Combinaisons de boules de mots pour la classification de séquences

Frédéric TANTINI, Alain TERLUTTE et Fabien TORRE

LORIA Nancy, INRIA Lille Nord Europe, CNRS LIFL

CAp 2010, Clermont-Ferrand

Plan de l'exposé

Classification supervisée avec VOLATA : points à définir

1. Représentation des objets à classer $\mathcal{X}$;
2. les classes discrètes à prédire $\mathcal{Y}$;
3. nature des hypothèses $\mathcal{H}$: $h \in \mathcal{H} : \mathcal{X} \to \mathcal{Y}$;
4. relation de subsomption $\succeq$ entre hypothèses.

Suffisent à définir une notion de *moindre généralisé*
et à instancier les méthodes d'apprentissage génériques de VOLATA.

Moindres généralisés

Définition : moindre généralisé

Étant donné un ensemble d'exemples $E \subseteq \mathcal{X}$, une hypothèse $h \in \mathcal{H}$ est dite *moindre généralisée* de E si et seulement si :

- $\forall e \in E : h \succeq e$;
- il n'existe pas h' vérifiant $\forall e \in E : h' \succeq e$ et $h \succeq h'$.

En cas d'unicité...

Deux vues algorithmiques possibles :

- $\mathrm{mg}(e_1, e_2, \ldots, e_n \in \mathcal{X})$ returns $h \in \mathcal{H}$;
- $\mathrm{mg}(h_{n-1}, e_n)$ returns $h \in \mathcal{H}$.

On préfère la deuxième version, plus efficace pour l'apprentissage.

VOLATA : une architecture à trois niveaux

1. Fournit l'opération mg permettant de calculer l'hypothèse moindre généralisée d'un ensemble d'exemples quelconque, découle de $\mathcal{H}$ et $\succeq$;
2. contrôle les généralisations de mg vis-à-vis des classes des exemples ; exemple : gc produit des hypothèses correctes, dépendantes de l'ordre des exemples ;
3. permet l'apprentissage d'un classifieur complet ; exemples : les méthodes d'ensemble GloBoost, Bagging-MG et AdaBoost-MG combinent les hypothèses fournies par gc ; importance de la diversité des hypothèses.

Seul le premier dépend des langages de représentation $\mathcal{X}$ et $\mathcal{H}$.

Cadres standard et moins standard

Cas idéal : moindre généralisé unique et calculable

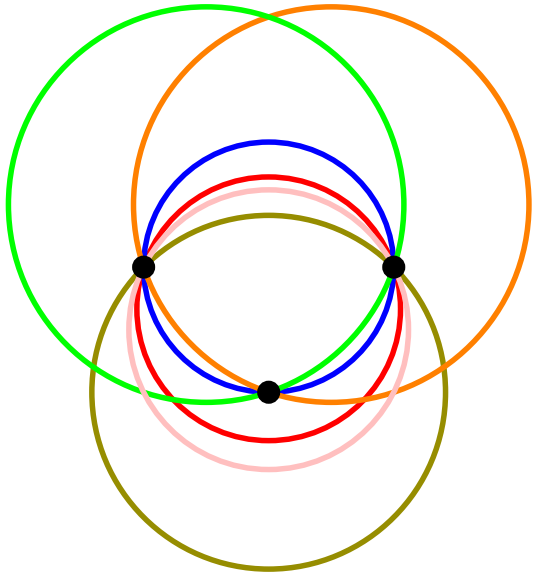
- rectangles en attributs-valeurs [Torre, 2005] ;
- automates (et preuves d'apprenabilité) pour la classification de séquences [Torre and Terlutte, 2009].

Deux difficultés possiblement induites par $(\mathcal{H}, \succeq)$

1. il existe un moindre généralisé unique mais son calcul est de complexité exponentielle (cadre relationnel, [Decoster et al., 2010]) ;
2. les moindres généralisés sont multiples.

Multiplicité des moindres généralisés : exemple et pistes

Une infinité de cercles moindres-généralisés pour capturer 3 points.



Prises en compte possibles des moindres généralisés multiples

- en maintenir plusieurs (*beam search*)... « pas VOLATA » ;
- choisir l'un des moindres généralisés ;
- considérer une autre généralisation.

