Bagging de caractéristiques pour l'authentification d'auteur

François-Marie Giraud — Thierry Artières

Laboratoire d'Informatique de Paris 6 (LIP6)
Université Pierre et Marie Curie, Paris 6 - 4 place Jussieu F-75252 PARIS cedex 05
{francois.giraud, thierry.artieres}@lip6.fr

RÉSUMÉ. *Les travaux en authentification d'auteur ont montré la difficulté de dépasser une stratégie simple telle qu'un classifieur linéaire opérant sur des représentations de type sac de caractéristiques des documents. Nous proposons pour surmonter cette difficulté d'utiliser les techniques de bagging de caractéristiques qui reposent sur l'apprentissage d'un ensemble de classifieurs appris sur des sous-ensembles aléatoires de caractéristiques, puis sur le vote de ces classifieurs en test.*

ABSTRACT. *The authorship attribution literature demonstrates the difficulty to design classifiers that outperform simple strategies such as linear classifiers operating on bag of features representation of documents. To overcome this difficulty we propose to use Bagging techniques that rely on learning classifiers on different random subsets of features, then to combine their decision by making them vote.*

MOTS-CLÉS : *Authentification d'auteur, Bagging de caractéristiques*

KEYWORDS: *Author identification, feature bagging*

1. Introduction

Une question clé dans le domaine de l'attribution et de la vérification d'auteur réside dans le choix de caractéristiques pertinentes, cela a motivé de nombreuses études (Stamatatos, 2009), (Koppel *et al.,* 2009), (Savoy, 2012). Malgré de nombreux efforts pour construire des caractéristiques pertinentes du style d'un auteur (e.g. (Kim *et al.,* 2011)) un choix très populaire reste de considérer un ensemble important de caractéristiques simples issues du monde de la recherche d'information (Comptages ou TFIDF de mots et/ou de n-grammes de caractères), dont on sélectionne une partie a priori pour limiter la complexité de l'apprentissage. Au delà du critère de fréquence (très utilisé) on peut réaliser la sélection de caractéristiques via d'autres critères (gain d'information, information mutuelle, corrélation χ^2, ...) (Savoy, 2012).

Les classifieurs les plus souvent rencontrés pour l'authentification d'auteurs sont des modèles linéaires tels que des machines à vecteurs support (SVM). Ces classificateurs opérant sur des représentations de type sac de caractéristiques des documents apparaissent en réalité assez difficiles à battre (Koppel *et al.,* 2009).

Notre travail est une tentative de dépasser cette approche simple mais efficace. Il est inspiré de deux constats. Tout d'abord, il a été observé dans (Sutton *et al.,* 2005) que l'apprentissage de modèles linéaires ou log-linéaires (dans lesquels il y a un paramètre par caractéristique) pouvait conduire à une forme de sous apprentissage, où certaines caractéristiques pertinentes pour la tâche de classification ne sont en fait pas prises en compte par le modèle après apprentissage (e.g. les poids correspondants sont proches de zéro). Cela peut se produire quand un petit nombre de caractéristiques est suffisant, à lui seul, à la discrimination parfaite des échantillons d'apprentissage. Dans ce cas, l'apprentissage d'un modèle linéaire peut se concentrer sur l'apprentissage des paramètres correspondant à ces caractéristiques et négliger d'autres caractéristiques éventuellement pertinentes sur les données d'apprentissage. Cela peut conduire à des erreurs en test qui pourraient être évitées si toutes les caractéristiques discriminantes étaient exploitées par le modèle après apprentissage. Deuxièmement, le travail par (Koppel *et al.,* 2007) suggère que le style d'un auteur peut être caractérisé par la façon dont se comporte le classifieur lorsque les caractéristiques les plus importantes (de poids fort) sont progressivement ignorées.

Nous proposons ici d'étudier comment exploiter ces deux résultats pour concevoir des classifieurs performants pour l'attribution d'auteur. Notre approche repose sur le bagging de caractéristiques où l'on combine les résultats d'un certain nombre de classifieurs qui sont appris chacun sur des sous-ensembles aléatoires (de taille limitée) de caractéristiques. Nous présentons tout d'abord un aperçu des travaux connexes dans la section 2 puis nous présentons les jeux de données dans la section 3, et en particulier les jeux de données du challenge PAN 2012 auquel nous avons participé. Ensuite, nous motivons et présentons l'idée générale du bagging de caractéristiques et en étudions l'intérêt dans des expériences en section 4. Enfin nous présentons une extension de l'approche par Bagging exploitant le résultat de (Koppel *et al.,* 2007) en section 4.2.

