PROMiDAT IBEROAMERICANO

Programa Iberoamericano de Formación en Minería de Datos

Aplicación del Análisis de Datos Simbólico a la Detección de Inusualidades

Dr. Oldemar Rodríguez Rojas

Inusualidad o atipicidad

Objetos Simbólicos:

- “Real-life objects are too ***complex*** to be represented by points in a vectorial space” [Bock&Diday, 2000]

- “Symbolic objects overcome this limitation by representing ***concepts*** rather than individuals” [Bock&Diday, 2000]

- “Knowledge extraction from large data bases is our main aim as in Data Mining“ [Diday, 1998]

- En análisis simbólico de datos transforma la Minería de Datos una herramienta para “Big Data”.

Tabla clásica:

Example 4.1.2: A classical data matrix with mixed variables

The classical data matrix in Table 4.2 describes characteristics of $N = 6$ students in terms of $p = 4$ variables:

Person u	Height Y_1	Weight Y_2	Subject Y_3	Sex Y_4
Anna	1.70	65.9	*math*	1
Berta	1.62	61.4	*med*	1
Claudia	1.65	60.1	*phys*	1
Daniel	1.75	77.3	*math*	0
Eric	1.80	74.3	*med*	0
Fred	1.68	67.0	*econ*	0

Table 4.2: *A classical data matrix* $\underset{\sim}{X} = (x_{uj})$ *with mixed variables*

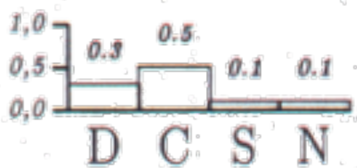
Symbolic Data Table:

$$\underline{X} = \left(\begin{array}{c|cccc|l} (80,100] & (\mathrm{D}\ 0.4; & \mathrm{C}\ 0.3; & \mathrm{S}\ 0.2; & \mathrm{N}\ 0.1) & \{\mathrm{CL},\ \mathrm{BNP}\} \\ (100,130] & (\mathrm{D}\ 0.1; & \mathrm{C}\ 0.3; & \mathrm{S}\ 0.4; & \mathrm{N}\ 0.2) & \{\mathrm{SPK},\ \mathrm{DB},\ \mathrm{BNP}\} \\ (8,10] & (\mathrm{D}\ 0.3; & \mathrm{C}\ 0.5; & \mathrm{S}\ 0.1; & \mathrm{N}\ 0.1) & \{\mathrm{SPK}\} \\ (10,13] & (\mathrm{D}\ 0.3; & \mathrm{C}\ 0.3; & \mathrm{S}\ 0.3; & \mathrm{N}\ 0.1) & \{\mathrm{CL},\ \mathrm{DB},\ \mathrm{BNP}\} \end{array} \right) \begin{array}{l} \leftarrow x_1' \\ \leftarrow x_2' \\ \leftarrow x_3' \\ \leftarrow x_4' \end{array}$$

where, e.g., the third row

$$x_3' = ((8,10],\ (\mathrm{D}\ 0.3;\ \mathrm{C}\ 0.5;\ \mathrm{S}\ 0.1;\ \mathrm{N}\ 0.1),\ \{\mathrm{SPK}\})$$

describes the town a_3 and could be equivalently displayed in a semi–graphical form such as

$x_3' = \Big((8,10],$  $, \{\mathrm{SPK}\}\Big)$ ← Symbolic Object

Creando tablas simbólicas:

Classical description of Schools

Schools	Town	Nb of pupils	Kind	Level
Jaurès	Paris	320	Public	1
Condorcet	Paris	450	Public	3
Chevreul	Lyon	200	Public	2
St Hélène	Lyon	380	Private	3
St Sernin	Toulouse	290	Public	1
St Hilaire	Toulouse	210	Private	2

Symbolic description of the towns by the schools variables

Town	Nb of pupils	Kind	Level
Paris	[320, 450]	(100%)Public	{1, 3}
Lyon	[200, 380]	(50%)Public , (50%)Private	{2, 3}
Toulouse	[210, 290]	(50%)Public , (50%)Private	{1, 2}

¿Cómo se construyen las tablas simbólicas?

