Computación y Sistemas Vol. 6 No.1 pp. 062 - 066
© 2002, CIC - IPN. *ISSN 1405-5546* Impreso en México

PH. D. THESIS ABSTRACT

Sensitivity Analysis in Logical-Combinatorial Pattern Recognition

Análisis de Sensibilidad en Reconocimiento Lógico Combinatorio de Patrones

Graduated: Jesús Ariel Carrasco Ochoa
Graduated on January 12, 2001
Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro No. 1, Sta Ma. Tenanzintla, Puebla
C.P. 72840, Apdo. Postal 51 y 216, 72000 México
e-mail: ariel@inaoep.mx
Advisor 1: José Ruiz Shulcloper
Instituto de Cibernética, Matemática y Física, Cuba
e-mail: recpat@cidet.icmf.inf.cu
Advisor 2: Juan Luis Díaz de León Santiago
Centro de Investigación en Computación-IPN, México
e-mail: jdiaz@cic.ipn.mx

Abstract

In Pattern Recognition problems (feature selection, supervised classification, unsupervised classification, etc.) data modifications are very common. This due to many reasons, for example: data acquisition error fixing; specialist changes sample classification; elimination of some data; new data; etc. Then, many times it is necessary to repeat some calculations in order to include this data modifications. This could be too expensive, in the computational sense, depending on the complexity of the recognition algorithm used. Because of this, it will be very useful if you can adjust already obtained results to the modified data, without have to apply the recognition algorithm again.

In this work, we study some logical combinatorial pattern recognition problems (Martínez-Trinidad and Guzmán-Arenas, 2001; Ruiz-Shulcloper et al., 1999) to find how do results change, when data is modified. We call this process Sensitivity Analysis. Sensitivity Analysis has as objective to find methods to adjust results when data is modified, but with lower complexity than original algorithm. We do this study for Zhuravlev typical testors, typical ε:testors, Goldman typical testors and crisp and fuzzy connected sets.

Keywords: Logical combinatorial pattern recognition, dynamic data, testor theory, unsupervised classification.

Resumen

En problemas de reconocimiento de patrones (selección de variables, clasificación supervisada, clasificación no supervisada, etc.) frecuentemente se presentan modificaciones en los datos. Esto debido a diversas razones, por ejemplo: errores en la adquisición de los datos; reconsideración por parte del especialista en cuanto a la clasificación de algunos objetos; eliminación de algún dato; aparición de nuevos datos; etc. Por lo cual, en muchas ocasiones es necesario repetir los cálculos para incluir dichas modificaciones. Esto puede ser muy costoso computacionalmente, dependiendo de la complejidad del algoritmo de reconocimiento utilizado. Debido a esto, resultaría de gran utilidad poder ajustar los resultados, ya obtenidos, a las nuevas condiciones (fruto de las modificaciones) sin tener que aplicar el algoritmo original sobre los datos modificados.

In este trabajo estudiamos algunos problemas dentro del reconocimiento lógico combinatorio de patrones (Martínez-Trinidad y Guzmán-Arenas, 2001; Ruiz-Shulcloper et al., 1999) para encontrar como cambian los resultados cuando los datos son modificados. Nosotros llamamos a este proceso Análisis de Sensibilidad. El proceso de Análisis de Sensibilidad tiene como objetivo encontrar métodos para ajusta los resultados cuando los datos son modificados, pero con una complejidad menor que la del algoritmo original. Este estudio lo realizamos para los testores típicos de Zhuravlev, ε:testores típicos, testores típicos de Goldman y componentes conexas duras y difusas.

Palabras clave: Reconocimiento lógico combinatorio de patrones, Datos dinámicos, Teoría de testores, Clasificación no supervisada.

1 Introduction

A pattern recognition algorithm is an algorithm to solve a problem of feature selection, supervised classification, partially supervised classification or unsupervised classification.

Suppose $A(\{p_1,\ldots,p_u\})$ is a pattern recognition algorithm with parameters $p_1,\ldots,p_u$, where the parameter set could be empty. Let M be a data set and let R be the result of applying $A(\{p_1,\ldots,p_u\})$ to M. This will be denoted by $A(\{p_1,\ldots,p_u\})(M)=R$.

We say that the *sensitivity of* R is the way R changes because of modifications to M, with fixed $A(\{p_1,\ldots,p_u\})$, or fixing M and changing some of A's parameters.

