Effets de la suppression des mots bruités en inférence grammaticale

Marc Sebban, Jean-Christophe Janodet et Aziz Yahiaoui

EURISE
Université Jean Monnet
UFR Sciences et Techniques, Département Informatique
42023 Saint-Etienne Cedex
$\{$*marc.sebban,janodet*$\}$*@univ-st-etienne.fr*

RÉSUMÉ. *Nous présentons dans ce travail une approche filter de réduction de données, visant à supprimer le bruit dans les modèles symboliques séquentiels. Les modèles étudiés ici sont des automates générés par inférence grammaticale à partir des mots d'un langage. Nous montrons qu'il est possible, en analysant les probabilités d'apparition des exemples, de détecter différentes catégories de mots non pertinents, catégories déja établies dans les modèles de réduction de données numériques. En supprimant ces exemples, il est ainsi possible de réduire la complexité des modèles inférés (taille des automates) et d'améliorer leurs performances en généralisation.*

MOTS-CLÉS : *apprentissage automatique, réduction de données, inférence grammaticale*

1. Introduction

Avec le développement des nouvelles technologies d'acquisition de données, les bases de données modernes, grandes mais souvent fortement bruitées, nécessitent des phases de pré-traitement afin d'être exploitables par des algorithmes d'apprentissage. La présence de données non pertinentes, mal typées, ou encore redondantes risque en effet d'avoir des conséquences sur la complexité et les performances des modèles induits. Durant la dernière décennie, de nombreux algorithmes de réduction de données en apprentissage automatique ont eu pour objectif de réduire la dimension de l'espace de représentation en sélectionnant les attributs pertinents [AHA 92, JOHN 94, LANGLEY 94] ou le nombre d'exemples d'apprentissage en ne retenant que les individus prototypes [SEBBAN 00, SKALAK 94, WILSON 97b]. Ces algorithmes sont généralement classés en deux sous-catégories [LANGLEY 94], appelées *wrapper*, pour ceux exploitant le taux de succès de l'algorithme d'apprentissage durant la sélection, et *filter*, pour ceux utilisant des mesures d'information ou des critères statistiques.

Il est à noter, et ce de manière surprenante, que la grande majorité de ces modèles est exclusivement réservée aux données numériques. Une des raisons vient probablement du fait que bon nombre de ces approches exploitent des distances pour mesurer leur critère, plus faciles à mettre en oeuvre dans des espaces continus, même si certains travaux ont proposé des fonctions de distance plus générales [WILSON 97a]. Les techniques *filter* de réduction de données deviennent même quasi-inexistantes dès qu'on aborde les données séquentielles de tailles variables, ce qui est le cas en inférence grammaticale, où les exemples à apprendre sont des mots. Les seules méthodes existantes sont en effet des modèles *wrapper*, autorisant le bruit pendant l'inférence (modèles stochastiques [CAR 94, THO 00]). Nous nous intéressons ici à la possibilité de supprimer les mots bruités en amont du processus d'inférence.

En inférence grammaticale, les objets manipulés sont donc des mots, et les concepts recherchés sont caractérisés par un certaine grammaire. L'objectif est de découvrir cette grammaire, sous la forme d'un automate, à l'aide d'un échantillon d'apprentissage, composé d'exemples positifs (représentant le concept à apprendre) et éventuellement d'exemples négatifs (contre-exemples du concept). Nous verrons que la présence de mots bruités a des impacts directs sur la complexité des automates inférés, et donc sur leur pouvoir de généralisation. Nous montrons dans cette présentation qu'il est possible de supprimer ces exemples *a priori*, en estimant la probabilité de chaque mot conditionnellement à sa classe d'appartenance, et en testant cette probabilité par rapport à un intervalle de confiance sur l'ensemble de mots négatifs et positifs. Les résultats expérimentaux montrent l'intérêt de notre approche.

2. Les exemples non pertinents

Il est d'usage, au sein de la communauté en réduction de données, de regrouper les exemples non pertinents en quatre catégories. La figure 1 représente ces exemples dans un espace de représentation continu.

1. Les exemples mal typés (ou *outliers*). Conserver ces exemples pénalisent considérablement les modèles inférés, qui risquent de faire de l'apprentissage par coeur (overfitting).

2. Les exemples situés aux frontières bayésiennes, et aux régions de chevauchement des classes (*overlap*). Ces exemples représentant les zones d'erreur bayésienne incompressible, les conserver n'apporte aucun intérêt.

