

16 Junio 2020. Text Mining Clasificación y Reglas de Asociación.

Herramienta: Tanagra. Parte 1

Descargar los archivos `text_mining_clas1.txt` y `text_mining_clas2.txt`.

Presentar un tutorial en que desarrolle la resolución de la tarea

Herramienta: Tanagra

Parte A

archivo `text_mining_clas1.txt`

1. Abrir Tanagra y cargar el archivo
2. Define status. Target: tipo Input: termino1, termino2, termino3
3. Ejecutar
4. Spv learning ID3
5. Parametros supervisados: Min size for split: 5 Min size for leaves 2
6. Presente el árbol de clasificación generado. Explique el proceso de clasificación
7. Explique la matriz de confusión. Qué representan las filas? qué representan las columnas?

Parte B

archivo `text_mining_clas2.txt`

1. Abrir Tanagra y cargar el archivo
2. Define status. Target: tipo Input: concepto1, concepto2, concepto3
3. Ejecutar
4. Spv learning ID3
5. Parametros supervisados: Min size for split: 5 Min size for leaves 2
6. Explique el proceso de clasificación
7. Explique la matriz de confusión

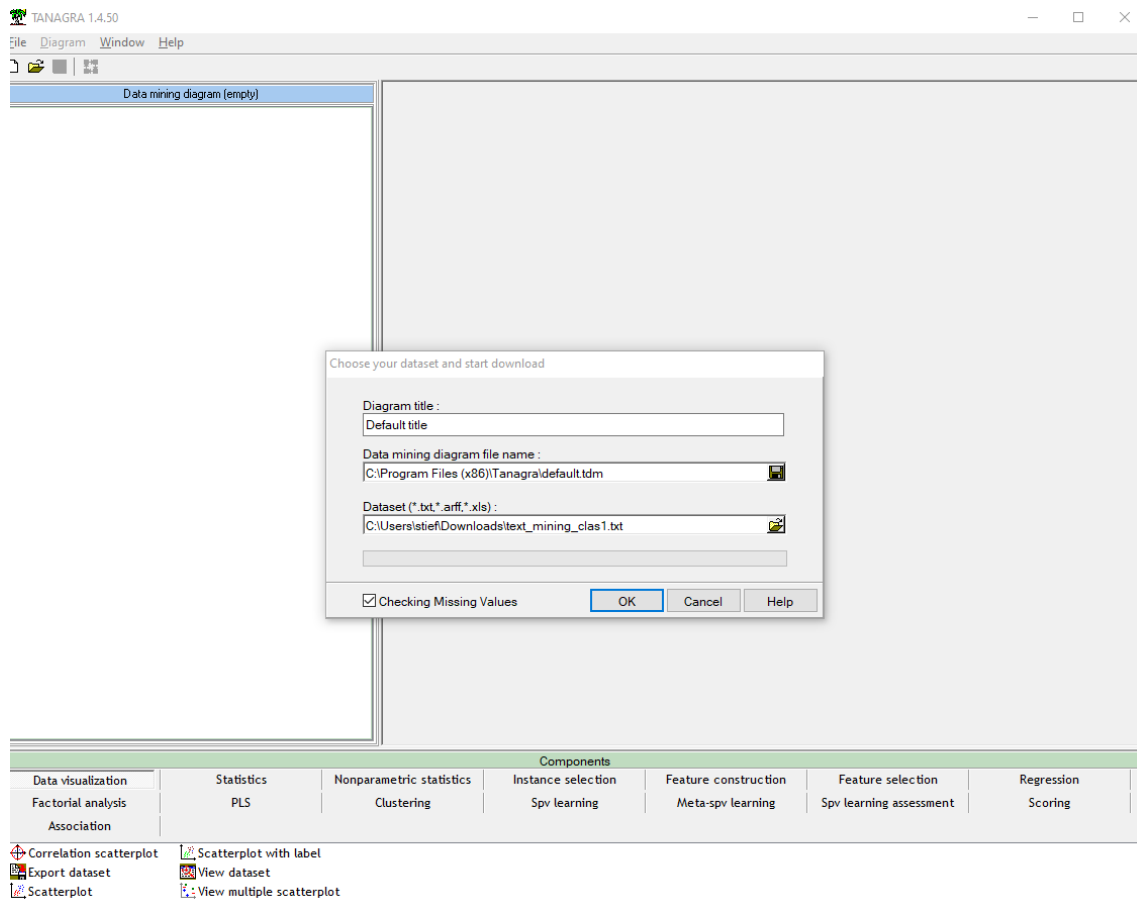
Parte C

Reglas de asociación supervisadas

1. Con los mismos datos y define status que en la Parte B
2. Association
3. Spv Assoc Rule Support 0.2 Confidence 0.5 class value: A
4. Idem con los restantes tipos en class value

PARTE A

Abrimos y cargamos los datos `text_mining_clas1.txt`.