Plan

Les boules de mots comme représentation de langages

Boules : définitions [Tantini, 2009]

- trois opérations d'édition de coût unitaire :
  - insertion : $aab \rightarrow aa\underline{b}b$
  - suppression : $a\underline{a}b \rightarrow ab$
  - substitution : $a\underline{a}b \rightarrow a\underline{b}b$
- distance d'édition : $d(w_1, w_2)$ est le nombre minimal d'opérations qui permettent de passer de w_1 à w_2 ;
- un langage est donné par un mot-centre et un rayon : $e \in B_2(bab)$ ou $B_2(bab) \succeq e$ ssi $d(bab, e) \leq 2$.

Remarques

- apprenabilité, positifs seuls, présence de bruit [Tantini, 2009] ;
- une boule de mots dénote un langage fini ;
- positionnement vis-à-vis des automates ?

Boules moindres généralisées multiples

Échantillon à moindres généralisés multiples

- Soit l'échantillon $E = [a, b, ab]$;
- $h = B_1(a)$ subsume les trois exemples de E ;
- $h' = B_1(b)$ subsume les trois exemples de E ;
- h contient aa, h' contient bb : h et h' sont incomparables ;
- chacune est un moindre généralisé (par exemple $B_1(\epsilon)$, incluse à la fois dans h et dans h', ne subsume pas $ab \in E$).

Nous ne sommes pas dans le cadre d'application idéal de VOLATA.

Distinguer une boule particulière

Distinguer une boule *moindre généralisée*

Celle de rayon minimal ?
Le théorème de [de la Higuera and Casacuberta, 2000] indique que trouver le centre de la plus petite boule contenant un ensemble fini de mots est *NP*-difficile.

Distinguer une boule *non moindre généralisée*

Celle centrée sur le premier exemple ?
Perte sévère de la diversité et de l'intérêt de la méthode VOLATA.

gball, proposition d'un algorithme de généralisation

Require: $e \in \mathcal{X}$ un exemple, $h = B_r(o) \in \mathcal{H}$ une hypothèse.
Ensure: $g \in \mathcal{H}$ une généralisation de e et h ($g \succeq h$ et $g \succeq e$).

1: $c = o \xrightarrow{*} e$ {chemin de longueur minimale}
2: soient x, y entiers et u un mot tels que $o \xrightarrow{x} u \xrightarrow{y} e$
3: $x = d(o, u)$, $y = d(u, e)$
4: $k = \max(r + x, y)$
5: **return** $B_k(u)$

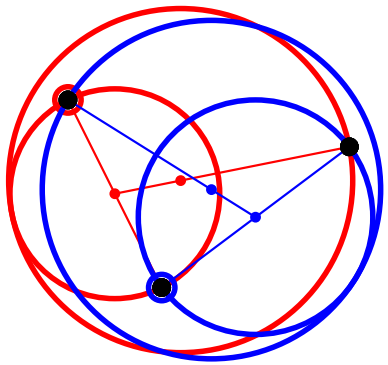
ajout de $e = \textit{chiens}$ à $h = B_3(\textit{niche})$

- chemin d'édition : *niche* ; *iche* ; *che* ; *ches* ; *chens* ; *chiens* ;
- $x = 1$, $y = 4$;
- nouveau centre u : *iche* ;
- nouvelle hypothèse proposée par gball(e, h) : $B_4(\textit{iche})$.

Généralisation en boules : stratégies

Différentes stratégies possibles :

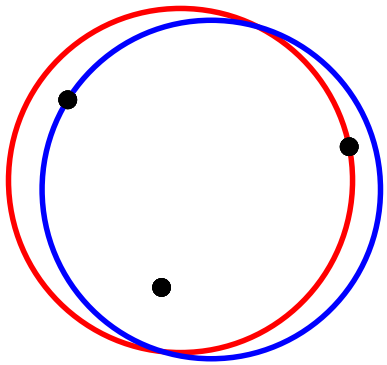
- *conservatrice* ($u = o$ ou $x = 0$) : stratégie déjà évoquée ;
- *conservatrice modérée* ($x = 1$) ;
- *aléatoire* ;
- *naturelle* ($r + x = y$) :



Plan

Sensibilité à l'ordre

Stratégie *naturelle* ($r + x = y$) :



Sensibilité à l'ordre avant la prise en compte des classes par gc !

Généralisation monotone...

Si $g = \text{gball}(e, h)$ alors $g \succeq e$ et $g \succeq h$.

Démonstration.

Rappel de l'algorithme : $k = \max(r + x, y)$.

