

# TANAGRA

Origine, architecture

Interface

Accès aux données

Définir les traitements

Bibliothèque

Évaluation et comparaisons

Expérimentations

Performances

Exploration graphique

Université de Lyon 2

<http://chirouble.univ-lyon2.fr/~ricco/tanagra/>

Culture « machine learning », stat. et stat. exploratoire

Code source libre en DELPHI (Licence private – 1.4.1)

Site WEB avec de nombreux didacticiels

Exécution stand-alone sous Windows (Exécutable)

# TANAGRA

Origine, architecture

Interface

Accès aux données

Définir les traitements

Bibliothèque

Évaluation et comparaisons

Expérimentations

Performances

Exploration graphique

The screenshot displays the TANAGRA 1.4.1 interface. The top window shows a workflow diagram with steps: Dataset [heart.txt], Define status 1, and Supervised Learning 1 (C-RT). A yellow box labeled "Chaîne de traitements" is overlaid on this window. The bottom window shows the results of the supervised learning process, including a table of results, a tree description, a decision tree, and a list of components.

	0.2111	0.2111	0.2000	0.2111
3	5	0.1667	0.2111	
2	7	0.1444	0.2000	
1	11	0.1111	0.2111	

**Tree description**

Number of nodes	7
Number of leaves	4

**Decision tree**

- type\_douleur in [D]
- depression < 0.5500
  - vaisseau in [D,B,C] then coeur = **presence** (73.33 % of 3 examples)
  - vaisseau in [A] then coeur = **absence** (73.68 % of 19 examples)
- depression >= 0.5500 then coeur = **presence** (90.91 % of 55 examples)
- type\_douleur in [C,B,A] then coeur = **absence** (78.02 % of 53 examples)

Computation time : 40 ms.

**Components**

Data visualization	Statistics	Nonparametric statist	Instance selection	Feature construction	Feature selection
Regression	Factorial analysis	PLS	Clustering	Svm learning	Meta-svm learning
Spv learning assesseme	Scoring	Association			

Binary logistic regression  
C-RT  
Decision List  
ID3  
K-NN  
Linear discriminant analysis  
Multilayer perceptron  
Multinomial Logistic Regression  
Naive bayes  
Prototype-NN  
Radial basis function  
SVM

**Composants**

Transcription de la notion de traitements alternatifs sur les mêmes données  
Rôle central du composant « Define Status »

# TANAGRA

Origine, architecture

Interface

Accès aux données

Définir les traitements

Bibliothèque

Évaluation et comparaisons

Expérimentations

Performances

Exploration graphique

Fichier texte, avec séparateur tabulation (export tableur par exemple)

```
sep_length»      sep_width»      pet_length»      pet_width»      type»
5.10» 3.50» 1.40» 0.20» Iris-setosa»
4.90» 3.00» 1.40» 0.20» Iris-setosa»
4.70» 3.20» 1.30» 0.20» Iris-setosa»
4.60» 3.10» 1.50» 0.20» Iris-setosa»
5.00» 3.60» 1.40» 0.20» Iris-setosa»
5.40» 3.90» 1.70» 0.40» Iris-setosa»
4.60» 3.40» 1.40» 0.30» Iris-setosa»
5.00» 3.40» 1.50» 0.20» Iris-setosa»
4.40» 2.90» 1.40» 0.20» Iris-setosa»
4.90» 3.10» 1.50» 0.10» Iris-setosa»
5.40» 3.70» 1.50» 0.20» Iris-setosa»
4.80» 3.40» 1.60» 0.20» Iris-setosa»
4.80» 3.00» 1.40» 0.10» Iris-setosa»
4.30» 3.00» 1.10» 0.10» Iris-setosa»
5.80» 4.00» 1.20» 0.20» Iris-setosa»
5.70» 4.40» 1.50» 0.40» Iris-setosa»
5.40» 3.90» 1.30» 0.40» Iris-setosa»
5.10» 3.50» 1.40» 0.30» Iris-setosa»
5.70» 3.80» 1.70» 0.30» Iris-setosa»
5.10» 3.80» 1.50» 0.30» Iris-setosa»
```



