Pattern Recognition Letters 80 (2016) 46–51

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Generating synthetic test matrices as a benchmark for the computational behavior of typical testor-finding algorithms☆

Eduardo Alba-Cabrera[a], Salvador Godoy-Calderon[b,*], Julio Ibarra-Fiallo[a]

[a] Colegio de Ciencias e Ingenierías, Universidad San Francisco de Quito (USFQ), Diego de Robles y Vía Interoceánica, Quito, Ecuador
[b] Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), Av. Juan de Dios Bátiz, Esq. Miguel Othón de Mendizábal. Col. Nueva Industrial Vallejo, Ciudad de México, México

ARTICLE INFO

Article history:
Received 4 September 2015
Available online 7 May 2016

Keywords:
Feature selection
Testor theory
Typical testor algorithms

ABSTRACT

Each typical testor-finding algorithm has a specific sensibility towards the number of rows, columns or typical testors within its input matrix. In this research a theoretical framework and a practical strategy for designing test matrices for typical testor-finding algorithms is presented. The core of the theoretical framework consists on a set of operators that allow the generation of basic matrices with controlled dimensions and for which the total number of typical testors is known in advance. After presenting the required theoretical foundation, and the logic for measuring a testor-finding algorithm's computational behavior, the proposed strategy is used to assess the behavior of three well-known algorithms: *BT*, *LEX*, and *FastCTExt*. Unexpected behaviors, observed during the test experiments, are analyzed and discussed, revealing previously unknown characterizations of the tested algorithms that neither a complexity analysis, nor a random experimentation protocol could have revealed beforehand.

1. Introduction

When dealing with supervised classification problems in pattern recognition, two common tasks are: (1) the determination of the informational relevance for each feature used to describe the objects under study, and (2) the selection of class representative objects from a supervision sample. Testor Theory [6], provides a solid framework under which both tasks can be tackled, positioning itself as a source of some of the most useful feature selection techniques. Typical testors play an important role when dealing with feature selection tasks [5,18] and have been used in solving practical problems like diagnosis of diseases [11], text categorization [12], document summarization [13] and document clustering [8].

The problem of finding the set of all typical testors in a basic matrix is an old problem that has had an important development during the last ten years. To support this statement, consider the number of papers presenting new algorithms related to this problem, for example [3,4,9,10,17]. All those algorithms are generally referred to as Typical Testor-finding Algoritms (*TTA*).

There are two classes of *TTA*: deterministic and meta-heuristic. Deterministic algorithms guarantee that they will find all typical testors in a given problem at the expense of an exponential time complexity. On the other hand, meta-heuristic algorithms have no guarantee to find all the typical testors in a given problem, but they are feasible to be used on extremely large search spaces where the time complexity of deterministic algorithms is simply unacceptable [16].

The complexity of deterministic *TTA* has not been sufficiently studied. This lack of sufficient study can be regarded as the cause of why most published works introducing new *TTA* fail, in the opinion of the authors of this paper, to properly justify their selection of basic matrices for comparative performance experimentation between different algorithms. On one hand, since the number of matrices selected for experimentation is considerably low, the obtained results lack statistical significance. On the other hand, by not using a specific strategy for comparatively testing algorithms, the characteristic behavior of each algorithm in the presence of certain stereotypical phenomena is not captured.

Fortunately a convenient strategy for selecting matrices for algorithm testing is certainly viable. In [1], a feasible strategy for generating test matrices was sketched for the first time, and in [2] it was used to benchmark some *TTA*. In the generated test matrices the set of typical testors can be determined in advance. This property allows the assessment of the computational behavior for the implementation of any deterministic *TTA*, as well as the validation of the answer completeness of any meta-heuristic *TTA*. Since both, the amount of typical testors and their length can be

☆ This paper has been recommended for acceptance by Eckart Michaelsen.

