Pattern Recognition Letters 41 (2014) 34–42

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

An evolutionary algorithm with acceleration operator to generate a subset of typical testors ☆

Guillermo Sanchez-Diaz [a,*], German Diaz-Sanchez [b], Miguel Mora-Gonzalez [c], Ivan Piza-Davila [d], Carlos A. Aguirre-Salado [a], Guillermo Huerta-Cuellar [e], Oscar Reyes-Cardenas [a], Abraham Cardenas-Tristan [a]

[a] Universidad Autonoma de San Luis Potosi, Dr. Manuel Nava 8, SLP, Mexico
[b] Centro de Investigacion y de Estudios Avanzados del I.P.N., Zapopan, Jal., Mexico
[c] Universidad de Guadalajara, Enrique Diaz de Leon 1144, Lagos de Moreno, Jal., Mexico
[d] Instituto Tecnologico y de Estudios Superiores de Occidente, Tlaquepaque, Jal., Mexico
[e] Diseño y Desarrollo Optomecatronico de Mexico, Capri 107, Leon, Gto., Mexico

ARTICLE INFO

Article history:
Available online 20 November 2013

Keywords:
Hill climbers
Feature selection
Typical testors
Pattern recognition

ABSTRACT

This paper is focused on introducing a Hill-Climbing algorithm as a way to solve the problem of generating typical testors – or non-reducible descriptors – from a training matrix. All the algorithms reported in the state-of-the-art have exponential complexity. However, there are problems for which there is no need to generate the whole set of typical testors, but it suffices to find only a subset of them. For this reason, we introduce a Hill-Climbing algorithm that incorporates an acceleration operation at the mutation step, providing a more efficient exploration of the search space. The experiments have shown that, under the same circumstances, the proposed algorithm performs better than other related algorithms reported so far.

1. Introduction

Data dimensionality reduction has become very important in machine learning over the past few decades. Many problems related to image processing, text mining and bioinformatics – among other disciplines – involve handling large datasets which instances can be described as a set of features.

A number of dimension–reduction techniques have emerged as a pre-processing step in tasks dealing with large datasets, such as: data analysis and supervised classification. Some of these techniques are about feature subset selection. The main difference between these techniques and other reduction techniques (like projection and compression) is that the first ones do not transform the input features, but they select a subset of them (Lozano et al., 2006).

Feature selection is a significant task in supervised classification and other pattern recognition areas. It identifies those features that provide relevant information for the classification process.

The problem of feature subset selection has been treated using metaheuristics (Inza et al., 1999; Kohavi and Jhon, 1997; Saeys et al., 2004), multi-objective point of view (Mierswa and Michael, 2006), etc. Nevertheless, results at this time are not conclusive.

Dmitriev et al. (1966) introduced the concept of test to pattern recognition problems. He defined a test as a subset of features that allows differentiating objects from different classes. This concept has been extended and generalized in several ways (Lazo-Cortes et al., 2001; Valev and Sankur, 2004).

In Logical Combinatorial Pattern Recognition approach (Martinez-Trinidad and Guzman-Arenas, 2001; Ruiz-Shulcloper et al., 2002), feature selection is addressed using Testor Theory (Lazo-Cortes et al., 2001).

In the eighties, Ruiz-Shulcloper introduced a typical testor characterization for computing all the typical testors of a training matrix, with object descriptions defined in terms of any kind of features, not only booleans (Bravo and Ruiz-Shulcloper, 1982; Ruiz-Shulcloper et al., 1983). The first algorithms to generate the entire set of typical testors of a training matrix were then developed (Ruiz-Shulcloper et al., 1985; Aguila-Feroz and Ruiz-Shulcloper, 1984; Bravo-Martinez, 1983).

The concept of testor and typical testor have also been used by V. Valev, under the names of descriptor and non-reducible descriptor, respectively (Valev and Zhuravlev, 1991).

Typical testors have been widely used in voting algorithms for object classification, based on partial-precedence determination (Ruiz-Shulcloper and Lazo-Cortes, 1999).

Besides, they have been used for evaluating the relevance of features on differential diagnosis of diseases (Ortiz-Posadas et al., 2001), and for estimating stellar parameters with remotely sensed

☆ This paper has been recommended for acceptance by J.Fco. Martínez-Trinidad.

* Corresponding author. Tel.: +52 444 8262330x6010.
E-mail address: guillermo.sanchez@uaslp.mx (G. Sanchez-Diaz).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.11.006

data (Santos et al., 2004). In addition, typical testors have been employed for: feature selection on natural-disaster texts classifications (Carrasco-Ochoa and Martnez-Trinidad, 2004), dimensionality reduction on image databases (Ochoa et al., 2008), text categorization (Pons-Porrata et al., 2007), and automatic summarization of documents (Pons-Porrata et al., 2003).