Análisis Clásico

Datos

Millones…

Id-trx	*Causal*	*Sucursal*	*Monto*	*# Tarjeta*
3457	*36*	*Curridabat*	*2,500.00*	*1000*
1251	*28*	*San Pedro*	*1,750.00*	*1001*
3245	*39*	*Grecia*	*2,400.00*	*1000*
7635	*35*	*San Pedro*	*1,900.00*	*1001*
3245	*35*	*Alajuela*	*1,850.00*	*1001*
5367	*27*	*Alajuela*	*1,900.00*	*1002*
6486	*34*	*Heredia*	*1,600.00*	*1002*

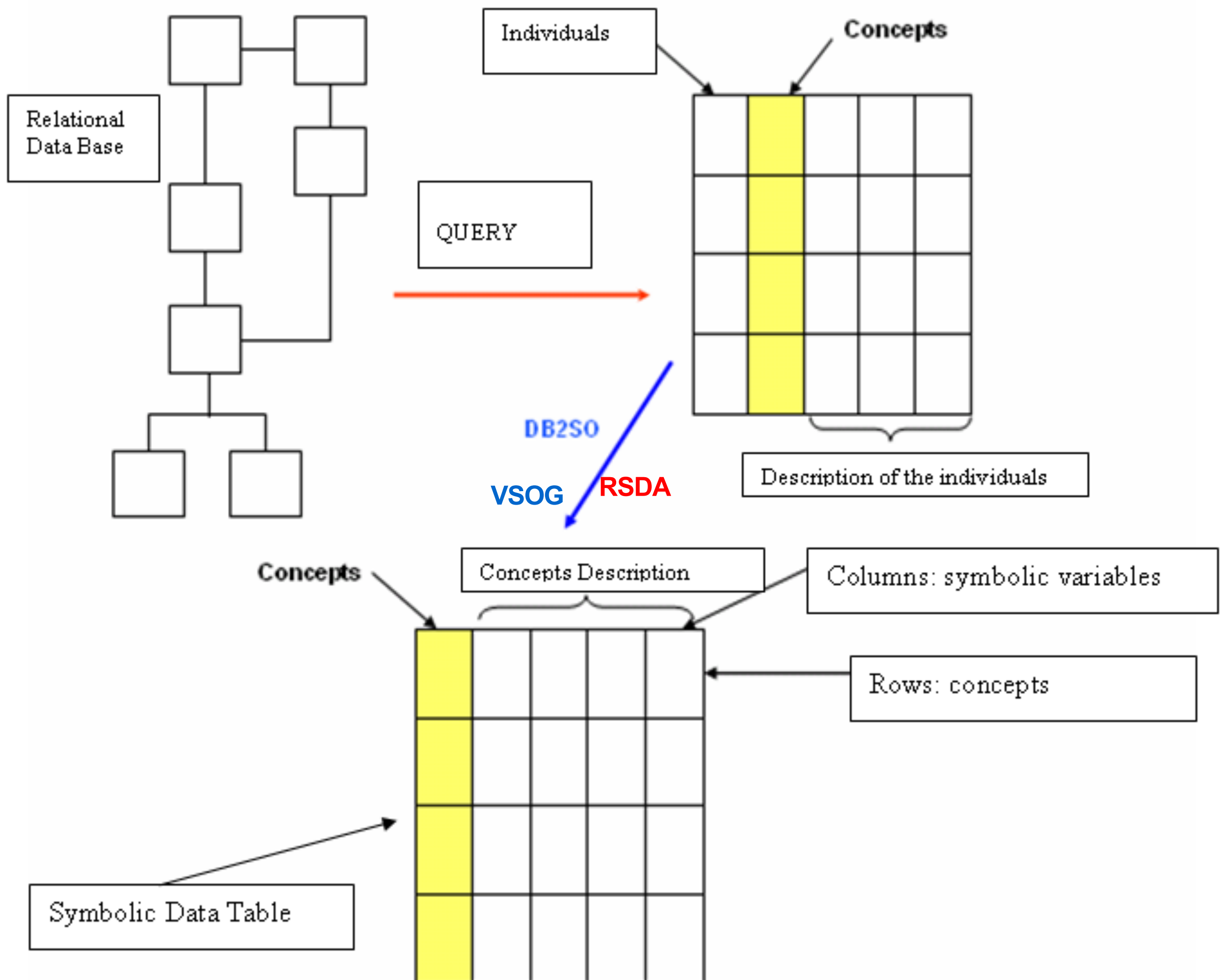
Análisis Multivariado Simbólico

Conceptos

Cientos…

# Tarjeta	*Causal*	*Sucursal*	*Monto*	**Salario**
1000	36(1/2),39(1/2)	*{Curr-50%,Gre-50%}*	[2.4,2.5]	255.4
1001	28(1/3),35(2/3)	*{SP-66%,Al-33%}*	[1.75,1.9]	122,2
1002	27(1/2),34(1/2)	*{Al-50%,Her-50%}*	[1.6,1.9]	534,5

From a Relational Database to a symbolic data table:



Symbolic Data Base:

Relational Data Base

100% knowledge

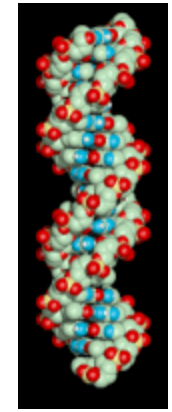
15 Gigabyte

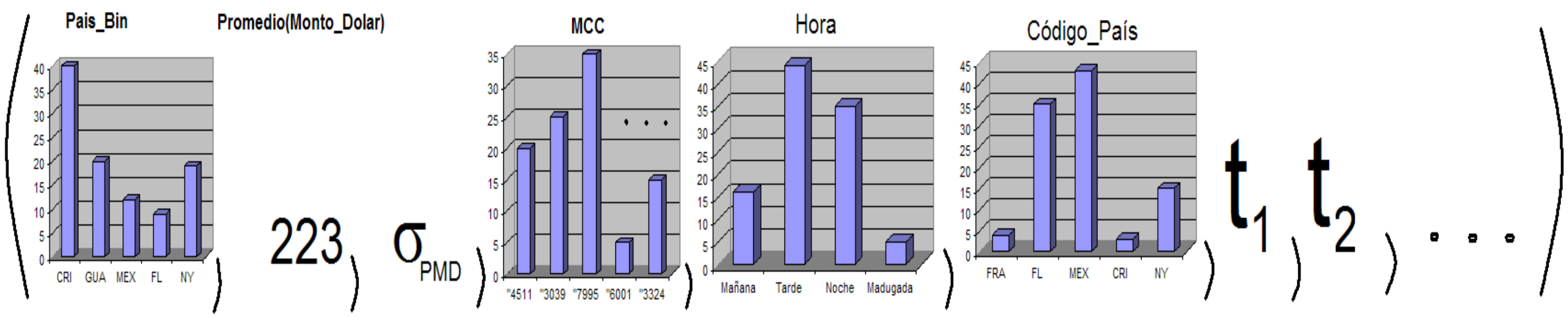
Symbolic Data Base

90 % knowledge

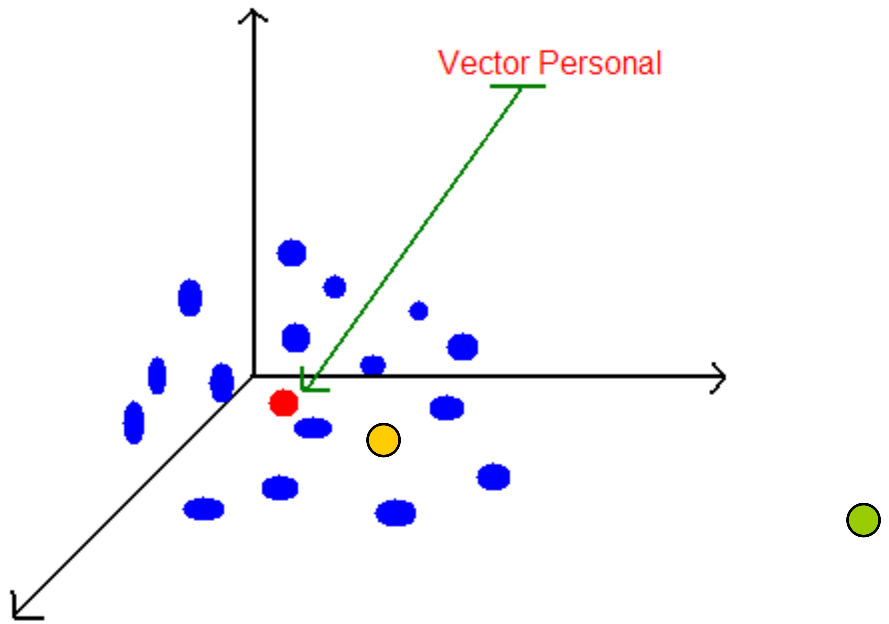
10.3 Megabyte

Perfil simbólico

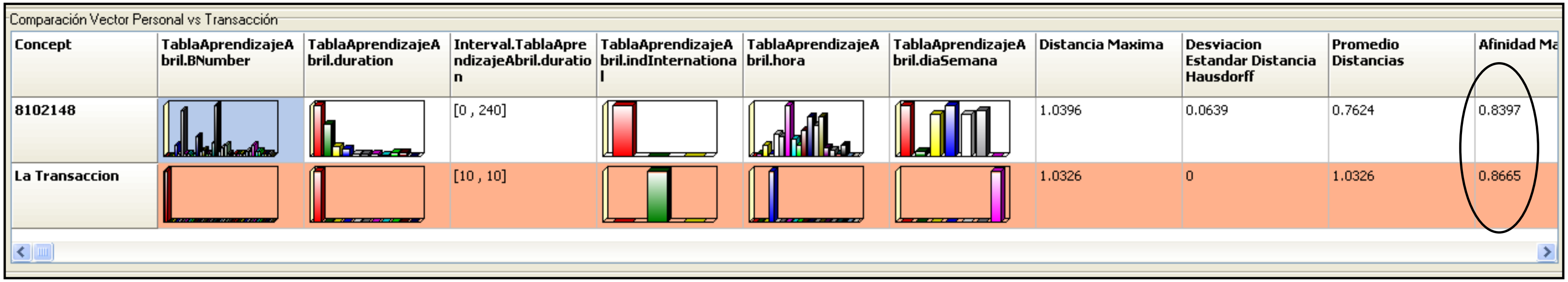




Espacio con el Perfil Simbólico y sus Transacciones



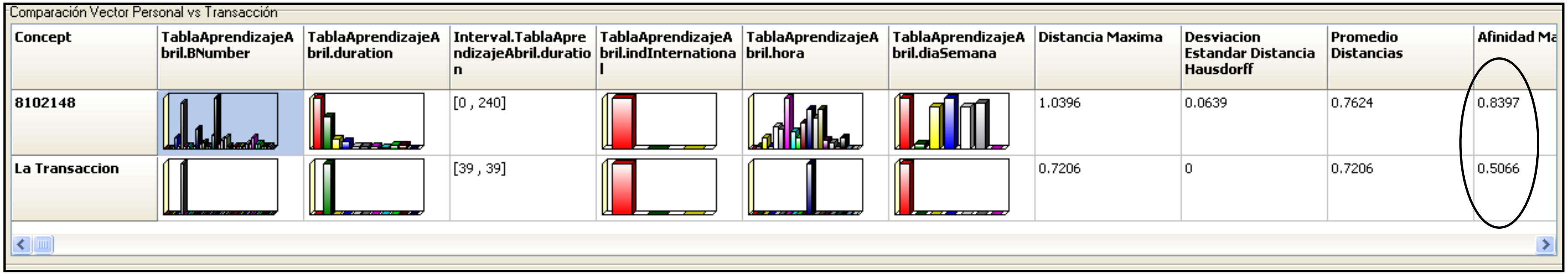
Transacción Inusual



Comparación Vector Personal vs Transacción

Concept	TablaAprendizajeAbril.BNumber	TablaAprendizajeAbril.duration	Interval.TablaAprendizajeAbril.duration	TablaAprendizajeAbril.indInternational	TablaAprendizajeAbril.hora	TablaAprendizajeAbril.diaSemana	Distancia Maxima	Desviacion Estandar Distancia Hausdorff	Promedio Distancias	Afinidad Ma
8102148			[0 , 240]				1.0396	0.0639	0.7624	0.8397
La Transaccion			[10 , 10]				1.0326	0	1.0326	0.8665