We call *Sensitivity Analysis* to the process of searching for methods to adjust R when M or some of the parameters change.

We do sensitivity analysis in testor theory for Zhravlev's typical testors, typical ε:testors and Goldman's typical testors. In all cases, behavior of the set of all typical testors was described by mean of a set of theorems. We propose and prove a theorem for describing the behavior of the set of all typical testors for each one of the different types of modification. Based on this theorems we define a method to adjust the set of all typical testors for each kind of modification. Complexity is analyzed and some experimental tests was done.

The solution of the sensitivity analysis for the case of data addition, allows us to define an incremental algorithm. So, also a new incremental algorithm for each kind of typical testor is defined. We do some experimental tests with these incremental algorithms.
Additionally, we do sensitivity analysis for the clustering problem. We found some results for crisp and fuzzy connected sets. A new method to deal with modifications is described and some experimental tests are done.

2 Sensitivity Analysis for Zhuravlev's Typical Testors

Let U be a universe of objects structured in classes. Suppose that a sample of each class is available. Each object in U is described in terms of $R=\{x_1,\ldots,x_n\}$, the feature set used to study these objects. As a description of an object O we understand the n-uple $I(O)=(x_1(O),\ldots,x_n(O))$ where $x_i(O)$ is the value of the feature x_i at the object O, with i=1,...,n. Analogously, as a subdescription of O in terms of features $x_{i_1},\ldots,x_{i_s}$ we understand the s-uple $(x_{i_1}(O),\ldots,x_{i_s}(O))$. From here on, we shall use the terms "object description" and "object" interchangeably. Similarly, we shall denote I(O) and O to refer to the description of the object O.

Suppose that a training sample has a matrix representation M, that is, object descriptions are stored in a matrix with as many columns as features and as many rows as objects in the sample. These rows are the object descriptions and they are grouped in r disjoint classes.

We use the next extended Zhuravlev's testor definition.

Let M be a training matrix with m rows, grouped in r classes (not necessarily disjoint), $K'_1,\ldots,K'_r$, $r\geq 2$, and n columns. The set $t=\{x_{i_1},\ldots,x_{i_s}\}\subseteq R$ is a testor of M if and only if after eliminating from M all columns except the ones in t, new equal subdescriptions in different classes do not appear.

From the set of all testors of a matrix M, there are some testors, which are *irreducible*. That is: if any feature is eliminated from them, then they stop being testors. That means, they confuse objects belonging to different classes.

These testors are called *typical testors*.

From here on we shall use x_i to refer to both the column with index "i" of M and the feature x_i.

In order to speed up algorithms for calculating all typical testors, we introduce some additional concepts and a new testor characterization.

Let C_i be a Boolean comparison criteria for the feature x_i, i=1,...,n. such that:

$$C_i(x_i(O_p),x_i(O_q))=\begin{cases}0 & \text{if } x_i(O_p) \text{ and } x_i(O_q) \text{ coincide}\\ 1 & \text{in other case}\end{cases}$$

The term *difference matrix* of M, is applied to a Boolean matrix MD composed by rows containing at least one 1, which is constructed as follows:

$$S_{ij}=(\alpha_1^{ij},\ldots,\alpha_n^{ij}),\ i\neq j,\ i,j,=1,\ldots,m,$$

where $\alpha_p^{ij}=C_p(x_p(O_i),x_p(O_j))$, p=1,...,n, and if $O_i\in K_s$, then $O_j\notin K_s$, s=1,...,r. The cardinal number of MD, denoted by m', is given as:

$$m'=\sum_{i=1}^{r-1}\sum_{t=i+1}^{r}|K'_i||K'_t|$$

where $|K'_i|$ is the number of objects of M belonging to K_i, i=1,...,r.

Proposition 2.1.- Feature set $\tau=\{x_{i_1},\ldots,x_{i_s}\}$ is a testor of M if and only if after eliminating from MD all columns except those of τ, there is no row with only zeros.

The proof is immediate. Indeed, if there is no row with only zeros, it means that there is no pair of objects belonging to different classes that coincide on all features from τ, and this is the testor definition.

This proposition is a testor characterization based on the difference matrix concept.

We denote by T(M) (TT(M)) the set of all (typical) testors of M. Analogously, we denote by T(MD) (TT(MD)) the set of all (typical) testors of MD.