D'autres accès sont disponibles : format WEKA (arff), EXCEL 97&2000

# TANAGRA

Origine, architecture

Interface

Accès aux données

**Définir les traitements**

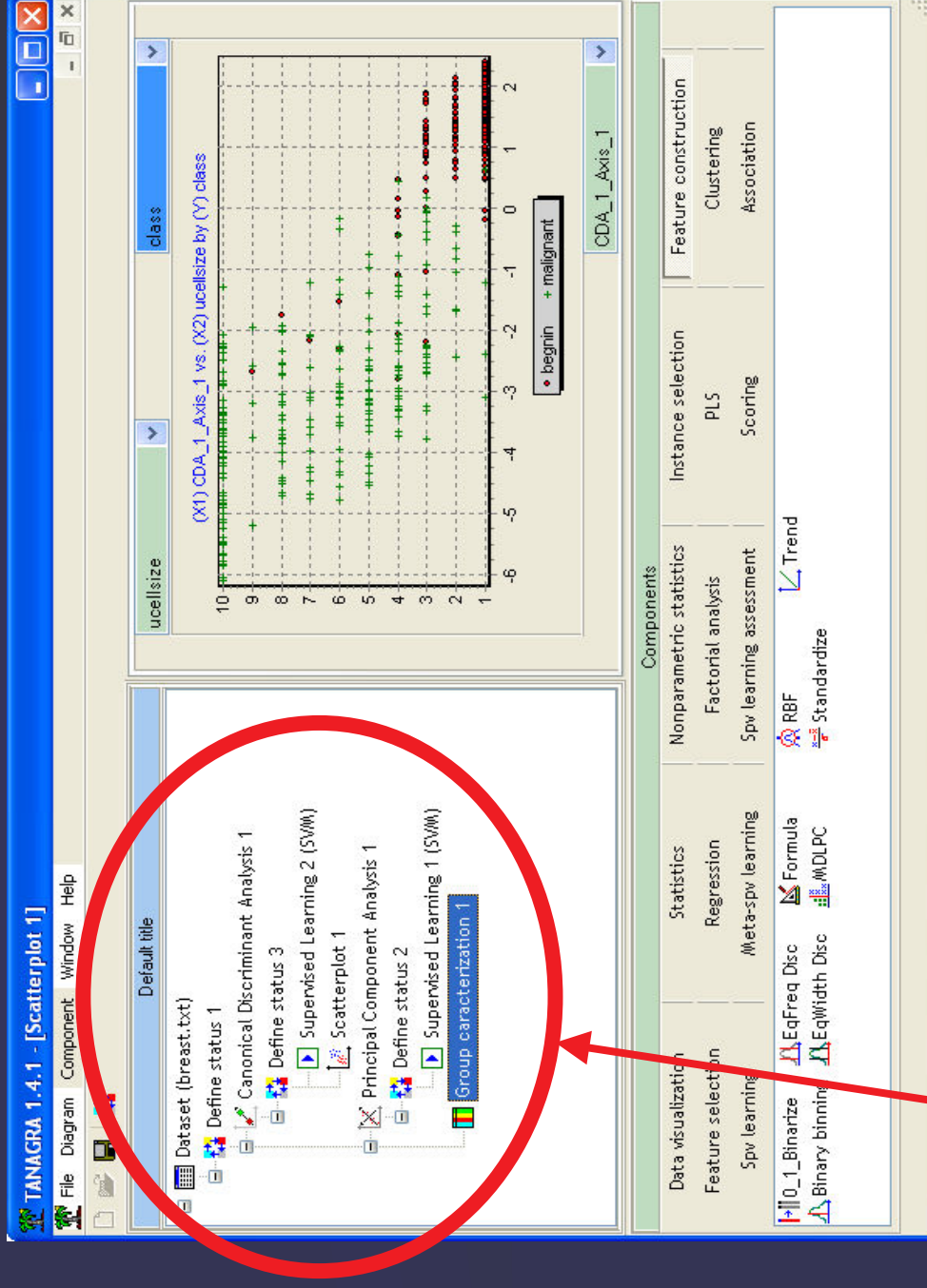
Bibliothèque

Évaluation et comparaisons

Expérimentations

Performances

Exploration graphique



Plusieurs traitements en parallèle  
Combinaison de traitements

# TANAGRA

Origine, architecture

Interface

Accès aux données

Définir les traitements

**Bibliothèque**

Évaluation et comparaisons

Expérimentations

Performances

Exploration graphique

- Description rapide des données et stat. descriptives 😊
- Traitement des données manquantes 😞
- Restrictions sur les individus (échantillon, condition) 😊
- Tests statistiques 😊
- Sélection des variables pour le supervisé 😊
- Construction de variables 😊
- Apprentissage supervisé (et régression) 😊
- Apprentissage non-supervisé 😊
- Règles d'association 😊
- Méthodes factorielles 😊
- Séries temporelles 😞



Un mix des cultures

# TANAGRA

Origine, architecture

Interface

Accès aux données

Définir les traitements

Bibliothèque

Évaluation et comparaisons (1/2)

Expérimentations

Performances

Exploration graphique

The screenshot displays the TANAGRA 1.4.1 software interface. The main window shows a workflow diagram with the following components:

- Dataset (breast.txt)
- Define status 1
- Supervised Learning 1 (Linear discriminant analysis)
- Cross-validation 1
- Supervised Learning 2 (Binary logistic regression)
- Cross-validation 2 (highlighted)
- Supervised Learning 3 (C-RT)
- Cross-validation 3

The results window for Cross-validation 2 shows the following data:

Value	Recall	1-Precision
benign	0.9746	0.0296
malignant	0.9435	0.0486

Overall cross-validation error rate: 0.0361

Confusion matrix:

	benign	malignant	Sum
benign	2229	58	2287
malignant	68	1135	1203
Sum	2297	1193	3490

Computation time : 2750 ms.  
Created at 14/11/2005 17:00:32

The bottom panel shows the Components list:

- Data visualization: Statistics
