

Extraction de motifs séquentiels à partir de données issues d'images satellitaires

Application à la caractérisation des paysages ruraux au Mali

Fadi Badra

UMR TETIS - Cemagref, Montpellier, France
badra@teledetection.fr

Extraction de motifs séquentiels à partir de données issues d'images satellitaires

Objectif

Mettre en œuvre des méthodes de caractérisation des paysages ruraux et de leurs systèmes de culture en appliquant des algorithmes d'extraction de motifs séquentiels sur des images satellites.

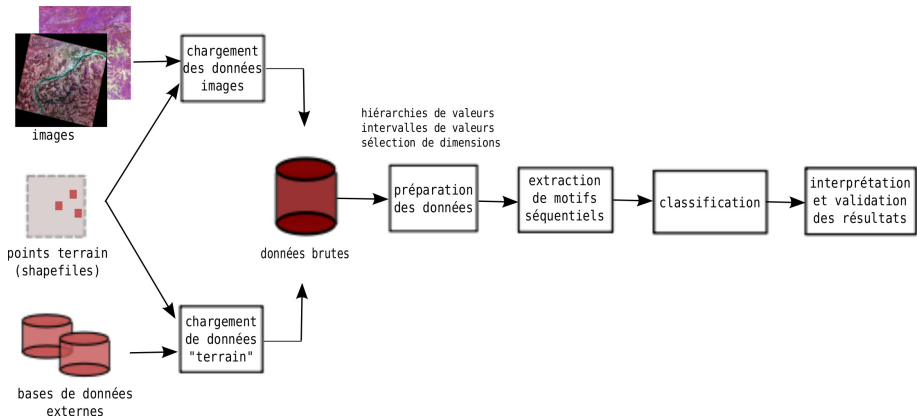
Partenaires

- Fadi Badra (UMR TETIS, Cemagref)
- Maguelonne Teisseire (UMR TETIS, Cemagref)
- Agnès Bégué (UMR TETIS, CIRAD)
- Elodie Vintrou (UMR TETIS, CIRAD)
- Christian Baron (UMR TETIS, CIRAD)

Motivations

- Adapter un mécanisme de recherche de motifs séquentiels à la fouille de données issues d'images satellites
 - données *multi-source* (terrain, image, BD externes)
 - description des données *multidimensionnelle* (informations spectrales, spatiales, temporelles)
- Proposer un système de classification (supervisée), afin d'offrir une aide à la cartographie d'occupation du sol

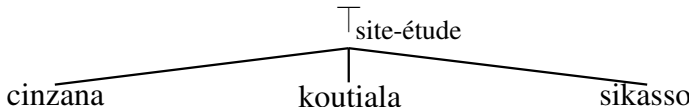
Processus d'extraction de connaissances



Extraction de motifs séquentiels

L'extraction de motifs séquentiels multidimensionnels

- ensemble de dimensions $D = \{D_R, D_A, D_T, D_I\}$
 - D_R : de référence
 - D_A : d'analyse
 - D_T : permettant d'introduire une relation d'ordre (le temps)
 - D_I : ignorées
- domaine de valeur $Dom(D_i), D_i \in D$



L'extraction de motifs séquentiels multidimensionnels

- *item multidimensionnel* $e = (d_1, d_2, \dots, d_m)$, $d_i \in D_A$

$e = (\text{sorgho}, \text{cinzana})$ $e' = (\text{sorgho}, \top_{\text{site-étude}})$
 $D_A = \{\text{type-de-culture}, \text{site-étude}\}$

- *itemset multidimensionnel* $i = \{e_1, e_2, \dots, e_n\}$

$i = \{(\text{sorgho}, \text{cinzana}), (\text{forêt}, \top_{\text{site-étude}})\}$

- *séquence multidimensionnelle* $s = \langle i_1, \dots, i_p \rangle$

$s = \langle \{(\text{sorgho}, \text{cinzana}), (\text{forêt}, \top_{\text{site-étude}})\}, \{(\text{sorgho}, \top_{\text{site-étude}})\} \rangle$