- $g \succeq e$ car $d(u, e) = y \leq k$ et donc $e \in B_k(u)$;
- $g \succeq h$ car, pour tout mot $w \in B_r(o)$ et par l'inégalité triangulaire :

$$d(u, w) \leq d(u, o) + d(o, w)$$

par suite $d(u, w) \leq x + r \leq k$ et $w \in B_k(u)$.

... mais pas moindre généralisée

Contre-exemple

- Soit $E = [a, b]$.
- première hypothèse = premier exemple = $h = B_0(a)$;
- deux possibilités pour choisir u sur le chemin $a \xrightarrow{1} b$;
- soit $\text{gball}(h, b) = B_1(a)$, soit $\text{gball}(h, b) = B_1(b)$;
- or la boule $B_1(\epsilon)$ contient bien E ;
- et est plus spécifique que les hypothèses calculées : $B_1(a) \succeq B_1(\epsilon)$ et $B_1(b) \succeq B_1(\epsilon)$.

Généralisation correcte non monotone

Rappel : gc généralise les exemples à l'aide de gball en testant systématiquement la correction.

Exemple

- Soient $E^+ = [\epsilon, b, a]$ et $E^- = [bb]$
- gc avec gball et $x = 1$ produisent successivement :
  - $B_0(\epsilon)$ (hypothèse initiale) ;
  - $B_1(b)$ (rejetée car elle couvre bb) ;
  - $B_1(a)$ (acceptée).
- or $B_1(a) \succeq b$ alors que l'ajout du mot b a été rejeté.

Conséquence : des implémentations moins efficaces d'algorithmes comme AdaBoost.

Récapitulatif des propriétés de gball

Comparaison avec un moindre généralisé unique :

	mg	gball
moindre généralisation	✓	✗
sensibilité ordre	✗	✓
monotonie de g	✓	✓
monotonie de gc	✓	✗

Plan

Expérimentations

Méthode utilisée : GloBoost + gball et stratégie *aléatoire*.

Jeux de données séquentiels de l'UCI [Blake and Merz, 1998]

- morpion, badges, prénoms américains, génomique ;
- concurrents :
  - Majorité, RPNI, Traxbar, Redblue ;
  - GloBoost + mgtssi et GloBoost + mgzr ;
- validation croisée 10 fois, 90% des données en apprentissage.

Jeu de données réel : reconnaissance de chiffres manuscrits

- base *Nist special database 3* ;
- 10 classes, 10 568 exemples ;
- validation croisée 10 fois, 10% des données en apprentissage ;
- concurrent : SeDiL [Boyer et al., 2008].

Jeux de données UCI

	morpion	badges	promoters	first-name	splice
Majorité	65.34 %	**71.43 %**	50.00 %	81.62 %	50.26 %
RPNI	91.13 %	62.24 %	-	81.42 %	-
TraxBar	90.81 %	57.48 %	56.60 %	81.37 %	**58.33 %**
Red-Blue	**93.89 %**	61.09 %	**63.02 %**	**82.83 %**	54.65 %
$\text{Gmgtssi}_{1\,000}$	91.47 %	**72.69 %**	**61.13 %**	**89.50 %**	**78.07 %**
$\text{Gmgzr}_{1\,000}$	**98.36 %**	71.43 %	50.00 %	83.07 %	-
$\text{Ggball}_{1\,000}$	92.95 %	80.41 %	87.63 %	87.10 %	93.76 %
$\text{Ggball}_{10\,000}$	94.69 %	**81.39 %**	88.43 %	88.80 %	**95.63 %**
$\text{Ggball}_{100\,000}$	**94.96 %**	81.36 %	**89.08 %**	**89.06 %**	95.62 %

Observations

- rapidité et diversité ;
- gains significatifs sur les problèmes de génomique.

Reconnaissance de chiffres manuscrits

SeDiL	**95.86 %**
$\mathrm{Ggball}_{1\,000}$	93.81 %
$\mathrm{Ggball}_{10\,000}$	95.93 %
$\mathrm{Ggball}_{100\,000}$	**96.32 %**

Bonnes performances de VOLATA malgré un protocole défavorable.

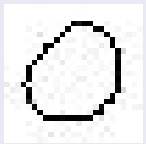
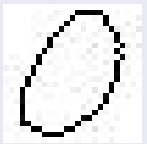
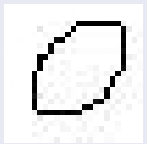
Une boule de chiffres

Une boule apprise pour la classe `zero`

$$B_{19}(22333234344444556656667660000000012111702311)$$

visualisation du centre :

Couverture de 36 exemples de zéro

etc.

tous à distance 19.