2. Etat de l'art

Les méthodes de classification les plus utilisées sont des méthodes linéaires, très souvent des machines à vecteurs supports ou SVMs. La littérature montre en effet que les SVM linéaires sont particulièrement efficaces dans le domaine de l'identification de l'auteur (comme ils le sont pour des tâches de catégorisation de textes) (Koppel *et al.,* 2009), et cela pour de nombreux types de caractéristiques. Seuls quelques travaux ont conclu a la supériorité de SVM non linéaires pour cette tâche (e.g. (Teytaud *et al.,* 2000)). En fait peu de travaux portent sur le développement de nouvelles méthodes de classification. La plupart des études publiées visent à construire de nouvelles caractéristiques pertinentes pour l'attribution d'auteur (e.g. (Kim *et al.,* 2011)), ou bien comparer des jeux de caractéristiques (Koppel *et al.,* 2009), (Stamatatos, 2009), ou enfin à proposer des méthodes de sélection des caractéristiques les plus pertinentes (Savoy, 2012).

La conception de bonnes caractéristiques semble être la vraie question clé pour l'identification d'auteur. Parmi les caractéristiques couramment employées on distingue des caractéristiques lexicales, syntaxiques, structurelles et contextuelles. Les caractéristiques lexicales incluent les tfidf, la longueur des mots, des phrases, la richesse du vocabulaire (Koppel *et al.,* 2009), les n-grammes de mots ou de caractères (Hoover, 2002). Ces caractéristiques étant souvent très nombreuses on en sélectionne un sous ensemble par des critères du type gain d'information (Savoy, 2012). Les caractéristiques syntaxiques sont des comptages ou des tfidf sur des mots particuliers (mots de liaison, conjonction, préposition, pronom, verbes modaux) ou sur des étiquettes POS. On peut également utiliser des n-grammes de POS tags (Argamon-Engelson *et al.,* 1998). Les caractéristiques structurelles concernent la taille de la police, la couleur, le nombre d'hyperliens etc (Abbasi *et al.,* 2005). Enfin les caractéristiques contextuelles concernent les sujets abordés par le document, l'élongation ou l'inflexion dans la langue arabe etc (Abbasi *et al.,* 2005). Parmi les nombreux travaux comparant la pertinence des divers jeux de caractéristiques, les conclusions sont souvent contradictoires. Par exemple les caractéristiques les plus discriminantes sont des caractéristiques lexicales dans (Koppel *et al.,* 2009) alors que ce sont des caractéristiques contextuelles dans l'étude de (Abbasi *et al.,* 2005).

Il semble d'après l'ensemble de la littérature sur le sujet que la complexité de l'authentification d'auteur vient d'une part du fait que la signature de l'auteur est mêlée à d'autres informations, plus prégnantes, qui concernent la nature du document, le sujet du texte, l'opinion de l'auteur, etc. Si bien que les caractéristiques lexicales sont souvent efficaces et qu'il est délicat de sélectionner manuellement des caractéristiques intuitivement liées au style. En outre, il est probable que les caractéristiques les plus discriminantes pour un auteur soient très dépendantes de l'auteur, la sélection de caractéristiques ne pourrait alors être réalisée efficacement a priori pour l'ensemble des auteurs.

3. Contexte expérimental

3.1. *Jeux de données*

Nous présentons des résultats expérimentaux obtenus sur les jeux de données de la compétition internationale PAN 2012 et sur deux ensembles de données additionnels sur lesquels nous avons cherché à caractériser le comportement de nos propositions.

– Corpus de *littérature anglaise*. Ce corpus a été utilisé dans certaines publications antérieures (Koppel *et al.,* 2009), (Koppel *et al.,* 2007). Il comprend 2 livres complets pour 9 auteurs. Chaque livre a été divisé manuellement en une centaine de documents, en gardant l'intégrité des chapitres et des sections de texte. Les documents obtenus sont longs d'environ 500 à 3000 mots.

– Corpus de *blogs*. Ce corpus compte originellement environ 18 000 auteurs (Koppel *et al.,* 2006), nous n'avons considéré que les 60 principaux auteurs, ceux qui ont posté au moins 20 messages de plus de 100 mots.

– Corpus de la compétition *PAN 2012*. La compétition PAN 2012 a proposé plusieurs tâches d'attribution d'auteur, nous nous sommes focalisés sur la tache traditionnelle d'attribution d'auteur de textes littéraires. Il y a avait trois corpus d'apprentissage (deux de nouvelles littéraires, le dernier de romans). Pour chacun des corpus il existait deux sous-tâches, l'une dite fermée ou un texte de test était écrit par l'un des auteurs du corpus d'apprentissage qu'il fallait reconnaître, l'autre ouverte dans laquelle cette condition n'était pas remplie (le classifieur devait donc être capable de rejeter un document). Nous avons essentiellement participé aux tâches fermées. Les jeux de données étaient les suivants :

- Corpus 1 : 3 auteurs, 2 documents par auteurs, entre 1700 et 6000 mots par documents.