3. Les exemples en marges des distributions. Il est d'usage en statistique descriptive de ne pas tenir compte de ces exemples potentiellement aberrants, comme c'est le cas dans les intervalles inter-quantiles.

4. Les exemples au centre de clusters. Dans des espaces numériques, retenir ces points n'apportent pas d'intérêt puisque un nouvel exemple à étiqueter dans cette région, serait correctement étiqueté par un modèle de prédiction.

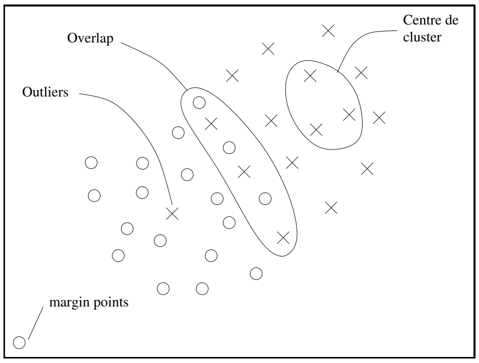

FIG. 1. *Exemples non pertinents destinés à la suppresion*

Il devient plus délicat de mettre en exergue les mêmes catégories d'exemples non pertinents dès que l'on travaille sur des mots d'un langage, décrit par un nombre d'attributs (les lettres) variable. Parmi ces catégories, il est néanmoins assez simple de décrire un mot *outlier*. Prenons l'exemple suivant : supposons que nous disposions de mots formés uniquement des "lettres" *0* et *1*. Supposons que le concept à apprendre soit la détection des mots contenant un nombre pair de *0* et de *1*. Même si cet exemple est intrinsèquement peu intéressant, il a le mérite de présenter clairement le problème du bruit dans les données. Partons d'un échantillon avec 9 exemples positifs et 11 négatifs :

$Ex^{+}=\{11, 00, 1001, 0101, 1010, 1000111101, 1001100001111010, 111111, 0000\}$
$Ex^{-}=\{0, 111, 010, 000000000, 1000, 01, 10, 1110010100, 010111111110, 0001, 011\}$

En appliquant un algorithme d'inférence grammaticale par fusion d'états (comme RPNI [ONC 92]) sur cette base de données non bruitée, on obtient l'automate présenté sur la figure 2.

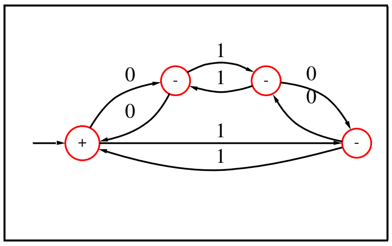

FIG. 2. *Automate inféré sur données non bruitées*

Supposons maintenant que l'exemple *00* devienne un *outlier* et soit transformé en exemple négatif, alors qu'il relève du concept à apprendre. L'effet sur la complexité de l'automate est immédiate. En effet, le nouvel automate inféré (voir figure 3) contient 7 états contre seulement 4 auparavant. La présence d'un seul *outlier* a détérioré le modèle et donc *a fortiori* sa capacité de généralisation.

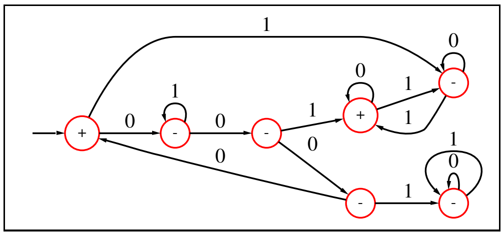

FIG. 3. *Automate inféré RPNI sur données bruitées*

3. Critère de sélection d'un mot

On définit un modèle stochastique comme un processus aléatoire qui peut changer d'états s_i à chaque étape q_t du temps. On peut donc représenter un mot $M = (q_1, q_2, ..., q_T)$ comme un tel processus, et supposer par soucis de simplification, qu'il vérifie également la propriété de Markov, c'est à dire :

$$P(q_1, q_2, ..., q_T) = P(q_1).P(q_2/q_1)....P(q_T/q_T - 1)$$

Ce qui veut dire que l'évolution globale est déterminée par la probabilité initiale et celles des transitions successives. Pour estimer la pertinence d'un mot, nous utiliserons donc dans notre approche sa probabilité d'apparition conditionnellement à sa classe. Par exemple, la probabilité d'apparition du mot *bb* sachant qu'il est positif sera :

$$p(bb/+) = p(b/+).p(b/b, +).p(fin/b, +)$$

A l'aide de cette probabilité conditionnelle, nous pouvons détecter aisément 3 des 4 catégories présentées précédemment.