Transacción Usual



Comparación Vector Personal vs Transacción

Concept	TablaAprendizajeAbril.BNumber	TablaAprendizajeAbril.duration	Interval.TablaAprendizajeAbril.duration	TablaAprendizajeAbril.indInternational	TablaAprendizajeAbril.hora	TablaAprendizajeAbril.diaSemana	Distancia Maxima	Desviacion Estandar Distancia Hausdorff	Promedio Distancias	Afinidad Ma
8102148			[0 , 240]				1.0396	0.0639	0.7624	0.8397
La Transaccion			[39 , 39]				0.7206	0	0.7206	0.5066

Distancias entre Objetos Simbólicos

FIG. 4.3 – *Comparaison d'intervalles*

Distancias entre Objetos Simbólicos

Notation	Intervalle	Qualitatif
A_k	$[a_l, a_u]$	$a_1, \ldots$
B_k	$[b_l, b_u]$	$b_1, \ldots$
inters ou $\|A_k - -B_k\|$	longueur $(\mathrm{A}_k \cap \mathrm{B}_k)$	card $(\mathrm{A}_k \cap \mathrm{B}_k)$
l_s ou $\|A_k + +B_k\|$	$\|max(a_u, b_u) - min(a_l, b_l)\|$	$l_a + l_b - inters$