L'extraction de motifs séquentiels multidimensionnels

- *bloc* : ensemble de n -uplets qui ont la même projection sur D_R

pt-id	date	type-de-culture	site-étude	NDVI-100
1	1	arachide	sikasso	très faible
2	1	mil	koutiala	faible
2	2	mil	koutiala	modéré
3	1	sorgho	koutiala	élevé

Base de données DB .

pt-id	date	type-de-culture	site-étude	NDVI-100
2	1	mil	koutiala	faible
2	2	mil	koutiala	modéré

Bloc $B_{(mil,koutiala)}$

$D_R = \{\text{type-de-culture, site-étude}\}$

- *support* d'une séquence s : nombre de blocs qui contiennent s

Recherche d'itemsets séquentiels multidimensionnels

Étant donné un seuil σ_{min} (support minimum), trouver toutes les séquences dont le support est supérieur ou égal à σ_{min} .

Les données

Les relevés terrain

Site d'étude	Zone de culture	Village	Total
Cinzana	345	70	415
Koutiala	316	48	364
Sikasso	152	20	172

Pour chaque relevé terrain :

- un type de culture (relevé en zone de culture) ou un nom de village (relevé village), et
- les coordonnées GPS du point

Les informations externes

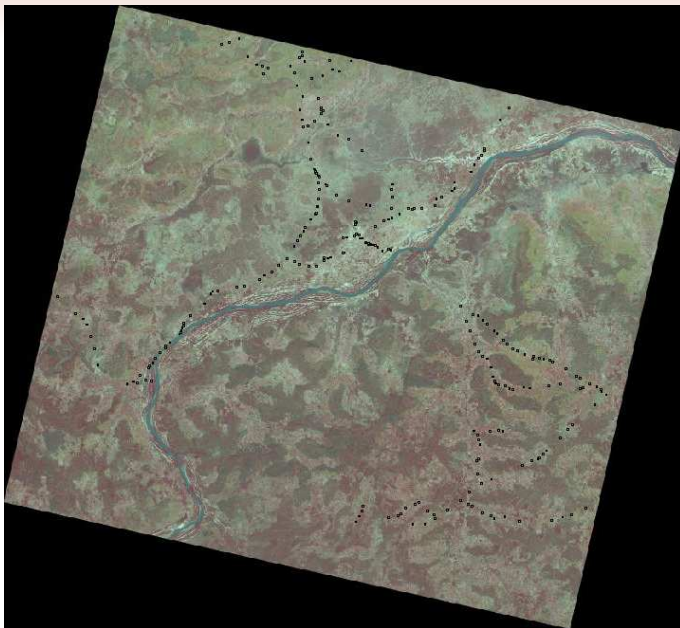
- le type de sol
- la pluviométrie
- le village référent
- la distance au village
- l'ethnie
- le nombre d'arbres
- . . .

Les données images

Id	Vecteur	Site	Date	Résolution
1	SPOT	Cinzana	16 juin 2007	multispectral à 10m
2	SPOT	Cinzana	20 novembre 2007	panchro 2,5m + multispectral à 10m
3	SPOT	Cinzana	20 novembre 2007	panchro 2,5m + multispectral à 10m
4	SPOT	Koutiala	12 juin 2007	multispectral à 10m
5	SPOT	Koutiala	14 novembre 2007	panchro 2,5m + multispectral à 10m
6	SPOT	Koutiala	14 novembre 2007	panchro 2,5m + multispectral à 10m
7	SPOT	Sikasso	14 novembre 2007	panchro 2,5m + multispectral à 10m
8	SPOT	Sikasso	14 novembre 2007	panchro 2,5m + multispectral à 10m
9	MODIS	Tout le Mali	1 ^{er} janvier 2007	250m
10	MODIS	Tout le Mali	3 février 2007	250m
11	MODIS	Tout le Mali	6 mars 2007	250m
12	MODIS	Tout le Mali	23 avril 2007	250m
13	MODIS	Tout le Mali	25 mai 2007	250m
14	MODIS	Tout le Mali	26 juin 2007	250m
15	MODIS	Tout le Mali	12 juillet 2007	250m
16	MODIS	Tout le Mali	29 août 2007	250m
17	MODIS	Tout le Mali	30 septembre 2007	250m
18	MODIS	Tout le Mali	16 octobre 2007	250m
19	MODIS	Tout le Mali	17 novembre 2007	250m

Les images satellites disponibles au 1^{er} mars 2010.