- Feature selection: Regression, Meta-svm learning, Bias-variance decomposition, Bootstrap
- Instance selection: Nonparametric statistics
- Feature construction: Feature construction
- PLS: Factorial analysis
- Scoring: Spv learning assessment, Cross-validation, Test, Train-test

(1) Visualisation séparée

(2) Mais possibilité de regrouper les résultats dans un seul fichier de sortie TXT

# TANAGRA

The screenshot displays the TANAGRA 1.4.1 software interface. The main window shows a project tree on the left with the following steps: Dataset (breast\_compare\_algorithms.xls), Select examples 1, Define status 1, Supervised Learning 1 (Linear discriminant analysis), Supervised Learning 2 (C-RT), Supervised Learning 3 (Multinomial Logistic Regression), Supervised Learning 4 (K-NN), Define status 2, and Test 1. The 'View dataset 1' window is active, showing a 'Dataset description' for 'breast\_compare...' with 63 ms computation time and 38 KB memory. It lists 11 attributes and 699 examples. The 'Attribute Category Informations' table is as follows:

Attribute	Category	Informations
clump	Continue	-
ucelsize	Continue	-
ucelshape	Continue	-
ngadhesion	Continue	-
sepics	Continue	-
bnuclei	Continue	-
bchromathn	Continue	-
normnucl	Continue	-
mitoses	Continue	-
class	Discrete	2 values
statut	Discrete	2 values

The 'Compute' window shows a table with columns 'uc1', 'mitoses', 'class', and 'statut'. A red circle highlights the 'statut' column, and a red arrow points from the 'statut' row in the 'Attribute Category Informations' table to the corresponding row in the table. The 'Results' window shows '300 selected examples from 699' and 'Attribute selection : statut', 'Value selection : apprentissage'. The bottom panel shows various analysis components like Data visualization, Statistics, PLS, Clustering, Cross-validation, Bias-variance decomposition, Bootstrap, Instance selection, Feature construction, Meta-spx learning, Spv learning, Feature selection, Spv learning assessment, Regression, and Scoring.