- Corpus 2 : 8 auteurs, 2 documents par auteurs, entre 1900 et 13000 mots par documents.

- Corpus 3 : 14 auteurs, 2 documents par auteur, entre 30 000 et 170 000 mots par documents.

3.2. *Conditions expérimentales*

Dans toutes les expériences rapportées, nous avons utilisé des classifieurs SVMs linéaires. Nous avons utilisé la bibliothèque libsvm (Chang *et al.,* 2011) où un classificateur multiclasse pour un problème de classification à N classes est mis en oeuvre par l'apprentissage de N $\times(N-1)/2$ SVM binaires. Tous les classifieurs sont appris avec un terme de régularisation L2 standard dont le poids est fixé sur la base de validation.

4. Bagging de caractéristiques

Nous considérons un problème de classification à N auteurs (classes), les documents sont représentés par des vecteurs à p dimensions.

4.1. *Motivation*

Il est connu que l'apprentissage de modèles de classification de forte capacité (par exemple des modèles linéaires sur des données en très grande dimension) peut conduire à une forme de sur-apprentissage, où le classifieur peut parfaitement discriminer les données d'apprentissage et avoir un comportement dégradé en généralisation (sur des données non vues en apprentissage). Dans certaines situations ce phénomène a également été qualifié de sous-apprentissage (Sutton *et al.,* 2005). Ce dernier suggère que certaines caractéristiques pertinentes peuvent ne pas être pleinement prises en compte par le modèle après apprentissage. Cela peut se produire quand plusieurs caractéristiques (ou plusieurs sous-ensembles de caractéristiques) sont suffisantes à elles seules pour une discrimination parfaite des échantillons d'apprentissage. Dans ce cas, l'apprentissage peut se concentrer sur l'exploitation de certaines de ces caractéristiques pertinentes tout en en négligeant d'autres. Du point de vue de l'apprentissage automatique un apprentissage de ce type, que l'on pourrait qualifier de partiel, correspondrait effectivement à une maximisation du critère (de marge, de probabilité a posteriori) et constituerait une solution valide. Malheureusement, si les caractéristiques discriminantes qui ont été négligées durant l'apprentissage apparaissent dans un échantillon de test, et pas les caractéristiques discriminantes sur lesquelles le classifieur a appris à construire sa prédiction, alors cet exemple de test sera mal classé. Alors qu'il inclut des caractéristiques pertinentes qui pourraient avoir été utilisées pour bien le classer. C'est en cela que cette forme de sur-apprentissage peut être interprétée comme du sous-apprentissage.

Ce type de phénomène a été observé en particulier dans le contexte du traitement de données textuelles avec des modèles log-linéaires (type CRFs ou Conditional Random Fields) qui sont traditionnellement appris avec un très grand nombre de caractéristiques pour des tâches où les données sont parfois linéairement séparables avec un petit sous-ensemble des caractéristiques.

En réalité nous avons effectivement observé sur une petite dizaine de corpus d'authentification d'auteurs que des SVMs travaillant sur des représentations des documents en grande dimension (e.g. de caractéristiques lexicales) atteignent très souvent une précision de 100% sur l'ensemble d'apprentissage tandis que les performances sur l'ensemble de test peuvent être nettement inférieures, ce qui est symptomatique d'un surapprentissage mais peut être mieux appréhendé comme du sousapprentissage.

Les figures 1 et 2 montrent la précision obtenue par un SVM linéaire en fonction du nombre (limité) de caractéristiques, X, utilisées pour représenter les documents. La valeur de X varie de 10 à 350. Les caractéristiques sont choisies parmi un ensemble de

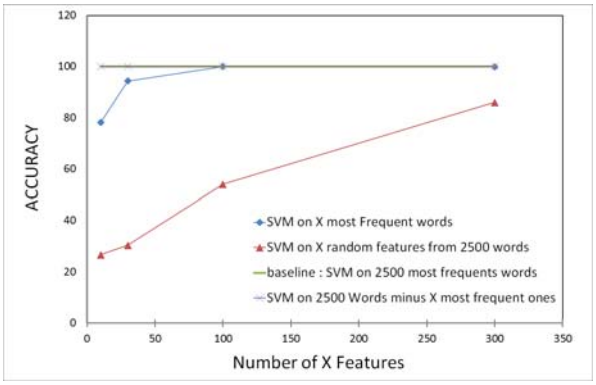

Figure 1. *Corpus de Littérature Anglaise (Koppel* et al., *2009). Taux de classification correcte sur les données d'apprentissage de SVMs linéaires en fonction du nombre X de caractéristiques utilisées, choisies au hasard ou les plus fréquentes (parmi un ensemble de 2 500 caractéristiques). La performance d'un SVM utilisant toutes les (2 500) caractéristiques ainsi que celle d'un SVM utilisant toutes les caractéristiques exceptées les X les plus fréquentes sont données également.*