* Corresponding author. Tel.: +52 55 5729 6000x56553; fax: +52 5556681250.
E-mail address: sgodoyc@cic.ipn.mx (S. Godoy-Calderon).

http://dx.doi.org/10.1016/j.patrec.2016.04.020
0167-8655/

preset, generated test matrices can be targeted for studying some specific computational behavior by varying only one parameter at a time. For example, we can consider the exponential increase in the number of matrix rows with only a linear increase in the number of typical testors, or a linear increase in the number of matrix columns, resulting in a polynomial growth of the number of typical testors.

In this paper, we worked along two main directions. First, we significantly extended the previously presented theoretical framework to allow for the generation of a whole new set of test matrices that is more flexible and versatile. Second, we show how those matrices can be used to study the behavior of a *TTA* in the presence of specific phenomena. We also selected three previously published *TTA*, tested them against carefully selected test matrices, and discussed the obtained results.

The rest of this paper is structured as follows. In Section 2, the theoretical background for the generation of test matrices is set. Section 3, presents the *TTA* selected for experimentation, as well as some of their known properties. The complete set of experiments with all three *TTA* is presented in Section 4. Finally, we draw some important conclusions for new algorithm developers.

2. Theoretical background

Several, if not all of the research works in Testor Theory, handle a matrix that holds the information about the comparison of objects belonging to different classes within a certain supervision sample. That matrix is called a comparison matrix. Since using boolean comparison functions for constructing this matrix is a common practice, the result is a Boolean matrix. The comparison matrix is known as a difference matrix (*DM*) when each entry 0 means that there is a pair of objects, within the supervision sample, with the exact same value in the feature corresponding to the column of that entry, and each entry 1 means that the value, for the corresponding feature, is dissimilar in those objects.

Let $\mathcal{R}_{DM} = \{a_1, \ldots, a_m\}$ and $\mathcal{C}_{DM} = \{x_1, \ldots, x_n\}$ be the set of rows and the set of columns of a *DM*, respectively. $T \subseteq \mathcal{C}_{DM}$ is called a testor in *DM* if the submatrix $DM|_T$, obtained by eliminating from *DM* all columns not in the subset *T*, does not have any row composed exclusively by entries 0. Also, *T* is called a typical testor (irreducible testor) if no subset of *T* can be found to be also a testor in *DM*. According to the previous definition, a typical testor is a minimal set of features capable of describing all objects in the supervision sample, without causing confusion among those belonging to different classes.

A row r_p within a difference matrix is considered as sub-row of another row r_q if the following two conditions hold: each position of r_p holds a value less than or equal to the value in r_q at the same position, and there is at least one position where r_p has a value strictly less than the corresponding one in r_q. A row r_p in a difference matrix *A*, is called a *basic row* if it has no sub-rows within the same matrix.

To reduce a difference matrix, and take advantage of the last definition, for each *DM*, a basic matrix (*BM*) can be constructed which contains all and exclusively the basic rows from that *DM*. Moreover, since a *BM* has equal or less rows than its original *DM*, and it has been demonstrated that the set of all typical testors is exactly the same in both matrices, a great majority of testor-finding algorithms work on the *BM* instead of the *DM* [7,14].

Let $A = [a_{ij}]_{m\times n}$ and $B = [b_{ij}]_{m'\times n'}$ be two basic matrices. Then three crucial operators on pairs of matrices *A* and *B* ($\theta(A, B)$, $\gamma(A, B)$, and $\varphi(A, B)$) can be defined as follows:

1. The $\theta(A, B)$ operation produces a new matrix where each row in *A* is left-concatenated with each row in *B*, consequently having $m \times m'$ rows (the product of the number of rows in *A* and *B*), and also having $n + n'$ columns.
2. The $\gamma(A, B)$ operation creates a new matrix which has matrix *A* on its upper-left corner, followed by zeroes on all columns of *B*, and also has the *B* matrix on its lower-right corner, preceded by zeroes on all columns of *A*.
3. Finally, the result of a $\varphi(A, B)$ operation is a new Boolean matrix obtained by concatenating *A* and *B* if they have the same number of rows. The resulting matrix has exactly the same number of rows of *A* and *B*, but it has $n + n'$ columns (the sum of the number of columns from *A* and *B*).

