_2. Thus, $|\Psi^*(Q_1)| > |\Psi^*(Q_2)|\ for\ N > 1$.

We have generated two types of matrices with equal size and different number of TT; however, we can generate 46 additional matrices with this features. Twenty-three $(4! - 1)$ matrices from $B_1 = I_4$ through transpositions and permutations of the columns of this matrix, and then applying the concatenation operator (φ). Similarly, 23 matrices from $B_2 = \theta(I_2, I_2)$ and then applying the concatenation operator. These kind of matrices have the same number of TT and the same size of $Q_1$and Q_2 respectively, but they look different.

Moreover, we can appreciate that the number of rows of Q_1 and Q_2 is 4. But, there exist some results that can help us to generate matrices with more rows:

Let n an even number $(n = 2, 4, 6, ...)$,

Proposition 4. $dim(I_{n^2}) = dim(\varphi(\underbrace{\theta(I_n, I_n), ..., \theta(I_n, I_n)}_{n/2\ times})) = n^2 \times n^2$

Proof. $dim(I_{n^2}) = n^2 \times n^2$, $dim(\theta(I_n, I_n)) = n^2 \times 2n$

$dim(\varphi(\underbrace{\theta(I_n, I_n), ..., \theta(I_n, I_n)}_{n/2\ times})) = n^2 \times \underbrace{2n + 2n + ... + 2n}_{n/2\ times}$

$= n^2 \times 2n\,[n/2] = n^2 \times n^2$

Proposition 5. $|\Psi^*(I_{n^2})| \neq \left|\Psi^*(\varphi(\underbrace{\theta(I_n, I_n), ..., \theta(I_n, I_n)}_{n/2\ times}))\right|$

Proof. $|\Psi^*(I_{n^2})| = 1$

$\left|\Psi^*(\varphi(\underbrace{\theta(I_n, I_n), ..., \theta(I_n, I_n)}_{n/2\ times}))\right| = \frac{n^n}{2^{n-1}}$ since $\theta(I_n, I_n)$ has two TT, each one with cardinality n. Thus, when we concatenate $\frac{n}{2}$ *times* , we get $\left[\frac{n}{2}\right]^n + \left[\frac{n}{2}\right]^n = 2\left[\frac{n}{2}\right]^n = \frac{n^n}{2^{n-1}}$.

Anyway, we can generate matrices with the same size and different number of TT, but with more than 4 rows.

Considering the above results we can generalize the construction of new matrices of this type using φ operator. Let $Q_3 = \varphi(I_{n^2}, ..., I_{n^2})$ and

$$Q_4 = \varphi\left[\underbrace{\varphi(\underbrace{\theta(I_n, I_n), ..., \theta(I_n, I_n)}_{n/2\ times}), ..., \varphi(\underbrace{\theta(I_n, I_n), ..., \theta(I_n, I_n)}_{n/2\ times})}_{N\ times}\right]$$

where:
$dim(Q_3) = dim(Q_4) = n^2 \times n^2 * N$, $|\Psi^*(Q_3)| = N^{n^2}$
$|\Psi^*(Q_4)| = 2N^n$, $|\Psi^*(Q_3)| \neq |\Psi^*(Q_4)|$

The next table shows the number of TT when we modify the number of concatenated matrices

Table 1. Number of typical testors when we modify the number of concatenated matrices

Concatenated Matrices (N)	$Dimension: n^2 \times n^2 * N$	$\lvert\Psi^*(Q_3)\rvert = N^{n^2}$	$\lvert\Psi^*(Q_4)\rvert = 2N^n$
1	$n^2 \times n^2$	1	2
2	$n^2 \times 2n^2$	2^{n^2}	$2 * 2^n$
⋮	⋮	⋮	⋮

5.2 Matrices with Different Dimensions and the Same Number of Typical Testors (TT)

Two matrices Q_1 and Q_2 with different dimensions and the same number of typical testors, satisfy two conditions: $dim(Q_1) \neq dim(Q_2)$, $|\Psi^*(Q_1)| = |\Psi^*(Q_2)|$

These matrices are $Q_1 = \varphi\left[\underbrace{I_{n_1}, ..., I_{n_1}}_{N_1\ times}\right]$ and $Q_2 = \varphi\left[\underbrace{I_{n_2}, ..., I_{n_2}}_{N_2\ times}\right]$

where $n_1 \neq n_2$, $N_1^{n_1} = N_2^{n_2}$
However, note if $n_1, n_2 \geq 2$ and
$n_2 = 2n_1$
$N_1 = N_2^2$
we get $n_1 \neq n_2$ and $N_1^{n_1} = \left[N_2^2\right]^{n_1} = N_2^{n_2}$

Thus, we can generate several matrices that satisfy this features. The next table illustrate some examples

Table 2. Test Matrices with the same number of TT and different size, for different values of $n_1, n_2, N_1\ y\ N_2$

n_1	n_2	N_1	N_2	$dim(Q_1) = n_1 \times n_1 * N_1$	$dim(Q_2) = n_2 \times n_2 * N_2$	$\lvert\Psi^*(Q_1)\rvert = N_1^{n_1}$	$\lvert\Psi^*(Q_2)\rvert = N_2^{n_2}$
2	4	4	2	2×8	4×8	4^2	2^4
2	4	9	3	2×18	4×12	9^2	3^4
2	4	16	4	2×32	4×16	16^2	4^4
2	4	25	5	2×50	4×20	25^2	5^4
2	4	36	6	2×72	4×24	36^2	6^4

On the other hand, there exist other way to generate matrices of this category. We observe that identity matrices of order $2, 3, 4, ...$ have different size, but the same number of TT. So, when we apply θ operator n times we get matrices with different number of TT ($from\ 1\ to\ n$) and different dimension.