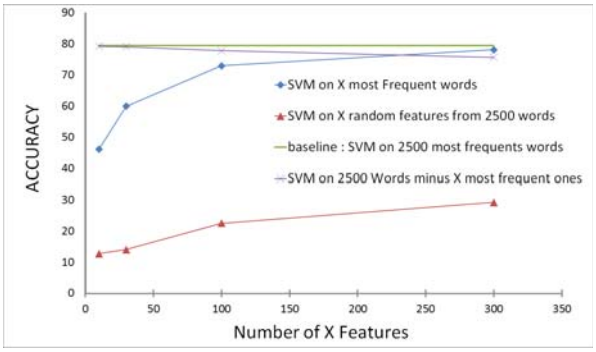

Figure 2. *Mêmes courbes que dans la figure 1 mais sur les données de test.*

2 500 caractéristiques qui sont les 2 500 trigrammes de caractères les plus fréquents. Les courbes sont tracées pour l'ensemble de données d'apprentissage (figure 1) et de test (figure 2) pour le cas où les X caractéristiques sont choisies aléatoirement parmi les 2 500 caractéristiques initiales ou bien par sélection des X caractéristiques les plus fréquentes. Deux courbes additionnelles sont données pour comparaison. L'une correspond à la performance d'un SVM exploitant les 2 500 caractéristiques l'autre

à la performance d'un SVM exploitant les 2 500 caractéristiques moins les X plus fréquentes. Ces résultats ont été obtenus sur le corpus de *Littérature Anglaise*.

On voit tout d'abord que les SVMs opérant sur l'ensemble des caractéristiques obtiennent une précision de 100% sur la base d'apprentissage aors que la performance plafonne à 80% en test.

Ensuite, la performance des classifieurs qui exploitent seulement quelques caractéristiques est très élevée sur l'ensemble d'apprentissage si l'on sélectionne les caractéristiques les plus fréquentes. Et si on augmente le nombre de caractéristiques utilisées la performance atteint rapidement une classification parfaite en apprentissage, (ce qui est également vrai lorsque l'on utilise toutes les caractéristiques), tandis qu'elle atteint le même plateau en test à environ 80% de précision. Il y a donc un écart clair entre la performance sur l'ensemble d'apprentissage et sur l'ensemble de test. On note également que la précision d'un SVM exploitant l'ensemble des 2 500 caractéristiques exceptées les X plus fréquentes est très élevée, tant sur l'ensemble d'apprentissage que sur l'ensemble de test, ce qui montre que ces caractéristiques contiennent également des informations discriminantes. Ainsi non seulement les X premières caractéristiques permettent une discrimination parfaite en apprentissage mais les caractéristiques de X à 2 500 le permettent aussi.

Par ailleurs si l'on observe la performance des classifieurs utilisant X caractéristiques prises au hasard dans l'ensemble des 2 500 caractéristiques, on voit que l'utilisation d'un nombre limité de caractéristiques aléatoires permet également de faire la distinction entre les auteurs jusqu'à un certain point, ce qui tend à montrer également que toutes les caractéristiques (y compris les moins fréquentes) contiennent des informations discriminantes. Potentiellement les courbes de ces figures laissent penser que de nombreuses caractéristiques véhiculent une information discriminante.

Or, il est probable que l'apprentissage d'un SVM mettra l'accent sur l'exploitation des caractéristiques les plus fréquentes de sorte qu'à la fin, on peut s'attendre à ce que les SVMs ne soient pas nécessairement en mesure d'exploiter pleinement toutes les caractéristiques discriminantes. En d'autres termes, il doit exister un certain nombre de caractéristiques discriminantes qui sont négligées par le processus d'apprentissage et qui pourraient améliorer les performances en généralisation si elles étaient réellement exploitées.

4.2. *Bagging de caractéristiques pour l'authentification d'auteur*

Sur la base de l'analyse précédente nous avons cherché à concevoir des approches capables d'exploiter pleinement le potentiel de toutes les caractéristiques disponibles. Nous nous sommes naturellement tournés vers des méthodes d'ensemble avec l'idée de combiner de multiples classifieurs appris sur des ensembles de caractéristiques différents.

De nombreuses méthodes ont été proposées pour combiner des classifieurs, le co-training, le boosting, le bagging, dont un certain nombre ont été conçues ou adaptées pour combiner des classificateurs opérant sur différents sous-ensembles de caractéristiques (Viola *et al.,* 2001), (Sutton *et al.,* 2005). En particulier, l'apprentissage de classifieurs appris sur des ensembles de caractéristiques divers a été étudié par quelques chercheurs dans le passé (O'Sullivan *et al.,* 2000), avec le cas particulier de (Viola *et al.,* 2001) qui ont utilisé des classifieurs faibles appris chacun sur une caractéristique.