## Évaluation et comparaisons (2/2)

Utiliser les mêmes individus en apprentissage, puis comparer sur les mêmes individus en test

# TANAGRA

Origine, architecture

Interface

Accès aux données

Définir les traitements

Bibliothèque

Évaluation et comparaisons

Expérimentations

Performances

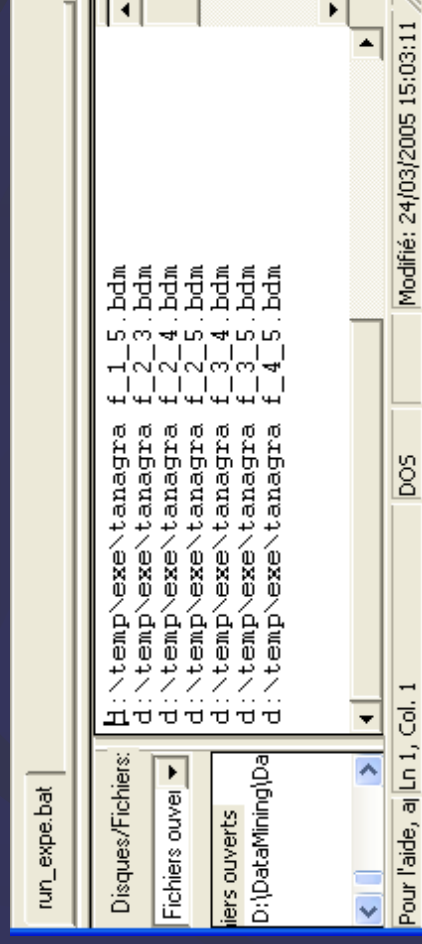
Exploration graphique

Exploiter les possibilités de la ligne de commande



```
end;
procedure TForm1.Button2Click(Sender: TObject);
var lst: TStringList;
    rep: integer;
begin
    lst:= TStringList.Create();
    lst.LoadFromFile(self.OpenDialog1.FileName);
    for rep:= 1 to self.spinMaxTest.Value do
    begin
        //modifier le paramètre
        lst.Values['x_best']:= IntToStr(rep);
        lst.SaveToFile('run.tdm');
        //info
        self.LMDLEDLabel1.Value:= rep;
        Application.ProcessMessages();
        //lancer l'apprentissage
        self.LMDStarter.Execute();
    end;
    lst.Free();
end;
end.
```

(1) En écrivant un programme qui exécute TANAGRA



```
run_expe.bat
Disques/Fichiers:
Fichiers ouverts
D:\DataMining\Da
H:\temp\exe\tanagra f_1_5.bdm
d:\temp\exe\tanagra f_2_3.bdm
d:\temp\exe\tanagra f_2_4.bdm
d:\temp\exe\tanagra f_2_5.bdm
d:\temp\exe\tanagra f_3_4.bdm
d:\temp\exe\tanagra f_3_5.bdm
d:\temp\exe\tanagra f_4_5.bdm
```

(2) En écrivant manuellement le fichier script



# TANAGRA

Origine, architecture

Interface

Accès aux données

Définir les traitements

Bibliothèque

Évaluation et comparaisons

Expérimentations

**Performances**

Exploration graphique

- Pas de limitation théorique (espace adressable) 😊
- Toutes les données en mémoire 😞
- Exécutable binaire sans moteur intermédiaire 😊
- **Organisation du code → Aide à la simplicité et au minimalisme**
  - Rapport code interface / code calcul 😊😊
  - Choix optimisés pour le temps de calcul 😊😊  
(importation, arbres de décision, ...)
  - Gestion mémoire moins performante 😞  
(règles d'association, calcul matriciel, ...)