Corollary 2.1.- T(M) = T(MD).

Corollary 2.2.- TT(M) = TT(MD).

It is clear that the search for all (typical) testors from MD has some advantages:

1. The comparison between two objects is made only once;
2. The search is made over a Boolean Matrix;
3. This testor characterization (proposition 2.1) could be programmed more efficiently than Zhuravlev's definition (extended).

Nevertheless, there is a difficulty, the number of rows in MD is quadratic whit respect to the number of rows in M. So, even though the search on MD is more advantageous than the search on M, m' is a very large number. In order to avoid this difficulty, we introduce a process that reduces both m' and the number of ones in MD.

If p and t are two rows of MD, then we say that p is a *sub-row* of t if and only if:

a) $\forall j\ (a_{pj}=1 \Rightarrow a_{tj}=1)$ (t has 1 everywhere p has 1)

b) $\exists k\ (a_{tk}=1 \wedge a_{pk}=0)$ (there is at least a column such that t has 1 and p has 0)

also we say that t is a *super-row* of p.

Let t be a row of MD. Then, t is called a *basic row* of MD if and only if MD does not have any row t' such that t' is a sub-row of t.

Given a difference matrix MD, the matrix MB that has exclusively basic rows of MD (without repetition) is called *basic matrix*.

Obviously, MB is also a Boolean matrix. Although reduction is not always the same (because it depends on distribution of ones in the difference matrix) we have found matrices with 32 and 125 rows, whose basic matrices have 4 and 15 rows respectively. This gives us an idea about how many rows could be eliminated. Likewise the number of ones in MB is smaller than in MD.

It is important to say that the process of calculating MD and MB from M is very fast, and using MB is an actual advantage for calculating all typical testors.
However, if we want to use MB to calculate all typical testors, first we must prove that the elimination of not basic rows of MD does not change the set of all testors, and as a consequence, neither does it change the set of all typical testors. This is shown in the next proposition and its corollaries.

Proposition 2.2.- T(MD) = T(MB), where T(MB) is the set of all testors of MB.

Proof.-
Obviously, if MB=MD, then T(MD)=T(MB). Then, suppose that MB≠MD.

a) Let t∈T(MD) be a testor of MD, then after eliminating all columns from MD except ones of t, there is no row with only zeros. But MB's rows are a subset of MD's, then neither does any row with only zeros appear in MB, and so t∈T(MB).

b) Let t∈T(MB) be a testor of MB, then after eliminating all columns from MB except ones of t, there is no row with only zeros. Like MB≠MD, there is a row f∈MD\MB, that is, a row of MD such that it is not a basic row. Then there is a row f'∈MB such that f' is a sub-row of f. It tells us that f has 1 wherever f' has 1 and at least in one other position. Since f'∈MB, f' has not only zeros in all positions corresponding to the features of t, and so neither has f. Note that it is for any row f∈MD\MB, from which we can deduce that in MD no row with only zeros appears either, after eliminating all columns except ones of t, that is, t∈T(MD).•

Corollary 2.3. - TT(MD) = TT(MB), where TT(MB) is the set of all typical testors of MB.

Based on the basic matrix concept, many algorithms to calculate the set of all typical testors have been developed, but all of them has exponential complexity in the worst case. Then, it is very expensive to apply any of these algorithms each time that a modification occurs. So, we do Sensitivity Analysis. First we analyze which kind of modifications can appear.
All possible alterations to a training matrix M can be summarized in 4 cases: delete a column; add a column; delete a row; add a row, or in successive compositions of them. For this reason, we only study how does the set of all typical testors change because of each of this 4 types of alterations. Figure 1 shows these cases and their effect over the basic matrix MB.

Alteration on M	Effect on MB
Delete a feature	The corresponding column is eliminated. If some row quits being basic, it is eliminated.
Add a feature	The corresponding column is added. If new basic rows appear, they are added.
Delete an object	Rows that come from the comparisons with the deleted object are eliminated. If new basic rows appear, they are added.
Add an object	Rows that quit being basic, when new rows appear in MD, are eliminated. If new basic rows appear, they are added.