Nous avons exploré une approche de Bagging de caractéristiques où l'on apprend un nombre important de classifieurs de base, chacun appris sur un sous-ensemble aléatoires de caractéristiques, que l'on fait voter en inférence pour produire une décision. Plus concrètement, pour obtenir des résultats performants sur nos corpus de classification d'auteurs nous avons utilisé plusieurs centaines à plusieurs milliers de modèles SVMs appris chacun sur un sous-ensemble aléatoire de quelques dizaines à quelques centaines de caractéristiques. Dans nos expériences tous les SVM sont appris avec la boîte à outils libsvm.

5. Résultats expérimentaux

5.1. *Résultats préliminaires*

Nous montrons tout d'abord quelques résultats préliminaires obtenus sur le corpus de *blogs*. Tous les résultats présentés dans cette section, exceptés ceux de la table 2, ont été obtenus pour un découpage unique apprentissage / validation / test pour limiter la complexité des expériences. La base de validation est utilisée pour déterminer la valeur optimale du paramètre de régularisation des classifieurs SVM.

Nous étudions tout d'abord l'influence du nombre de caractéristiques aléatoires X exploitées par chacun des classificateurs de base et du nombre de classifieurs de base, M. La figure 3 montre l'évolution et la précision du système en fonction du nombre de modèles de base M. Chaque courbe correspond à une valeur de X. Les caractéristiques utilisées sont choisies parmi l'ensemble des 3 000 trigrammes de caractères les plus fréquents dans le corpus. Comme on le voit la valeur de X une influence clairement sur la performance de l'approche globale et il semble préférable d'utiliser ici une valeur plutôt petite, qui permet probablement d'induire une plus grande variabilité entre les classificateurs de base. Par ailleurs, il semble que la performance du système final augmente avec le nombre de classifieurs de base, en particulier dans le cas de classifieurs de base opérant sur un petit nombre de caractéristiques.

La table 1 fournit les performances minimales, moyennes et maximales parmi un ensemble de 1 300 classifieurs exploitant 100 caractéristiques choisies aléatoirement, sur les bases d'apprentissage, de validation et de test. On y voit que les performances en apprentissage sont très élevées en moyenne et que le gap entre performance en apprentissage et en test est effectivement important dans tous les cas.

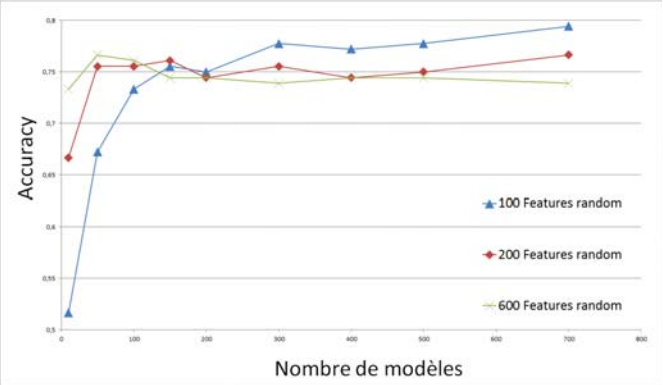

Figure 3. *Performance (précision) de l'approche par Bagging en fonction du nombre M de modèles utilisés (de 1 à 700) pour trois tailles de sous ensembles de caractéristiques ($X = 100, 200, 600$).*

La table 2 montre les résultats de l'approche par bagging de caractéristiques sur le même jeu de données, avec 1 300 classifieurs exploitant chacun 100 caractéristiques aléatoires, mais cette fois les résultats ont été obtenus avec une procédure de validation croisée à 6 folds, ce qui permet d'obtenir des indices de significativité. On voit que l'approche Baggging apporte des gains significatifs par rapport à l'approche de référence, un SVM linéaire opérant sur l'ensemble des caractéristiques. La dernière ligne montre la valeur de la statistique obtenue par un t-test pour valider que la performance de l'approche Bagging est significativement supérieure à celle du SVM. Ces valeurs montrent que l'approche par Bagging est supérieure avec un degré de confiance de l'ordre de 95%.