Podemos usar la herramienta “View Dataset” para ver los datos con los que estamos trabajando.

Vamos a hacer clasificación porque queremos saber en base a las palabras de un documento, que tipo de documento es (medicina, tecnología..)

Definimos un nuevo “Define status”, en el que el “arget será “tipo”, y las input serán “termino1, termino2 y termino3”.

TANAGRA 1.4.50 - [View dataset 1 [All] (100 examples, 5 attributes)]

File Diagram Component Window Help

Default title

Dataset (text_mining_clas1.txt)

View dataset 1

Define status 1

	documento	termino1	termino2	termino3	Tipo
1	documento12	2	11		A
2	documento22	0	12		A
3	documento32	0	10		A
4	documento41	0	8		A
5	documento52	0	7		A
6	documento62	0	11		A
7	documento73	1	2		A
8	documento82	2	8		A
9	documento93	1	13		A

Define attribute statuses

Parameters

Attributes :

documento
termino1
termino2
termino3
Tipo

Target Input Illustrative

Tipo

Clear all Clear selected

OK Cancel Help

31 documento37 5 8 A
32 documento37 5 11 B
33 documento36 3 14 B

Components

Data visualization Factorial analysis Association

Statistics PLS

Nonparametric statistics Clustering

Instance selection Spv learning

Feature construction Meta-spv learning

Feature selection Spv learning assessment

Regression Scoring

Correlation scatterplot
Export dataset
Scatterplot

Scatterplot with label
View dataset
View multiple scatterplot

Confirmamos que se han establecido tal y como queríamos las variables:

TANAGRA 1.4.50 - [Define status 1]

File Diagram Component Window Help

Default title

Dataset (text_mining_clas1.txt)

View dataset 1

Define status 1

Define status 1

Parameters

Target : 1
Input : 3
Illustrative : 0

Results

Attribute	Target	Input	Illustrative
documento	-	-	-
termino1	-	yes	-
termino2	-	yes	-
termino3	-	yes	-
Tipo	yes	-	-

Computation time : 0 ms.
Created at 17/06/2020 0:00:15

Components

Data visualization Factorial analysis Association

Statistics PLS

Nonparametric statistics Clustering

Instance selection Spv learning

Feature construction Meta-spv learning

Feature selection Spv learning assessment

Regression Scoring

Binary logistic regression
BVM
C4.5

C-PLS
C-RT
CS-CRT

CS-MC4
C-SVC
CVM

Decision List
ID3
K-NN

Linear discriminant analysis
Log-Reg TRIRLS
Multilayer perceptron

Multinomial Logistic
Naive bayes
Naive bayes contr

Ahora nos iremos a la pestaña “Spv Learning”, ya que asociación y clasificación tienen una variable objetivo. En ella elegiremos el algoritmo ID3, seleccionándolo y arrastrándolo hasta colgarlo debajo de nuestro “define status”.

Pulsaremos sobre el con el botón derecho, y seleccionaremos “supervised parameters”.

Aquí cambiaremos los siguientes parámetros:

Min size for split: 5 Min size for leaves 2

Una vez hecho esto lo ejecutamos.

Classifier characteristics

Data description

Target attribute	Tipo (3 values)
# descriptors	3

Tree description

Number of nodes	13
Number of leaves	7

Decision tree

- termino1 < 4,0000 then Tipo = A (100,00 % of 30 examples)
- termino1 >= 4,0000
 - termino1 < 9,5000
 - termino3 < 10,5000
 - termino3 < 7,5000 then Tipo = B (100,00 % of 24 examples)
 - termino3 >= 7,5000
 - termino3 < 8,5000 then Tipo = B (75,00 % of 4 examples)
 - termino3 >= 8,5000 then Tipo = B (100,00 % of 6 examples)
 - termino3 >= 10,5000
 - termino3 < 11,5000 then Tipo = B (75,00 % of 4 examples)
 - termino3 >= 11,5000 then Tipo = B (100,00 % of 11 examples)
 - termino1 >= 9,5000 then Tipo = C (100,00 % of 21 examples)

Computation time : 0 ms.

Created at 17/06/2020 2:51:02

Aquí vemos el árbol de clasificación que se ha creado, teniendo en cuenta los parámetros cambiados.