Fig. 1. Possible alterations on M and their effects on MB

An immediate consequence of the typical testor concept is the following:
Let $t=\{ x_{i_1},..., x_{i_s} \}$ be a testor of MB. t is a typical testor of MB if and only if there is in MB a set of s rows (associated to t) $F_t=\{ f_{i_1},..., f_{i_s} \}$ such that the submatrix constructed only with these rows and columns is a diagonal matrix save for permutations. That is, each row and each column have only one 1. So if any column is eliminated then appears a row with only zeros. Given a t, F_t could not be unique. The set of the entire possible F_t is denoted by $\mathbb{F}_t$.

Without loss of generality, we are going to consider to the basic matrix as a set of rows, and each row as a set of those features, which there is a 1 in the corresponding column.

Behavior of the set of all typical testors when there is a change on the basic matrix is shown in the following theorems.

Theorem 2.1.- Let MB' be the resultant basic matrix after eliminating the column x_i from MB, then $T'=T\backslash T_i$ where:

T' is the set of all typical testors of MB'
T is the set of all typical testors of MB
T_i is the set of all typical testors of MB, which contain to x_i.

Theorem 2.2.- Let MB' be the resultant basic matrix after adding the column x_{n+1} to MB, then $T'=T\cup T'_{n+1}$ where:

T' is the set of all typical testors of MB'
T is the set of all typical testors of MB
T'_{n+1} is the set of all typical testors of MB which contain to x_{n+1}.

Let MR be the resultant matrix after eliminating a row f_i from a basic matrix MB and all columns that have 1 in f_i. MR is not necessarily a basic matrix, but MR contains a basic matrix, we denote this basic matrix as MB''.

Theorem 2.3.- Let MB' be the resultant basic matrix after eliminating the row f_i from MB, then $T'=T\backslash T_0 \cup T_1$ where:

T' is the set of all typical testors of MB'
T is the set of all typical testors of MB
$T_0 = \{t\in T|$ there is a set $F_t\in \mathbb{F}_t$ with $f_i\in F_t$ and there is not a row $f_j\neq f_i$ such that $(F_t\backslash\{f_i\})\cup\{f_j\}\in \mathbb{F}_t\}$
T_1 is the set of all typical testors of MB''

Theorem 2.4.- Let MB' be the resultant basic matrix after adding the row f_{k+1} to MB, then $T'=T\setminus T_0 \cup T_1$ where:

T' is the set of all typical testors of MB'

T is the set of all typical testors of MB

$T_0 = \{t \in T \mid t \cap f_{k+1} = \emptyset\}$

$T_1 = \{t' \mid t' = t \cup \{x_i\},\ t \in T_0,\ x_i \in f_{k+1},$ there is no $t_0 \in T\setminus T_0$ with $t_0 \subseteq t'\}$

Proofs for these theorems can be found in (Carrasco-Ochoa, 2001)

Based on theorem 3.4 we define a new incremental algorithm, adding all of the rows of MB, one at each time.

2.1 Experimental Tests

In order to show the performance of the proposed methods (Incremental Algorithm and Sensitivity method) its runtimes were compared with CC and CT runtimes. CC and CT was algorithms with the best performance (Sánchez-Díaz, 1997). Tests was done with matrices from 15 to 30 features. Figure 2 shows results for the case of row addition.

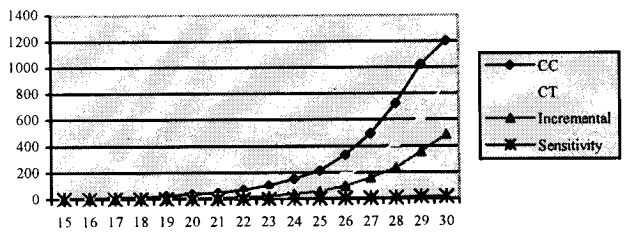

Fig. 2. Graph for runtimes of CC, CT, new Incremental algorithm and Sensitivity method for row addition

3 Sensitivity Analysis for Other Problems

We do Sensitivity Analysis for typical ε:testors and Goldman typical testors (Lazo-Cotés *et al.*). In both case we describe behavior using theorems for each kind of modification. Also we define a new incremental algorithm based on the row addition theorem. Again, experimental tests was done with excellent results.

Additionally, Sensitivity analysis was done for fuzzy and crisp connected sets (Carrasco-Ochoa, 2001). In this case we propose sensitivity methods based on depth-first spanning trees and graph theory.