Exemple d'image SPOT (site de Cinzana)



Premières expériences

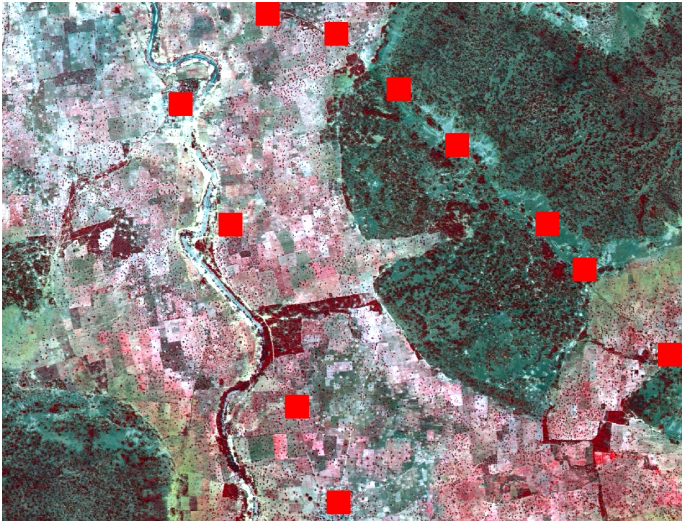
Ensemble d'apprentissage

Points terrain :

- en zone de culture (\neg village)
- situés sur une image SPOT

Site d'étude	Culture	Non culture	Total
Cinzana	138	85	223
Koutiala	105	78	183
Sikasso	46	46	92

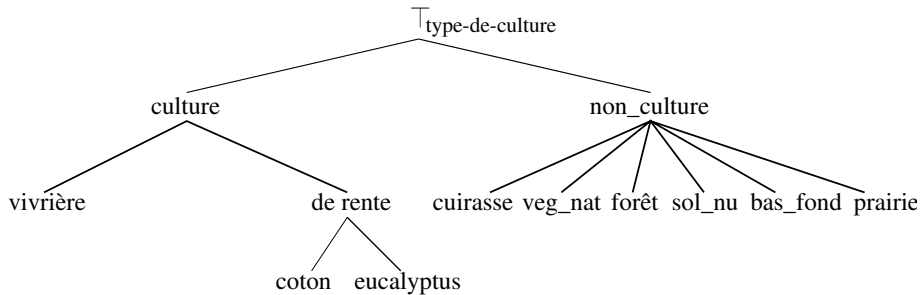
Extraction des indices images



Les dimensions considérées

dimension D_i	intervalles de valeurs				
id-pt	{1}, {2}, ..., {498}				
date	{1}				
site-étude	{cinzana, koutiala, sikasso}				
type-de-culture	{riz, sorgho, maïs, ...}				
nom-village	{dioforongo, tigui, sanando, ...}				
distance-village	proche	éloigné			
	[0,3000]	[3001,+∞[
NDVI-100	très faible	faible	modéré	élevé	
	[-1,0.2]	[0.2,0.3]	[0.3,0.5]	[0.5,1]	
NDVI-200	[-1,0.2]	[0.2,0.3]	[0.3,0.5]	[0.5,1]	
variance-100	très faible	faible	modéré	élevé	très élevé
] $-\infty$,0.56]	[0.56,1.22]	[1.22,2.38]	[2.38,4.00]	[4.00,+∞[
variance-200] $-\infty$,0.56]	[0.56,1.22]	[1.22,2.38]	[2.38,4.00]	[4.00,+∞[
homogénéité-100] $-\infty$,0.67]	[0.67,0.74]	[0.74,0.80]	[0.80,0.87]	[0.87,+∞[
homogénéité-200] $-\infty$,0.67]	[0.67,0.74]	[0.74,0.80]	[0.80,0.87]	[0.87,+∞[
dissimilarité-100] $-\infty$,0.26]	[0.26,0.40]	[0.40,0.54]	[0.54,0.72]	[0.72,+∞[
dissimilarité-200] $-\infty$,0.26]	[0.26,0.40]	[0.40,0.54]	[0.54,0.72]	[0.72,+∞[
contraste-100] $-\infty$,0.27]	[0.27,0.44]	[0.44,0.72]	[0.72,1.05]	[1.05,+∞[
contraste-200] $-\infty$,0.27]	[0.27,0.44]	[0.44,0.72]	[0.72,1.05]	[1.05,+∞[