l_a ou $\|A_k\|$	$\|a_u - a_l\|$	card(A_k)
l_b ou $\|B_k\|$	$\|b_u - b_l\|$	card(B_k)
$\|U_k\|$	longueur du domaine	cardinal du domaine

Mesure proposée par Ichino [Ichino88]

En reprenant les notations du tableau 4.1 (page 146), la mesure s'écrit de la façon suivante :

$$\Phi(A_k, B_k) = |A_k + +B_k| - |A_k - -B_k| + \gamma(2|A_k - -B_k| - |A_k| - |B_k|)$$

Distancias entre Objetos Simbólicos

Distance de Hausdorff [Bandemer et Nather92]

Entre deux ensembles A et B, la distance s'écrit :

$$D_h(A,B) = \max\{\sup_{a\in A}\inf_{b\in B} d(a,b), \sup_{b\in B}\inf_{a\in A} d(a,b)\}$$

La mesure du Khi2 est la suivante:

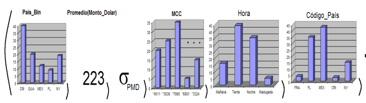
$$d(1,2) = \sum_{k=1}^{p}\left(\sum_{i=1}^{q}\frac{1}{f_{ik}^1+f_{ik}^2}\left(\frac{f_{ik}^1}{\sum_{i=1}^q f_{ik}^1} - \frac{f_{ik}^2}{\sum_{i=1}^q f_{ik}^2)}\right)^2\right)^{\frac{1}{2}}$$

