

Volata, système de classification tout terrain

Fabien Torre

Mostrare (INRIA Lille Nord Europe et CNRS LIFL)
Université Lille Nord de France

VOLATA¹ vise à apprendre à classer, à l'aide d'hypothèses venant d'un ensemble \mathcal{H} , des objets issus d'un ensemble \mathcal{E} que l'on peut supposer inclus dans \mathcal{H} . De plus, pour comparer les éléments de \mathcal{H} , nous nous donnons une relation de généralité notée \succeq .

Sur ces hypothèses, le système VOLATA est organisé en trois niveaux algorithmiques :

- le premier niveau fournit une implémentation de \succeq et d'une opération MG permettant de calculer l'hypothèse moindre généralisée d'un ensemble d'exemples quelconque (MG découle directement des choix de \mathcal{H} et \succeq) ;
- le deuxième prend en compte les classes des exemples pour produire des hypothèses correctes, ou quasi-correctes si du bruit de classe est présent (MGC) ;
- le dernier niveau permet l'apprentissage d'un classifieur complet, par combinaison de règles élémentaires apprises au niveau précédent.

Il est important de noter que seul le premier niveau dépend des langages de représentation \mathcal{E} et \mathcal{H} . Seules les opérations \succeq et MG sont donc à définir pour bénéficier de l'architecture globale et en particulier des méthodes du dernier niveau : techniques gloutonnes rapides, algorithmes minimisant le nombre de règles apprises pour une meilleure compréhension, méthodes d'ensemble comme le *bagging* et le *boosting* pour de meilleures prédictions.

À travers VOLATA, cette méthodologie est mise en œuvre dans des domaines distincts de la classification supervisée :

- la classification attribut-valeur est considérée dans (Torre, 2004) avec une généralisation des exemples-vecteurs sous forme d'hyper-rectangles ;
- la classification de séquences est envisagée sous deux angles :
 - avec \mathcal{H} qui est une classe d'automates connue en inférence grammaticale et dotée d'un algorithme de généralisation MG, par exemple les automates *réversibles* de Angluin (1982) ou les *k-TSS* de García & Vidal (1990), dans (Torre & Terlutte, 2009) ;
 - avec les hypothèses qui sont des *boules de mots* dans (Tantini *et al.*, 2010) ;

¹<http://www.grappa.univ-lille3.fr/~torre/Recherche/Softwares/volata/>

- pour la classification d’arbres, **Decoster et al. (2010)** utilisent un codage des arbres en logique du premier ordre qui permet des algorithmes polynomiaux pour le calcul de \succeq et MG (il s’agit ici de la θ -subsomption et du *lgg* définis par **Plotkin (1970)** et de complexité exponentielle en toute généralité) ;
- la classification de graphes est traitée à l’aide des opérations de subsomption et de généralisation proposées par **Liquière (2007)**.

Ainsi, selon le domaine d’application, l’opération MG choisie et l’apprenant de haut niveau impliqué, VOLATA produit pour chaque classe une combinaison (pondérée ou non) d’hyper-rectangles, d’automates, de boules de mots, de motifs d’arbres ou de graphes.

La démonstration proposée permettra d’appréhender les différentes variations possibles avec VOLATA et de lancer quelques expérimentations au gré des interlocuteurs. Il sera également possible de discuter certaines théories préalablement apprises sur des problèmes classiques de la classification supervisée.

Références

- ANGLUIN D. (1982). Inference of reversible languages. *Journal of the ACM*, **29**(3), 741–765.
- DECOSTER J., STAWORKO S. & TORRE F. (2010). Apprentissage relationnel polynomial pour la classification d’arbres. In E. MEPHU NGUIFO, Ed., *12ème Conférence francophone sur l’Apprentissage automatique (CAp’2010)*, Clermont-Ferrand : PUG.
- GARCÍA P. & VIDAL E. (1990). Inference of k-testable languages in the strict sense and application to syntactic pattern recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, **12**(9), 920–925.
- LIQUIÈRE M. (2007). Arc consistency projection : A new generalization relation for graphs. In U. PRISS, S. POLOVINA & R. HILL, Eds., *15th International Conference on Conceptual Structures (ICCS 2007)*, volume 4604 of *Lecture Notes in Computer Science*, p. 333–346 : Springer.
- PLOTKIN G. (1970). A note on inductive generalization. In B. MELTZER & D. MITCHIE, Eds., *Machine Intelligence*, volume 5, p. 153–165. Edinburgh University Press.
- TANTINI F., TERLUTTE A. & TORRE F. (2010). Combinaisons de boules de mots pour la classification de séquences. In E. MEPHU NGUIFO, Ed., *12ème Conférence francophone sur l’Apprentissage automatique (CAp’2010)*, Clermont-Ferrand : PUG.
- TORRE F. (2004). GloBoost : Boosting de moindres généralisés. In M. LIQUIÈRE & M. SEBBAN, Eds., *Actes de la Sixième Conférence Apprentissage CAp’2004*, p. 49–64 : Presses Universitaires de Grenoble.
- TORRE F. & TERLUTTE A. (2009). Méthodes d’ensemble en inférence grammaticale : une approche à base de moindres généralisés. In Y. BENNANI & C. ROUVEIROL, Eds., *11ème Conférence francophone sur l’Apprentissage automatique (CAp’2009)*, p. 33–48, Hammamet (Tunisie) : PUG.