Hierarchies de valeurs



Premières expériences

algorithme M^3SP

PLANTEVIT M., LAURENT A., LAURENT D., TEISSEIRE M., CHOONG Y. W., *Mining Multidimensional and Multilevel Sequential Patterns*, ACM Transactions on Knowledge Discovery from Data TKDD, vol. 4, no 1, 2010.

- pas de dimension temporelle $D_T = \{\text{date}\}$, $\text{date} = 1$
- $D_R = \{\text{id-pt}\}$ (nb blocs = nb points)
- filtrage de l'ensemble d'apprentissage avant la fouille (site d'étude)

Premières expériences

Cinzana, $D_A = \{\text{type-de-culture}, \text{distance-village}\}$

id-pt	date	type-de-culture	distance-village
1	1	sol_nu	proche
2	1	sol_nu	proche
...
222	1	mais	proche
223	1	verger	eloigne

Premiers résultats

$$D_A = \{\text{type-de-culture, distance-village}\}$$

Site d'étude	Itemset	Support
Cinzana (223 points)	$s_1 = \langle \{(\text{culture}, \top_{\text{distance-village}})\} \rangle$	138 (62%)
	$s_2 = \langle \{(\text{culture}, \text{proche})\} \rangle$	121 (54%)
Koutiala (183 points)	$s_3 = \langle \{(\text{culture}, \top_{\text{distance-village}})\} \rangle$	105 (57%)
	$s_4 = \langle \{(\text{culture}, \text{proche})\} \rangle$	80 (44%)
Sikasso (92 points)	$s_5 = \langle \{(\text{culture}, \top_{\text{distance-village}})\} \rangle$	46 (50%)
	$s_6 = \langle \{(\text{culture}, \text{proche})\} \rangle$	27 (29%)

Interprétation

« Dans les 3 sites étudiés, les cultures sont généralement cultivées autour des villages, dans un rayon de 2 à 3 km pour la majorité. »

Premiers résultats

$$D_A = \{\text{type-de-culture, NDVI-100}\}$$

Site d'étude	Itemset	Support
Cinzana (223 points)	$s_7 = \langle\langle\langle\text{culture, très faible}\rangle\rangle\rangle$	74 (33%)
Koutiala (183 points)	$s_8 = \langle\langle\langle\text{culture, modéré}\rangle\rangle\rangle$	56 (31%)
	$s_9 = \langle\langle\langle\text{culture, faible}\rangle\rangle\rangle$	33 (18%)
Sikasso (92 points)	$s_{10} = \langle\langle\langle\text{culture, faible}\rangle\rangle\rangle$	25 (28%)
	$s_{11} = \langle\langle\langle\text{culture, modéré}\rangle\rangle\rangle$	20 (22%)

Interprétation

« Ceci reflète bien le gradient bioclimatique au Mali. Il pleut moins au Nord qu'au Sud, et donc les plantes ont une activité photosynthétique inférieure à Cinzana qu'à Sikasso, en moyenne.