# TANAGRA

Origine, architecture

Interface

Accès aux données

Définir les traitements

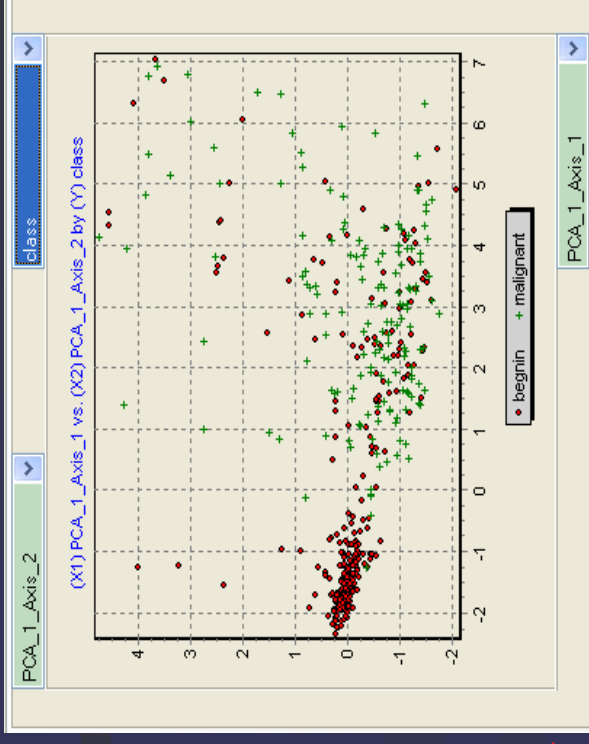
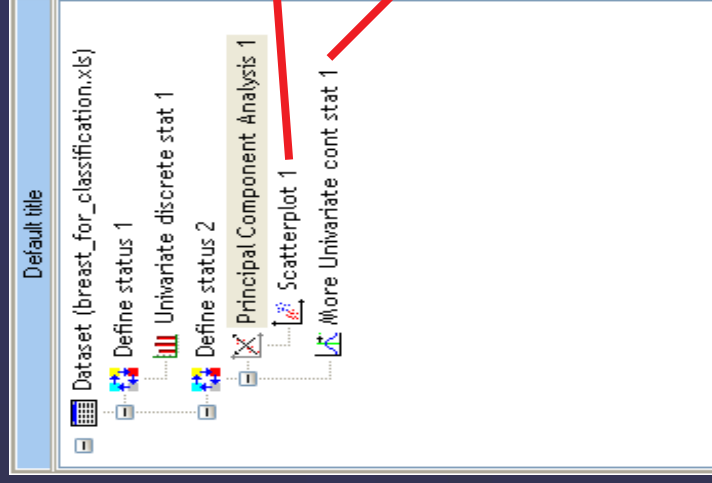
Bibliothèque

Évaluation et comparaisons

Expérimentations

Performances

Exploration graphique



Values	Count	Percent	Histogram
X_<_1,9000	353	50,50%	
1,9000_<=X_<_2,8000	59	8,44%	
2,8000_<=X_<_3,7000	56	8,01%	
3,7000_<=X_<_4,6000	44	6,29%	
4,6000_<=X_<_5,5000	34	4,86%	
5,5000_<=X_<_6,4000	30	4,29%	
6,4000_<=X_<_7,3000	30	4,29%	
7,3000_<=X_<_8,2000	28	4,01%	
8,2000_<=X_<_9,1000	7	1,00%	
X>=9,1000	58	8,30%	

Minimaliste et sans interactivité

# TANAGRA

## Bilan des fréquentations Le virage statistique en août 2005

Découverte d'une autre communauté avec ses logiciels gratuits « féériques » :

- R
- DATAPLOT
- OPEN STAT
- VISTA
- WINIDAMS
- etc.

<http://members.aol.com/johnp71/javasta2.html>

<http://freestatistics.altervista.org/stat.php>

Bilan de fréquentation de votre site		Le 4/12/2005	
Visiteurs uniques		43	
Visites		44	
Pages vues		51	
Nbr de pages vues par visiteurs		1.19	
Nbr de pages vues par visites		1.16	
Nbr de visites par visiteurs		1.02	
Durée moyenne d'une visite		6' 00"	
Nbr de visites à une seule page		36	
Nbr de pages visitées différentes		4	
<b>Estimation pour la journée en cours</b>			
		J-1	
Visiteurs uniques	51	-16.4 %	
Visites	52	-18.8 %	
Pages vues	61	-22.8 %	
<b>Progression &amp; Bilan annuel</b>			
Mois	Visiteurs	Visites	Pages vues
<b>Juillet 2005</b>	906	989	1348
<b>Août 2005</b>	1596 +76.16 %	1781 +80.08 %	2510 +86.20 %
<b>Septembre 2005</b>	1806 +13.16 %	1972 +10.72 %	2621 +4.42 %
<b>Octobre 2005</b>	1988 +10.08 %	2163 +9.69 %	2950 +12.55 %
<b>Novembre 2005</b>	2300 +15.69 %	2495 +15.35 %	3289 +11.49 %