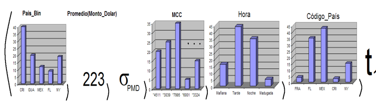
Índice de Afinidad

Now, if we weight the importance of a variable Y_j with a weight w_j (with $0 \leq w_j \leq 1$ and $\sum_{j=1}^{p} w_j = 1$), we define the (weighted) **affinity similarity coefficient** $\mathit{aff}(k,k') \equiv a(k,k')$ between the units (groups) $k, k' \in E$ by the weighted average:

$$a(k,k') \quad := \quad \sum_{j=1}^{p} w_j \cdot \mathit{aff}\,(\xi_{kj}, \xi_{k'j}) = \sum_{j=1}^{p} w_j \cdot \sum_{l=1}^{m_j} \sqrt{\frac{n_{kjl}}{n_{kj\cdot}} \cdot \frac{n_{k'jl}}{n_{k'j\cdot}}}. \qquad (8.68)$$

Perfil del Cliente

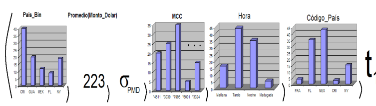
$VC\_1 = ($  $, t_1, t_2, \ldots) \longrightarrow d_1$

$VC\_2 = ($  $, t_1, t_2, \ldots) \longrightarrow d_2$

.

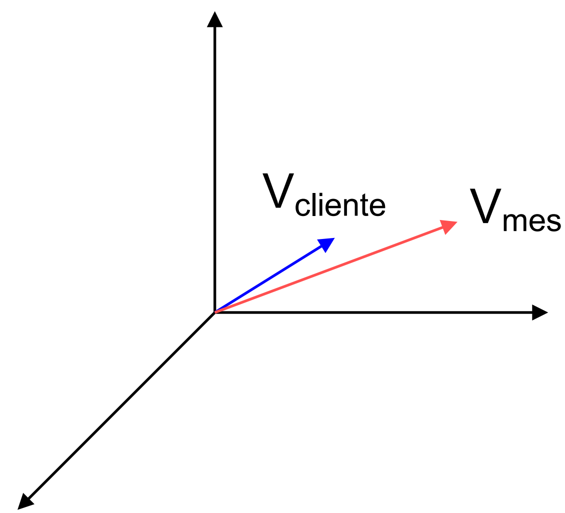
.

.