4 Conclusions

The main result of this work is the fact that the new sensitivity methods reduce the needed work to calculate the new set of all typical testors, when there are modifications on the basic matrix. This, for Zhuravlev's typical testors, typical ε:testors and Goldman typical testors. Also it is possible to adjust connected sets after modifications.

Now we are working on Sensitivity Analysis for compact sets and strong compact sets.

5 References

Carrasco-Ochoa J. A., Sensitivity Analysis in Logical-Combinatorial Pattern Recognition, PhD. Tesis, CIC-IPN Mexico, 2001 (In Spanish).

Martínez-Trinidad J. F. and **Guzmán-Arenas A.**, The logical combinatorial approach to pattern recognition, an overview through selected works, Journal of Pattern Recognition, 34/4 1-11, (2001).

Ruiz-Shulcloper J., **Guzmán-Arenas A.** and **Martínez-Trinidad J. F.**, Logical Combinatorial Pattern Recognition I. Feature Selection and Supervised Classification, Advances in Pattern Recognition Series, IPN, ISBN 970-18-2384-1, 1999 (In Spanish).

Sánchez Díaz, G., Develop and programming efficient algorithms (Secuential and Parallel) to calculate typical testors of a basic Matrix, MS. Tesis, Benemérita Universidad Autónoma de Puebla, 1997 (In Spanish).

Lazo-Cortés M., Ruiz-Shulcloper J. and **Alba-Cabrera E.,** *An overview of the concept testor*, (To appear in Pattern Recognition).

Jesús-Ariel Carrasco-Ochoa, *was born in Mexico in 1965. He received his PhD degree in Computer Science form the Center for Computing Research of the National Polytechnic Institute (CIC-IPN), Mexico, in 2001. He works as associated researcher at the National Institute for Astrophysics, Optics and Electronics of Mexico. His current research interests include Sensitivity Analysis, Logical Combinatorial Patter Recognition, Testor Theory, Feature Selection and Clustering*

José Ruiz-Shulcloper, *was born in 1948 in Havana, Cuba. He received the BS in Mathematics from the Havana University, and his Ph. D. degree in Mathematics at the Moscow State University, Lomonosov and the Center of Calculus of the Sciences Academy from USSR in 1978. He was awarded with the National Medal "P. Miguel" of the Cuban Society of Mathematics and Computer Sciences (SCMC) in 1980 and with the Research National Prize of the Cuban Academy of Sciences in 1998. He had been member of the National Committee of SCMC since 1982, member of the Scientific Council of the Institute of Cybernetics, Mathematics and Physic (ICIMAF) (1980-1992), member of the National Scientific Council of the Cuban Academy of Sciences (1983-1987), member of the Scientific Council of the International Mathematical Center "S. Banach" in Poland (1981-1991). He was Associate Professor at the Havana University, Invited Research Professor in several institutions in Latin America, Europe, and the United States. Since 1978 he has been the leader of the Logical Combinatorial Pattern Recognition Group of the ICIMAF. He had published more than 90 papers in journals and proceedings, and also he is author of several monographs, textbooks, and others publications. He was the founder of the Logical Combinatorial Pattern Recognition Groups in several Cuban and Mexican institutions. In this moment, he is Senior Research and Leader of the Pattern Recognition Group at the ICIMAF, Cuba, and Associated Editor of the Journals Ciencias Matemáticas, and Computación y Sistemas. His current research interests include conceptual clustering, symbolic objects, fuzzy models of Pattern Recognition, Testor Theory, and Data and Text Mining.*

Dr. Juan Luis Díaz de León *, was born in Veracruz, Ver., México. He is an engineer from the Regional Institute of Technology at Veracruz. He obtained his MSc. and PhD. on Electrical Engineering at CINVESTAV-IPN. He was invited Profesor at the University of Florida in Gainesville, USA, at the Computer Sciences and Engineering Department. He was Director of Technology at the National Public Security Sistem for the Mexican Goberment. He is Full Professor and the actual Director of the Center for Computing Research in the IPN. In 1998 he was awarded with the National Prize for Research and Technological Development of Exelence "Luis Enrique Erro". In 1999 he was awarded with the Prize of Culture by the Goberment of Veracruz. He has received many others institutional and regional awards due to his achievements in several fields of science.*
His main interests are: Computer Vision, Artificial Inteligence, Automatic Control, Mobile Robotics and Pattern Recognition.