El árbol se va dividiendo primero dependiendo del numero de veces que aparezca el término 1. Tras esto se divide en función del número de veces que aparece el término 3. Observamos que el término 3 no influye en el árbol de decisión.

Por ejemplo, observamos que cuando el termino1 se presentaba menos de 4,0000 veces, es porque el documento era de tipo A.

En cambio, si el término1 es mayor o igual a 4,0000 y menor que 9,5000, y el término 3 es menor que 7,5000, es porque el documento era tipo B.

Podemos observar que la variable objetivo es el tipo, tal y como habíamos decretado.

Por último, vamos a presentar la matriz de confusión:

Classifier performances

Error rate			0,0200				
Values prediction			Confusion matrix				
Value	Recall	1-Precision		A	B	C	Sum
A	0,9677	0,0000	A	30	1	0	31
B	1,0000	0,0408	B	0	47	0	47
C	0,9545	0,0000	C	0	1	21	22
			Sum	30	49	21	100

Podemos ver que la mayor parte de los elementos están correctamente clasificados. Estos se vería en la diagonal principal.

Sin embargo, vemos que un elemento que era tipo A, se clasificó mal como tipo B.

También observamos que un elemento que era tipo C, se clasificó mal como un tipo B.

Las columnas representan la predicción, y las filas representan la realidad.

PARTE B

En este caso, vamos a trabajar con el archivo text_mining_clas2.txt.

Seguimos los pasos de la parte anterior para abrir el archivo, pero en esta ocasión, usaremos con inputs “concepto1, concepto2, concepto3”, y como variable objetivo (target), “tipo”.

The screenshot displays a software interface for configuring a supervised learning task. On the left, a tree view shows a dataset named 'text_mining_clas2.txt' with a sub-entry 'Define status 1'. The main panel is titled 'Define status 1' and contains a 'Parameters' section with the following settings: Target : 1, Input : 3, and Illustrative : 0. Below this is a 'Results' section showing a table with the following data:

Attribute	Target	input	Illustrative
documento	-	-	-
concepto1	-	yes	-
concepto2	-	yes	-
concepto3	-	yes	-
Tipo	yes	-	-

At the bottom of the interface, it shows 'Computation time : 0 ms.' and 'Created at 17/06/2020 3:30:26'.

Seguimos el mismo procedimiento de la parte A, poniendo los mismos valores en los parámetros supervisados a la hora de configurar el algoritmo ID3, y lo ejecutamos:

Decision tree

- concepto1 in [c1] then Tipo = A (100,00 % of 30 examples)
- concepto1 in [c2]
 - concepto3 in [c7] then Tipo = B (100,00 % of 25 examples)
 - concepto3 in [c8] then Tipo = B (81,82 % of 11 examples)
 - concepto3 in [c9] then Tipo = B (100,00 % of 13 examples)
- concepto1 in [c3] then Tipo = C (100,00 % of 21 examples)

Computation time : 0 ms.

Created at 17/06/2020 3:33:27

En este caso el árbol de decisión tiene menos nodos y menos hojas que en la parte A.

La bifurcación principal se presenta si concepto1 ha sido C1, C2 o C3.

Si ha sido C1, entonces era tipo A.

Si ha sido C3, entonces era tipo C.

Si ha sido C2, entonces hay una nueva bifurcación, que depende de si concepto3 ha sido C7, C8 o C9.

Si concepto1 ha sido C2, y concepto 3 ha sido C7, entonces era tipo B.

Si concepto1 ha sido C2, y concepto 3 ha sido C8, entonces era tipo B.

Si concepto1 ha sido C2, y concepto 3 ha sido C9, entonces era tipo B.

PARTE C

En esta parte usaremos los mismos datos y el mismo define status que en la parte B, pero en esta ocasión usaremos un algoritmo de asociación supervisado.

Para ello, en la pestaña “Association” y elegimos “spv association rule”, la cual arrastraremos hasta colgarla debajo de nuestro define status. Haciendo click derecho, seleccionamos “Parameters..”, y cambiamos los siguientes parámetros:

Support 0.2 Confidence 0.5 class value: A

Spv Assoc Rule Parameters

Parameters

Support : 0,2

Confidence : 0,5

Max card itemsets : 4

Lift : 1,1

Learning set ratio : 1

Repetition : 1

Class value : A

VT Boundary Filtering : 2

OK Cancel Help

Pulsamos ok, y ejecutamos el algoritmo. Observamos lo siguiente:

Filtered = 2 rules

Rules evaluation

N°	Antécédent	Conséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift	Leverage	Importance	Conviction	Surprise	VT-H
1	"concepto2=c4"	"Tipo=A"	100	30	31	30	0,30000	1,00000	3,22581	0,20700	4,24850	99,99000	0,96774	0,00000
2	"concepto1=c1"	"Tipo=A"	100	30	31	30	0,30000	1,00000	3,22581	0,20700	4,24850	99,99000	0,96774	0,00000

All rules

Rules evaluation

N°	Antécédent	Conséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift	Leverage	Importance	Conviction	Surprise	VT-H
1	"concepto1=c1"	"Tipo=A"	100	30	31	30	0,30000	1,00000	3,22581	0,20700	4,24850	99,99000	0,96774	0,00000
2	"concepto2=c4"	"Tipo=A"	100	30	31	30	0,30000	1,00000	3,22581	0,20700	4,24850	99,99000	0,96774	0,00000
3	"concepto1=c1"	"Tipo=A"	100	30	31	30	0,30000	1,00000	3,22581	0,20700	4,24850	99,99000	0,96774	0,00000

Como podemos ver, si concepto2 es C4, entonces el tipo es A.

Si concepto1=c1 entonces el tipo es A.

Vamos a cambiar ahora únicamente el apartado “Class value” de los parámetros supervisados.

Observamos que al cambiar “Class Value” por B, obtenemos las siguientes reglas de asociación:

Default title														
Dataset (text_mining_clas2.txt)														
Define status 1														
Spv Assoc Rule 1														

Rules evaluation														
N°	Antécédent	Conséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift	Leverage	Importance	Conviction	Surprise	VT-H 100
1	"concepto2=c5"	"Tipo=B"	100	49	47	47	0,47000	0,95918	2,04082	0,23970	99,99000	12,98500	0,95745	0,000
2	"concepto1=c2"	"Tipo=B"	100	49	47	47	0,47000	0,95918	2,04082	0,23970	99,99000	12,98500	0,95745	0,000

All rules														
Rules evaluation														
N°	Antécédent	Conséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift	Leverage	Importance	Conviction	Surprise	VT-H 100
1	"concepto1=c2"	"Tipo=B"	100	49	47	47	0,47000	0,95918	2,04082	0,23970	99,99000	12,98500	0,95745	0,000
2	"concepto2=c5"	"Tipo=B"	100	49	47	47	0,47000	0,95918	2,04082	0,23970	99,99000	12,98500	0,95745	0,000
3	"concepto1=c2"	"Tipo=B"	100	25	47	25	0,25000	1,00000	2,12766	0,13250	1,22645	99,99000	0,53191	6,436
4	"concepto2=c5"	"Tipo=B"	100	25	47	25	0,25000	1,00000	2,12766	0,13250	1,22645	99,99000	0,53191	6,436
5	"concepto1=c2"	"Tipo=B"	100	49	47	47	0,47000	0,95918	2,04082	0,23970	99,99000	12,98500	0,95745	0,000
6	"concepto2=c5"	"Tipo=B"	100	25	47	25	0,25000	1,00000	2,12766	0,13250	1,22645	99,99000	0,53191	6,436

Computation time : 0 ms.
Created at 17/06/2020 4:31:11

Como podemos ver, si concepto2 es C5, entonces el tipo es B.

Si concepto1 es C2, entonces el tipo es B.

Procedemos ahora a cambiar “Class Value” por C. Observamos que se forman las siguientes reglas de asociación:

Filtered = 2 rules

Rules evaluation

N°	Antécédent	Conséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift	Leverage	Importance	Conviction	Surprise	VT-H 100
1	"concepto2=c6"	"Tipo=C"	100	21	22	21	0,21000	1,00000	4,54545	0,16380	4,36945	99,99000	0,95455	0,000
2	"concepto1=c3"	"Tipo=C"	100	21	22	21	0,21000	1,00000	4,54545	0,16380	4,36945	99,99000	0,95455	0,000

All rules

Rules evaluation

N°	Antécédent	Conséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift	Leverage	Importance	Conviction	Surprise	VT-H 100
1	"concepto1=c3"	"Tipo=C"	100	21	22	21	0,21000	1,00000	4,54545	0,16380	4,36945	99,99000	0,95455	0,000
2	"concepto2=c6"	"Tipo=C"	100	21	22	21	0,21000	1,00000	4,54545	0,16380	4,36945	99,99000	0,95455	0,000
3	"concepto1=c3"	"Tipo=C"	100	21	22	21	0,21000	1,00000	4,54545	0,16380	4,36945	99,99000	0,95455	0,000

Vemos que si concepto2 es C6, entonces el tipo es C.

Si concepto1 es C3, entonces el tipo es C.