Boules de parties de morpions

Boule correcte apprise :

$$B_5^+(xxxoxbxxbxxx)$$

couvrant 80 parties gagnantes (toutes à distance de 5 du centre) :

- xxxoxboob
- xxxoboobx
- xxxobobox
- xxxobboxo
- xxoxbbxoo
- xxooxbobx
- xxooxbbox
- xxobxboxo
- xoxoxbxob
- xoxoxbobx
- xxxooboxb
- xxxobobxo
- xxxobbxoo
- xxxobboox
- xxooxboxb
- xxooxbbxo
- xxobxobox
- xoxoxbxbo
- xoxobxbox
- etc.

Boules de prénoms

L'une des boules apprises :

$$B_7(LRLRTSVKCA)$$

qui couvre 346 prénoms féminins (tous à distance 7) et pas un masculin :

- ALBERTHA
- BERTA
- DRUSILLA
- ELSA
- FRANCESCA
- HORTENSIA
- JESSIKA
- KRYSTINA
- LORENZA
- MIRTA
- NERISSA
- OCTAVIA
- PARTICIA
- REBBECA
- SYLVIA
- TERESSA
- URSULA
- VERONICA
- etc.

Dimension VC des boules

Theorem ([Janodet, 2010])

La VC-dimension des boules, sur un alphabet $|\Sigma| = 2$, est infinie.

Éléments de preuve pour $n = 5$

- cinq mots : *baaaa*, *abaaa*, *aabaa*, *aaaba*, *aaaab* ;
- un étiquetage avec deux positifs : *baaaa*$(+)$, *abaaa*$(+)$, *aabaa*$(-)$, *aaaba*$(-)$, *aaaab*$(-)$;
- la boule $B_2(bbaaa)$ contient les positifs et pas les négatifs.

Remarques

- preuve généralisable à tout n et pour tout étiquetage ;
- la preuve utilise une boule creuse, sans par cœur ;
- une boule creuse *ressemble* à un moindre généralisé.

Bilan des boules dans VOLATA et perspectives

Pour les boules

- généralisation et classification rapides ;
- grande diversité ;
- *on peut en calculer beaucoup* ;
- bons résultats expérimentaux ;

Perspectives : boules creuses ? applications en génomique ?

Pour VOLATA

- intégration réussie d'un calcul de généralisation qui n'avait a priori pas les bonnes propriétés ;

Perspective : autres classes avec moindres généralisés multiples ?

Bibliographie I


Blake, C. and Merz, C. (1998).
UCI repository of machine learning databases
[http ://archive.ics.uci.edu/ml/].


Boyer, L., Esposito, Y., Habrard, A., Oncina, J., and Sebban, J. (2008).
Sedil : Software for edit distance learning.
In Daelemans, W., Goethals, B., and Morik, K., editors, *Proceedings of the 19th European Conference on Machine Learning*, pages 672–677. Springer.


de la Higuera, C. and Casacuberta, F. (2000).
Topology of strings : median string is NP-complete.
Theoretical Computer Science, 230 :39–48.

Bibliographie II


Decoster, J., Staworko, S., and Torre, F. (2010).
Apprentissage relationnel polynomial pour la classification d'arbres.
In Mephu Nguifo, E., editor, *12ème Conférence francophone sur l'Apprentissage automatique (CAp'2010)*, pages 189–200, Clermont-Ferrand. PUG.


Janodet, J.-C. (2010).
The vapnik-chervonenkis dimension of balls of strings is infinite.
Personal Communication.


Tantini, F. (2009).
Inférence grammaticale en situations bruitées.
PhD thesis, Université Jean Monnet de Saint-Étienne.

Bibliographie III


Torre, F. (2005).
Globoost : Combinaisons de moindres généralisés.
Revue d'Intelligence Artificielle, 19(4-5) :769–797.


Torre, F. and Terlutte, A. (2009).
Méthodes d'ensemble en inférence grammaticale : une approche à base de moindres généralisés.
In Bennani, Y. and Rouveirol, C., editors, *11ème Conférence francophone sur l'Apprentissage automatique (CAp'2009)*, pages 33–48, Hammamet (Tunisie). PUG.