Tableau 1. *Statistiques sur les $M = 1\,300$ classifieurs exploitant 100 caractéristiques.*

Minimum			Mean			Maximum		
Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
67,3	17,8	16,1	99,5	28,5	26,9	100	42,2	41,7

Enfin, pour explorer la capacité de rejeter efficacement des documents (par exemple pour rejeter un document non écrit par un des auteurs connus du système) nous avons construit une courbe de précision rappel en ordonnant les documents par leur score calculé comme le nombre de votes de la classe reconnue. Nous avons considéré les exemples bien classés comme des exemples positifs et les exemples mal classés comme des exemples négatifs. On obtient des courbes de précision rappel du type

Tableau 2. *Comparaison de la performance de l'approche Bagging et d'un SVM. La dernière ligne est la valeur de la statistique obtenue par un test t-test pairé sur la supériorité de l'approche par Bagging (e.g. une valeur inférieure à 0,05 indique que l'approche Bagging est supérieure avec une certitude de 95%).*

Model	Train	Valid	Test
Bagging (1 300 modèles - 100 features)	100	83,6	82,1
Single SVM with all 3 000 features	100	80,3	79,8
Valeur de la statistique	-	0,0015	0,0535

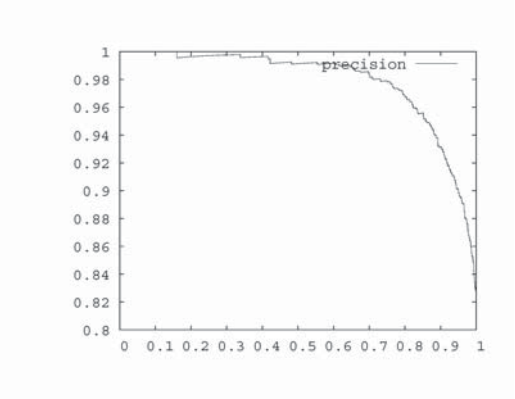

Figure 4. *Courbe de précision en fonction du rappel pour l'approche de Baggging. Les exemples sont rangés suivant le nombre de votes qu'ils ont reçus. Les exemples positifs sont les exemples bien reconnus et les exemples négatifs sont les exemples non reconnus.*

de la figure 4. On voit que le nombre de votes est un critère qui semble assez efficace pour rejeter.

5.1.1. *Résultats sur la compétition PAN*

Pour le défi PAN, il nous était donné un ensemble de données d'entrainement. Nous avons choisi d'apprendre des modèles de classification sur un certain nombre, S (de l'ordre de 10), de partitions du corpus d'entrainement en ensemble d'apprentissage et ensemble de validation. Pour chacune de ces partitions, nous avons appris M modèles exploitant chacun X caractéristiques choisies aléatoirement parmi les caractéristiques initiales, soit des comptages mots soit des comptages de ngrams de caractères. Nous avons ensuite généré des prédictions en faisant voter les $S \times M$ modèles

obtenus. Pour les problèmes fermés, nous avons simplement utilisé un vote majoritaire.

Lors de cette compétition, 3 méthodes ont obtenu les meilleurs résultats. La première est le résultat de la collaboration entre l'université de Bucarest et l'institut Fraunhofer FOKUS (Popescu *et al.,* 2012). Leur méthode utilise des SVM à noyaux et plus particulièrement des "string kernels". La seconde méthode provient de l'université de Bar-Ilan (Akiva, 2012) et utilisent des SVMs sur une représentation qui capture la présence et absence de mots courants. Enfin, l'université de Duquesne (Ryan *et al.,* 2012) a proposé une approche faisant voter 3 techniques sur des représentations en sac de mot : le plus proche voisin en distance L1, un SVM, une technique de vote sur un découpage en micro-documents de 3000 caractères. Sur les tâches fermées de la compétition nous avons fini en 3ème position derrière (Popescu *et al.,* 2012) et (Akiva, 2012).

6. Approche par similitude de profils

Nous avons voulu également exploiter les résultats de (Koppel *et al.,* 2007) qui a suggéré que le style d'un auteur se caractérise par la façon dont se comporte la performance d'un classifieur au fur et à mesure que l'on ignore en test les caractéristiques les plus importantes identifiées en apprentissage (i.e. celles dont les poids sont les plus importants dans un modèle linéaire). Peu importe de quelle caractéristique il s'agit la vitesse à laquelle la performance décroit est révélatrice de l'auteur.

Nous avons voulu exploiter ce résultat dans le cadre de notre approche par Bagging. Le système que nous proposons est un système en deux étapes. Dans un premier temps, nous apprenons comme dans notre approche par Bagging M SVMs linéaires multiclasses exploitant des sous-ensembles aléatoires de X caractéristiques. Nous notons $S_i \subset [1, p]$ l'ensemble des indices des caractéristiques utilisées par le i^{eme} classifieur SVM, que nous notons SVM_i. Tous ces classifieurs sont appris à affecter un document à l'un des N auteurs.