1. Un mot M déclaré positif est un *outlier* si $p(M/+)$ est très "faible", et que p(M/-) est très "forte". Une valeur faible (resp. forte) représentera une valeur inférieure (resp. supérieure) à la borne basse (resp. haute) de l'intervalle de confiance (IC) autour de la moyenne des probabilités des exemples positifs (resp. négatifs). Le raisonnement sera inversé si M est connu comme étant négatif.

2. Un mot M positif est un point d'*overlap* si les probabilités $p(M/+)$ et $p(M/-)$ ne sont pas faibles, c'est à dire si elles appartiennent toutes deux à l'IC autour de la moyenne de leur classe. Le raisonnement est le même si M est connu comme étant négatif.

3. Un mot M positif est un *margin point* si $p(M/+)$ et $p(M/-)$ sont significativement faibles par rapport aux autres exemples de la classe, c'est à dire inférieures aux bornes basses des IC.

Un exemple d'application est présenté sur la figure 4. Sur un échantillon de 100 exemples contenant 8 mots bruités, notre stratégie permet de supprimer la quasi-totalité du bruit, et simplifier considérablement la complexité de l'automate. En effectuant une 10 cross-validation, la taille de l'automate avant sélection est de 17.3 ± 2.9 contre 9.5 ± 1.9 après. Son taux de succès en généralisation passe de 69.16 ± 8.1 à 90 ± 10.1. Des résultats plus significatifs sur de grandes bases de données seront présentés.

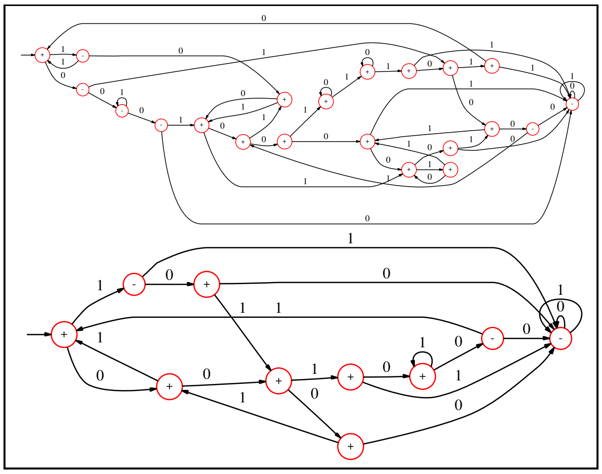

FIG. 4. *Automate construit sur l'échantillon d'apprentissage d'origine complet (haut) et réduit (bas)*

4. Bibliographie

[AHA 92] AHA D., Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms, *Intern. Journal Man Mach. Studies*, , 1992, p. 267-287.

[CAR 94] CARRASCO R. C., ONCINA J., Learning Stochastic Regular Grammars by Means of a State Merging Method, CARRASCO R. C., ONCINA J., Eds., *Grammatical Inference and Applications (ICGI-94)*, p. 139-152, Springer, Berlin, Heidelberg, 1994.

[JOHN 94] JOHN G., KOHAVI R., PFLEGER K., Irrelevant features and the subset selection problem, *Eleventh International Conference on Machine Learning*, 1994, p. 121-129.

[LANGLEY 94] LANGLEY P., Selection of relevant features in machine learning, *AAAI Fall Symp. on Relevance*, 1994.

[ONC 92] ONCINA J., GARCIA P., Identifying regular languages in polynomial time, 1992.

[SEBBAN 00] SEBBAN M., NOCK R., Instance Pruning as an Information Preserving Problem, *Seventeenth International Conference on Machine Learning*, 2000.

[SKALAK 94] SKALAK D., Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithms, *Eleventh International Conference on Machine Learning*, 1994, p. 293-301.

[THO 00] THOLLARD F., DUPONT P., DE LA HIGUERA C., Probabilistic DFA Inference using Kullback-Leibler Divergence and Minimality, *Proc. 17th International Conf. on Machine Learning*, Morgan Kaufmann, San Francisco, CA, 2000, p. 975–982.

[WILSON 97a] WILSON D., MARTINEZ T., Improved heterogeneous distance functions, *J. Arti. Intell. Res.*, , 1997, p. 1-34.

[WILSON 97b] WILSON D., MARTINEZ T., Instance Pruning techniques, *Fourteenth International Conference on Machine Learning*, 1997, p. 404-411.