D'autre part, pour le NDVI du mois de novembre, les cultures sont déjà entièrement récoltées à Cinzana, et partiellement récoltées à Koutiala et à Sikasso. »

Premiers résultats

$$D_A = \{\text{type-de-culture, variance-100,} \\ \text{homogénéité-100, dissimilarité-100, contraste-100}\}$$

Site d'étude	Itemset	Support
Cinzana (223 points)	$S_{12} = \{\{(\text{culture}, \top_{\text{variance-100}}, \top_{\text{homogénéité-100}}, \top_{\text{dissimilarité-100}}, \text{élevé})\}\}$	56 (25%)
Koutiala (183 points)	$S_{13} = \{\{(\text{culture}, \top_{\text{variance-100}}, \top_{\text{homogénéité-100}}, \top_{\text{dissimilarité-100}}, \text{faible})\}\}$	35 (19%)
Sikasso (92 points)	$S_{14} = \{\{(\text{culture}, \top_{\text{variance-100}}, \text{très élevé}, \text{très faible}, \text{très faible})\}\}$	23 (25%)
	$S_{15} = \{\{(\text{culture}, \top_{\text{variance-100}}, \top_{\text{homogénéité-100}}, \top_{\text{dissimilarité-100}}, \text{faible})\}\}$	18 (20%)

Interprétation

« Le contraste qui diminue du Nord au Sud peut s'expliquer par une différence de densité d'arbres dans les champs cultivés. Il serait en effet plus commun de trouver des arbres comme le Balanzan, le Néré ou le Karitier dans des champs de la région de Cinzana, qu'à Koutiala ou Sikasso, ce qui expliquerait les brusques changements de radiométrie, et donc un indice de contraste élevé. »

Conclusion et Perspectives

Travaux actuels

- Processus d'extraction de connaissances
 - ✓ acquisition des données (sauf indices images)
 - ✓ préparation des données
 - ✓ fouille de données : algorithme M^3SP (Plantevit, 2010)
 - ✗ classification
 - ✓ interprétation et validation

Perspectives de recherche

- appliquer sur des séries temporelles MODIS
- construire un classifieur qui exploite les motifs extraits et les connaissances du domaine ?
- (comment visualiser (sur une carte) les motifs extraits ?)

Cimetière de transparents

L'extraction de motifs fréquents

Détecter dans les données des motifs (ensembles d'items, séquences, arbres) qui interviennent fréquemment ensemble.

- $\mathcal{O} = \{o_1, o_2, \dots, o_n\}$ un ensemble d'objets
- $\mathcal{A} = \{a_1, a_2, \dots, a_m\}$ un ensemble d'attributs ou items
- $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{A}$, où $\mathcal{R}(o, a)$ signifie que l'objet o possède l'attribut a

Objets \ Attributs	7	8	9	10	valet	dame	roi	as
Jérôme	x				x	x		x
Jean-Jacques		x		x			x	x
Clémentine	x	x			x		x	
Thibaut		x		x		x	x	

L'extraction de motifs fréquents

- *motif* : ensemble d'attributs $m \subseteq \mathcal{A}$
- *support* d'un motif : nombre d'objets de \mathcal{O} qui partagent ce motif
- Un motif est dit *fréquent* si son support est supérieur à un seuil σ donné, appelé *seuil de support*.

Exemple

$m = \{8, 10, \text{roi}\}$ est un motif de support 2

Objets \ Attributs	7	8	9	10	valet	dame	roi	as
Jérôme	x				x	x		x
Jean-Jacques		x		x			x	x
Clémentine	x	x			x		x	
Thibaut		x		x		x	x	

L'extraction de motifs séquentiels multidimensionnels

- *séquentiels* : séquences de motifs ordonnés dans le temps
- *multidimensionnels* : dans des BDs multidimensionnelles

Exemple

$D_{\mathcal{A}} = \{\text{culture, NDVI, aire, ethnie}\}$

$m = \langle \{\text{mil}, [0; 0.2], [0; 30], \text{bambaras}\}, \{\text{mil}, [0.4; 0.6], [0; 30], \text{bambaras}\} \rangle$

pt	date	culture	NDVI	aire	arbres	ethnie
1	1	mil	[0 ;0.2]	[0 ;30]	0	bambaras
1	3	mil	[0.4 ;0.6]	[0 ;30]	1	bambaras
2	1	mil	[0.2 ;0.4]	[120 ;150]	1	bambaras
3	2	riz	[0 ;0.2]	[30 ;50]	3	dogons
3	3	sol nu	[0.6 ;0.8]	[180 ;+∞]	0	bambaras