Here are the formal specifications for all the above operations:

$$\theta(A,B) = \begin{bmatrix} a_{11\ldots}a_{1n} & b_{11\ldots}b_{1n'} \\ \vdots & \vdots \\ a_{11\ldots}a_{1n} & b_{m'1\ldots}b_{m'n'} \\ \vdots & \vdots \\ a_{m1\ldots}a_{mn} & b_{11\ldots}b_{1n'} \\ \vdots & \vdots \\ a_{m1\ldots}a_{mn} & b_{m'1\ldots}b_{m'n'} \end{bmatrix} \tag{1}$$

$$\gamma(A,B) = \begin{bmatrix} a_{11\ldots}a_{1n} & 0 \\ \vdots & \vdots \\ a_{m1\ldots}a_{mn} & 0 \\ \vdots & \vdots \\ 0 & b_{11\ldots}b_{1n'} \\ \vdots & \vdots \\ 0 & b_{m'1\ldots}b_{m'n'} \end{bmatrix} \tag{2}$$

and if $m = m'$ then

$$\varphi(A,B) = \begin{bmatrix} a_{11\ldots}a_{1n} & b_{11\ldots}b_{1n'} \\ & \vdots \\ a_{m1\ldots}a_{mn} & b_{m1\ldots}b_{mn'} \end{bmatrix} \tag{3}$$

The most important property of the φ, θ and γ operators is that, when applied to basic matrices, the resulting matrix is also basic, since they preserve the portion of the matrix that guarantees that rows are incomparable. Also, if their arguments are all basic matrices, then all three operators are associative. As a consequence, we will write $\varphi^N(A)$ to represent the resulting matrix of applying the φ operator over *N* matrices *A* (with $\varphi^1(A) = A$); we write $\theta^N(A)$ to represent the result of applying the θ operator *N* times consecutively (with $\theta^1(A) = A$); and we write $\gamma^N(A)$ when applying γ *N* times over matrix *A* (with $\gamma^1(A) = A$).

Now, let $\mathcal{C}_A = \{x_1, \ldots, x_n\}$ be the set of columns in basic matrix *A*, and let $x_j \in \mathcal{C}_A$. We will write $[x_j]_N$ to denote the class of all columns in $\varphi^N(A)$ exactly equal to x_j. In other words, $[x_j]_N = \{x_j, x_{j+n}, \ldots, x_{j+(N-1)n}\}$. Given $S \subseteq \mathcal{C}_A$ and $S = \{x_{j_1}, \ldots, x_{j_s}\}$, $[S]_N$ will denote the family of all subsets of columns from $\varphi^N(A)$ that can be obtained by replacing one or more columns in *S* with any other column in the same class, that is, $[S]_N = [x_{j_1}]_N \times \cdots \times [x_{j_s}]_N$. Then it is easy to verify that $|[S]_N| = N^{|S|}$.

Therefore, if *A* and *B* are basic matrices such that the sets $\Psi^*(A)$ and $\Psi^*(B)$ of all typical testors in matrices *A*, and *B* are known, then the next three propositions establish how the sets $\Psi^*(\varphi^N(A))$, $\Psi^*(\theta(A, B))$, and $\Psi^*(\gamma(A, B))$ can be analytically obtained:

Proposition 1. $\Psi^*(\varphi^N(A)) = \{[T]_N \mid T \in \Psi^*(A)\}$.

Proposition 1 states that the set of typical testors in a matrix *A*, concatenated *N* times with itself, is exactly the set of all classes of typical testors in *A*. This proposition can be proved by observing