Ensuite, nous utilisons les M SVMs appris pour construire de nouveaux vecteurs que nous appelons des profils. Il y a un profil par couple (auteur, document). Pour tout auteur $a \in [1, N]$ et pour tout document d, on construit un nouveau vecteur à p dimensions (un profil), noté $u(d, a)$, dont la j^{eme} composante est définie suivant :

$$\forall j \in [1, p]\,, u_j(d, a) = \frac{1}{Z(j)} \sum_{i=1:M} \delta(j \in S_i) \times \delta(SVM_i(d) == a) \qquad [1]$$

où $SVM_i(d)$ correspond à la sortie (un numéro de classe dans [1 .. N]) de SVM_i pour le document d, où $\delta(P)$ est égal à 1 si le prédicat P est vrai et à 0 sinon, et où $Z(j)$ est un facteur de normalisation $Z = \sum_{i=1:M} \delta(j \in S_i)$. Ainsi $u_j(d, a)$ représente le pourcentage des classifieurs, parmi ceux qui exploitent la j^{eme} caractéristique, qui

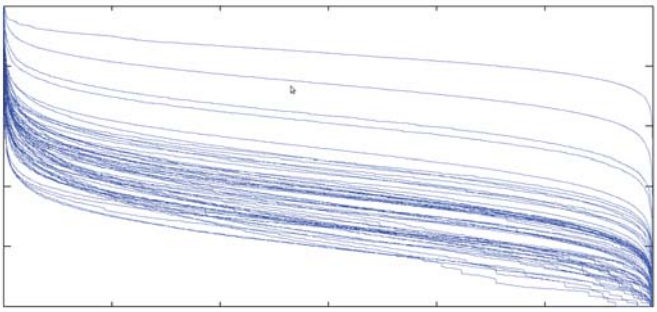

Figure 5. *Profils moyens positifs (triés) sur le corpus de blogs, où l'on voit une courbe par auteur. L'ordonnée est la valeur d'une composante d'un profil et l'abscisse est le numéro de la composante.*

prédisent l'auteur a. A chaque document correspondent N profils, un pour chaque auteur, dont un est un profil dit positif (celui qui correspond à l'auteur réel du document) et les autres sont des profils négatifs (ceux calculés pour les autres auteurs).

Ces profils peuvent être triés (de la valeur la plus élevée à la plus basse) de sorte que la numérotation des composantes est perdue. La figure 5 montre les profils positifs triés (et normalisés pour que la valeur de la première composante soit égale 1) pour les 60 auteurs du corpus *blog*. D'une certaine façon le résultat de (Koppel *et al.,* 2007) conduit à penser que la forme de cette courbe et sa pente est caractéristique d'un profil positif ou négatif.

A ce stade, nous pouvons apprendre un nouveau classifieur pour discriminer entre les profils positifs et les profils négatifs et construire un classifieur en N classes opérant sur les profils. Nous avons ici utilisé une méthode très simple dans laquelle chaque auteur est représenté par le profil positif moyen sur les documents de validation écrits par cet auteur, noté u_a^{val}. Pour classifier un document de test on calcule la sortie des M SVMs puis on construit N profils du document (un par auteur) $\{u(d,a)|a = 1..N\}$ puis on détermine l'auteur pour lequel le profil positif moyen est le plus corrélé au profil correspondant du document :

$$\hat{a} = argmax_a \left[correl(u_a^{val}, u(d,a)) \right] \qquad [2]$$

où $correl(x,y)$ est la corrélation entre les deux vecteurs x et y.

Le tableau 3 montre que cette seconde approche donne des résultats équivalents à l'approche SVM simple qu'elle ne parvient pas à dépasser. Si l'on utilise des profils non triés la performance est même très dégradée. Mais si cette approche ne permet pas de dépasser la méthode de Bagging, elle opère sur une représentation tout à fait

différente et l'on peut chercher à les combiner, ce que nous avons réalisé en calculant une combinaison linéaire des scores :

$$score^{Comb}(u,a) = \alpha \times score_{Profile}(d,a) + (1-\alpha) \times score_{Bagg}(d,a) \quad [3]$$

où $score_{profile}(d,a)$ est la mesure de corrélation précédente et $score_{Bagg}(d,a)$ est un score produit par l'approche par Bagging (un nombre de votes). La figure 6 montre l'évolution de la performance de la méthode de combinaison en fonction du paramètre de mélange α.

Tableau 3. *Comparaison de l'approche par profils et de l'approche combinée avec l'approche SVM simple. A noter également les résultats de l'approche par profils sur des profils non triés. La valeut T stat calculée dans la dernière ligne correspond à la valeut de la statistique d'un t-test pairé sur la supériorité de la méthode combinée (avec un paramètre de mélange optimale) par rapport à la méthode de Bagging, elle montre la significativité à 95% de ce résultat.*

Méthode	Précision en test
SVM unique	79,8
Bagging	82,1
Profils (Non triés)	48,9
Profils (Triés)	78,9
Méthode combinée	83,5 (T stat=0,0415)

7. Conclusion

Nous avons proposé d'utiliser une approche de bagging de caractéristiques pour l'authentification d'auteurs et montré que cette approche dépasse une stratégie très populaire et efficace pour cette tâche. Cette méthode a obtenu des résultats intéressants dans la compétition internationale PAN 2012 et s'est placé à la troisième place parmi onze participants sur les tâches d'identification fermées. Nous avons également proposé une seconde méthode qui si elle s'est avérée moins efficace que la première peut lui être avantageusement combinée.