Other way to generate matrices of this category, is applying θ operator to n identity matrices of different order. The general form of this kind of matrix is $\theta(I_{s_1}, I_{s_2}, ..., I_{s_n})$ where s_i ,with i=1, 2, 3, ... , can take the next values: $2, 3, 4, ...$ Its number of TT is n and its dimension is $\prod_{i=1}^{n} s_i \times \sum_{i=1}^{n} s_i$.

For example, when $n = 3$, $s_1 = 5$, s_2= 7, $s_3 = 12$; the resulting matrix is $\theta(I_5, I_7, I_{12})$ whose dimension is $(5 * 7 * 12) \times (5 + 7 + 12)$, and it has 3 TT.

Note when s_i=$s_j (i \neq j)$ we get matrices with high dimensions and few typical testors.

6 Conclusions

The objective of this work is applying a general method to create test matrices for evaluating the performance of strategies to search typical testors. The method proposed is based on theoretical results that are simple in their formulation. However, these results are a solid support to obtain different instances that allow to test several edges of the typical testor search problems.

In this paper, we have described new ways to generate the following test matrices: Matrices of equal size and different number of TT and Matrices with different dimensions and the same number of TT. We found that θ operator helps increase the size of the matrix, but does not change the cardinality of its TT. Moreover, when we combine this operator with operator φ, we have showed that we can generate several kinds of test matrices.

We have presented some examples that illustrate how to generate the instance needed to test the influence of a particular feature on the typical testor search algorithms. Taking into account that we have introduced a flexible and general tool for the construction of test matrices, we expect that this work will allow standarize the procedures to evaluate the performance of Typical Testor algorithms.

References

1. Dmitriev, A.N., Zhuravlev, Y.I., Krendeleiev, F.: On the mathematical principles of patterns and phenomena classification. Diskretnyi Analiz. 7, 3–15 (1966); Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. J. Mol. Biol. 147, 195–197 (1981)
2. Ruiz, J., Lazo, M., Alba, E.: An overview of the concept of testor. Pattern Recognition 34, 13–21 (2001)
3. Lazo, M., Ruiz, J.: Determining the feature relevance for non classically described objects and a new algorithm to compute typical fuzzy testors. Pattern Recognition Letters 16, 1259–1265 (1995)
4. Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F.: Feature Selection for Natural Disaster Texts Classification Using Testors. In: Yang, Z.R., Yin, H., Everson, R.M. (eds.) IDEAL 2004. LNCS, vol. 3177, pp. 424–429. Springer, Heidelberg (2004)
5. Vázquez, R., Godoy, S.: Using testor theory to reduce the dimension of neural network models. Special Issue in Neural Networks and Associative Memories 28, 93–103 (2007)
6. Santos, J., Carrasco, A., Martínez, J.: Feature selection using typical testors applied to estimation of stellar parameters. Computación y Sistemas 8, 15–23 (2004)
7. Pons-Porrata, A., Gil-García, R.J., Berlanga-Llavori, R.: Using Typical Testors for Feature Selection in Text Categorization. In: Rueda, L., Mery, D., Kittler, J. (eds.) CIARP 2007. LNCS, vol. 4756, pp. 643–652. Springer, Heidelberg (2007)
8. Pons-Porrata, A., Ruiz-Shulcloper, J., Berlanga-Llavori, R.: A Method for the Automatic Summarization of Topic-Based Clusters of Documents. In: Sanfeliu, A., Ruiz-Shulcloper, J. (eds.) CIARP 2003. LNCS, vol. 2905, pp. 596–603. Springer, Heidelberg (2003)
9. Li, F., Zhu, Q., Lin, X.: Topic discovery in research literature based on non-negative matrix factorization and testor theory. In: Asia-Pacific Conference on Information Processing, vol. 2, pp. 266–269 (2009)
10. Sánchez, G.: Efficient algorithms to calculate typical testors from a basic matrix. In: Design and Program. Master Thesis, BUAP, México (1997)
11. Morales-Manilla, L.R., Sanchez-Diaz, G.: FS-EX Plus: A New Algorithm for the Calculation of Typical FS-Testor Set. In: Rueda, L., Mery, D., Kittler, J. (eds.) CIARP 2007. LNCS, vol. 4756, pp. 380–386. Springer, Heidelberg (2007)
12. Sánchez, G., Lazo, M., Fuentes, O.: Genetic algorithm to calculate minimal typical testors. In: Proceedings of the IV Iberoamerican Symposium on Pattern Recognition, pp. 207–214 (1999)
13. Alba, E., Santana, R., Ochoa, A., Lazo, M.: Finding typical testors by using an evolutionary strategy. In: Proceedings of the V Ibero American Symposium on Pattern Recognition, pp. 267–278 (2000)
14. Garey, M., Johson, D.: Computers and intractability: A guide to the theory of NP-completeness. W. H. Freeman and Company, New York (1979)
15. Alba, E., Santana, R.: Generación de matrices para evaluar el desempeño de estrategias de búsqueda de testores típicos. Avances en Ciencias e Ingenierías, 30–35 (2010)
16. Sanchez-Diaz, G., Piza-Davila, I., Lazo-Cortes, M., Mora-Gonzalez, M., Salinas-Luna, J.: A Fast Implementation of the CT_EXT Algorithm for the Testor Property Identification. In: Sidorov, G., Hernández Aguirre, A., Reyes García, C.A. (eds.) MICAI 2010, Part II. LNCS, vol. 6438, pp. 92–103. Springer, Heidelberg (2010)