$VC\_n = ($  $, t_1, t_2, \ldots) \longrightarrow d_n$

$\longrightarrow V_{cliente} = (d_1, d_2, \ldots, d_n)$

Distancia Máxima = $\| V_{cliente} \|$

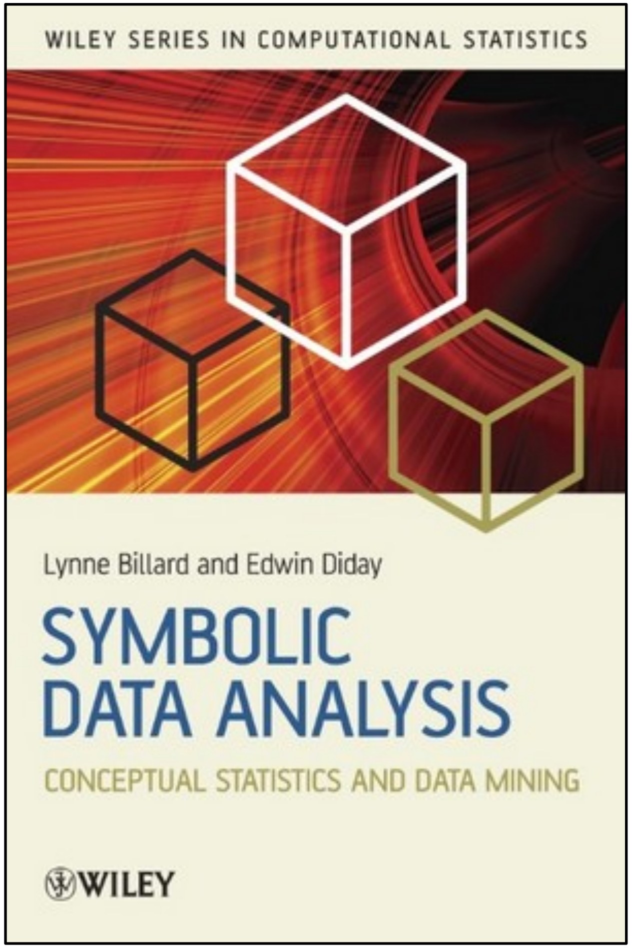


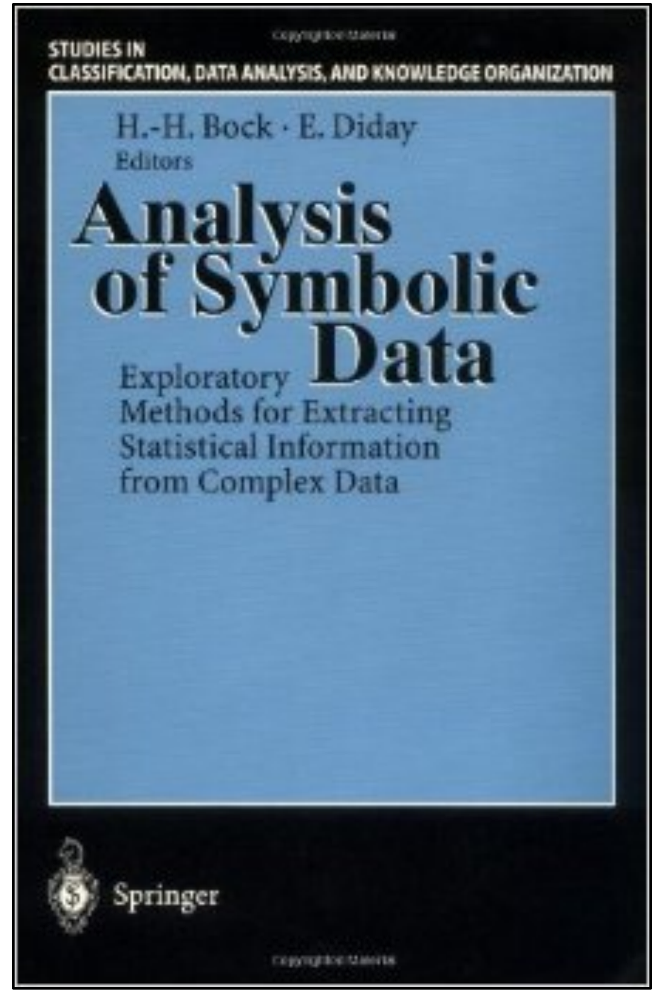
Inusualidad o atipicidad

Comparación Vector Personal vs Transacción

Concept	TablaAprendizajeAbril.BNumber	TablaAprendizajeAbril.duration	Interval.TablaAprendizajeAbril.duration	TablaAprendizajeAbril.indInternational	TablaAprendizajeAbril.hora	TablaAprendizajeAbril.diaSemana	Distancia Maxima	Desviacion Estandar Distancia Hausdorff	Promedio Distancias	Afinidad Ma
8102148			[0 , 240]				1.0396	0.0639	0.7624	0.8397
La Transaccion			[10 , 10]				1.0326	0	1.0326	0.8665


WORLD INTELLECTUAL PROPERTY ORGANIZATION
WIPO
OMPI
ORGANISATION MONDIALE DE LA PROPRIETE INTELLECTUELLE


WILEY SERIES IN COMPUTATIONAL STATISTICS
Lynne Billard and Edwin Diday
SYMBOLIC DATA ANALYSIS
CONCEPTUAL STATISTICS AND DATA MINING
WILEY


STUDIES IN CLASSIFICATION, DATA ANALYSIS, AND KNOWLEDGE ORGANIZATION
H.-H. Bock · E. Diday
Editors
Analysis of Symbolic Data
Exploratory Methods for Extracting Statistical Information from Complex Data
Springer

PROMiDAT

IBEROAMERICANO

Programa Iberoamericano de Formación en Minería de Datos

Gracias…..