There are some real world problems which do not require the entire set of typical testors, but only a subset. Some examples include:

- Determination of risk factors associated to pregnant Mexican women (Torres et al., 2006). In this work, a problem of finding the most relevant features concerning neonatal morbidity on pregnant women is introduced. A genetic algorithm to find typical testors was used. Some of the features considered in this problem include: mother's age and weight, number of pregnancies, number of deliveries, bled, Apgar test within the first minute of the baby's life, and gestational age. The matrix employed to generate the typical testors has 32,768 rows and 29 columns.
- Determination of factors associated with Transfusion Related Acute Lung Injury (TRALI) (Torres et al., 2014). This paper describes the determination of informational weight of features related to TRALI, using a hybrid genetic algorithm for the identification of risk factors and the establishment of an assesment to each variable. In this problem, each typical testor denotes a set of features that best differentiates patients who will present TRALI from those who will not. The matrix used to generate the typical testors has 174 rows and 31 columns.
- Medical electrodiagnostic using pattern recognition tools (Lopez-Perez et al., 1997). This work introduces a medical diagnosis problem using neuroconduction studies, electromyography, signs and symptoms. The objects are assigned one of the following classes: lumbosacral radiculopathy, neuropathies, Guillain–Barre, myopathies, traumatic injuries of sciatic and Charcot-Marie-Tooth. This work used typical testors as support sets system, in the second step of a voting classification algorithm. The matrix used to generate the typical testors has 1,215 rows and 105 columns.

The number of rows of the matrix employed in the first example is too large. An algorithm capable to generate the whole set of typical testors takes several days.

The second example introduces a cut-off criterion for calculating the informational weight of features obtained from the generated typical testors. This criterion can be automatically calculated.

In the last example presented, the entire set of typical testors has not been found yet. The authors divided the matrix in three parts to find other typical testors, but without taking into account all features described in the problem. This fact affects the accuracy of the classification.

The computation of the entire set of typical testors requires exponential time (Valev and Asaithambi, 2003). In general, two approaches have been developed to address this problem: (a) algorithms that generate the entire set (LEX (Lexicographic Order Algorithm) (Santiesteban-Alganza and Pons-Porrata, 2003), CT_EXT (Complete elements extended) (Sanchez-Diaz and Lazo-Cortes, 2007), BR (binary operations) (Lias-Rodriguez and Pons-Porrata, 2009), and Fast-CT_EXT (Fast-Complete elements extended) (Sanchez-Diaz et al., 2010)); and (b) algorithms that find only a subset of typical testors (GA (Simple Genetic Algorithm) (Sanchez-Diaz et al., 1999), UMDA (Evolutionary Strategy) (Alba et al., 2000) and AGHPIA (Genetic algorithm with evolutionary mechanisms) (Torres et al., 2009)).

Nevertheless, these global-search heuristics become too slow as the number of features grows significantly. One reason is because the goal of this techniques is to reach the global maximum which, in this case, refers to the entire set of typical testors. However, each typical testor can be considered a local maximum for this particular problem.

This paper introduces a local-search heuristic based on the Hill-Climbing algorithm, that incorporates an acceleration operation, useful to find a subset of the entire set of typical testors. The goal of this Hill Climbing technique is to generate a single typical testor, iteratively, across the space search.

Preliminary results of this algorithm were presented in Diaz-Sanchez et al. (2011), but this work explains in detail typical-testor concepts, and shows experimentally the stability of the proposed algorithm when different values of its parameters are handled, using different basic matrices.

The classic concept of testor, in which classes are assumed to be both hard and disjointed, is used. The comparison criteria used for all features are Boolean, regardless of the feature type (qualitative or quantitative). The similarity function used for comparing objects demands similarity in all features. These concepts are formalized in the following section.

2. Background

Let $TM = \{O_1, O_2, \ldots, O_m\}$ be a training matrix containing m objects, each belonging to a class $K_i \in \{K_1, K_2, \ldots, K_c\}$, described in terms of n features $R = \{x_1, x_2, \ldots, x_n\}$. Each feature $x_i \in R$ takes values in a set $M_i, i = 1, \ldots, n$. A comparison criterion of dissimilarity $D_i : M_i \times M_i \rightarrow \{0, 1\}$ is associated to each x_i (0 = similar, 1 = dissimilar) (Diukova, 1976; Ruiz-Shulcloper et al., 1980).

An example of training matrix which was taken from Valev and Sankur (2004) is the following:

Example. A medical doctor can tell whether a patient suffers from a step throat or from a flu by the presence or absence of the following symptoms: sore throat, cough, cold and fever.

In this example, patients are the objects $(O_1, O_2, \ldots, O_7)$, symptoms are the features (x_1, x_2, x_3, x_4), and diseases are the classes (K_1, K_2).

The training matrix (shown in Table 1) stores the information of seven patients; the first two suffers from strep throat (class K_1), and the last five suffers from a flu (class K_2).

Each row in the training matrix denotes the presence (1) and absence (0) of every symptom on a patient.

Definition 1. If a feature subset $T \subseteq R$ allows to distinguish objects belonging to different classes, then T is called a testor (or descriptor) (Dmitriev et al., 1966).

Definition 2. If a given testor T, does not allow to distinguish objects belonging to different classes after removing any attribute $x_i \subset R$, then T is called typical testor (or non-reducible descriptor), and it is denoted by TT (Dmitriev et al., 1966).

Table 1
Training matrix of patients.

Objects	x_1	x_2	x_3	x_4	Class
O_1	1	1	0	0	K_1
O_2	1	0	1	0	K_1
O_3	0	0	1	1	K_2
O_4	1	0	1	1	K_2
O_5	0	0	1	0	K_2
O_6	0	1	1	0	K_2
O_7	0	1	1	1	K_2