8. Remerciements

Merci à Moshe Koppel de Bar-Ilan Univeristy (Israel) pour nous avoir fourni ses corpus. Ce travail a été réalisé dans le cadre du projet SAIMSI financé par l'ANR (ANR-09-CSOSG-SAIMSI).

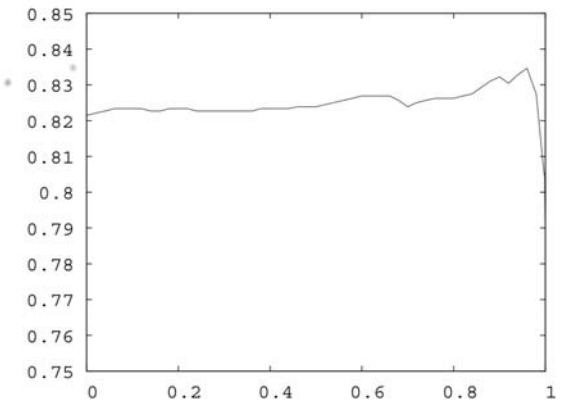

Figure 6. *Performance de la méthode combinant les approches Bagging et Profil en fonction du coefficient de mélange* α. $\alpha = 0$ *est équivalent à la méthode de Bagging et* $\alpha = 1$ *est équivalent à la méthode par Profils.*

9. Bibliographie

Abbasi A., Chen H., « Applying Authorship Analysis to Extremist-Group Web Forum Messages », *IEEE Intelligent Systems*, vol. 20, n° 5, p. 67-75, September, 2005.

Akiva N., « Authorship and Plagiarism Detection Using Binary BOW Features », *In Notebook for Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN) at CLEF*, 2012.

Argamon-Engelson S., Koppel M., Avneri G., « Style-based text categorization : What Newspaper Am I Reading ? », *Proceedings of the AAAI Workshop on Text Categorization*, p. 1-4, 1998.

Chang C.-C., Lin C.-J., « LIBSVM : A library for support vector machines », *ACM Transactions on Intelligent Systems and Technology*, vol. 2, p. 27 :1-27 :27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Hoover D. L., « Frequent Word Sequences and Statistical Stylistics », *Literary and Linguistic Computing*, vol. 17, p. 157-180, 2002.

Kim S., Kim H., Weninger T., Han J., Kim H. D., « Authorship classification : a discriminative syntactic tree mining approach », *SIGIR*, 2011.

Koppel M., Schler J., Argamon S., « Computational methods in authorship attribution », *Journal of the American Society for Information Science and Technology*, 2009.

Koppel M., Schler J., Argamon S., Messeri E., « Authorship attribution with thousands of candidate authors », *Proceedings of the 29th annual international ACM SIGIR conference*

on Research and development in information retrieval, SIGIR '06, ACM, New York, NY, USA, p. 659-660, 2006.

Koppel M., Schler J., Bonchek-Dokow E., « Measuring differentiability : unmasking pseudonymous authors », *Journal of Machine Learning Research*, 2007.

O'Sullivan J., Langford J., Caruana R., Blum A., « FeatureBoost : A Meta Learning Algorithm that Improves Model Robustness », *In Proceedings of the Seventeenth International Conference on Machine Learning*, p. 703-710, 2000.

Popescu M., Grozea C., « Kernel Methods and String Kernels for Authorshiip Analysis », *In Notebook for Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN) at CLEF*, 2012.

Ryan M., Jr J. N., « Mixture of Experts Authorship Attribution », *In Notebook for Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN) at CLEF*, 2012.

Savoy J., « Authorship Attribution Based on Specific Vocabulary », *ACM Trans. Inf. Syst.*, vol. 30, n° 2, p. 12 :1-12 :30, May, 2012.

Stamatatos E., « A survey of modern authorship attribution methods », *Journal of the American Society for Information Science and Technology*, vol. 60, n° 3, p. 538-556, 2009.

Sutton C., Sindelar M., Mccallum A., Feature bagging : Preventing weight undertraining in structured discriminative learning, Technical report, CIIR, 2005.

Teytaud O., Jalam R., « Kernel-Based Text-Categorization », *In International Joint Conference on Neural Networks (IJCNN)*, 2000.

Viola P., Jones M., « Rapid object detection using a boosted cascade of simple features », *Computer Vision and Pattern Recognition, 2001*, 2001.