Attribution d'auteur par ensembles de séparateurs

Jacques Savoy

Institut d'informatique, Université de Neuchâtel
rue Emile Argand 11, 2000 Neuchâtel (Suisse)
Jacques.Savoy@unine.ch

RÉSUMÉ. *L'attribution d'auteur peut être analysée comme une tâche particulière en catégorisation de textes. Dans cette perspective, on définit d'abord une liste d'attributs pertinents (vocables dans cet article). Ensuite, on entraîne un modèle de classification afin de discriminer entre les auteurs potentiels. Pour améliorer la performance moyenne on peut s'appuyer sur un ensemble de séparateurs, la solution retenue étant celle de la majorité (bagging). Afin de générer ce groupe de classifieurs, nous présentons deux formes de variations possibles, d'une part en perturbant les profils d'auteurs et, d'autre part, la liste des attributs. Afin de comparer l'efficacité de ces approches, nous avons extrait deux corpus d'articles de presse (Glasgow Herald) écrits par cinq journalistes, un dans le domaine du sport (1 948 articles) et le second en politique (987 articles). Sur la base de la performance obtenue par la méthode de divergence Kullback-Leibler (Zhao & Zobel, 2007), les stratégies plus complexes n'apportent pas toujours les améliorations escomptées.*

ABSTRACT. *The authorship attribution problem can be viewed as a categorization problem. To discriminate between different writers (or categories), we must first select a list of useful features (word types in this study), and then we train our classifier. To improve effectiveness, we can consider an ensemble of models instead of a single classifier (bagging). In the current study, we propose two forms of variation: varying the author profiles on the one hand, and on the other, varying the list of selected features. To compare the effectiveness of these solutions, we have extracted two corpora from the Glasgow Herald written by five columnists, the first one is on sports (1,948 articles), and the second on politics (987 articles). Using the KLD model (Zhao & Zobel, 2007), we found that a simple classification scheme tends to produce results comparable to those obtained from using more complex ones.*

MOTS-CLÉS : *Méthode d'ensemble, attribution d'auteur, catégorisation de textes.*

KEYWORDS: *Ensemble learning, bagging, authorship attribution, text categorization.*

1. Introduction

L'attribution d'auteur cherche à déterminer l'auteur d'un écrit anonyme ou celui dont l'attribution reste incertaine (Juola, 2006). Durant les siècles précédents, plusieurs techniques ont été proposées (Love, 2002) et les plus anciennes remontent

à Leon Battista Alberti (1404-1472) (Ycart, 2012). En s'appuyant sur des méthodes statistiques et sur les capacités de traitement de l'informatique, des méthodes plus sophistiquées ont été proposées. Récemment, plusieurs approches suggèrent de représenter le style particulier d'un auteur en retenant un nombre limité de mots fonctionnels fréquents. Comme hypothèse sous-jacente, on admet que la fréquence d'apparition de certains mots n'est pas sous le contrôle conscient de l'auteur. De plus, cette fréquence d'emploi varie d'une personne à l'autre. Enfin, les mots fonctionnels possèdent l'avantage d'être relativement indépendants des thèmes abordés.

En suivant cette idée, toutes les solutions proposées font l'hypothèse que les écrits disponibles sont fixes et stables. Or, un regard plus attentif révèle que les auteurs modifient leurs œuvres d'une édition à l'autre. Par exemple, on peut moderniser la formulation et l'orthographe d'une édition ancienne. Dans d'autres cas, la reconnaissance des caractères (OCR) ou le traitement informatique peuvent introduire des erreurs ou des modifications. Ces variations peuvent influencer et faire varier les profils d'auteurs (concaténation de tous les textes connus d'un écrivain). Afin de discriminer entre auteurs, la liste des mots retenus fluctue d'un modèle de classification à l'autre. Par exemple, Zhao & Zobel (2007) proposent une liste de 363 mots tandis que Hughes *et al*. (2012) énumèrent 307 vocables qui ne sont pas un sous-ensemble de la première liste. Ces faits indiquent qu'il est illusoire de concevoir les profils d'auteurs ou les listes d'attributs comme absolument fixes et stables.

Basé sur ces idées, l'objectif de cet article est de présenter comment un système d'attribution d'auteur performant (le modèle KLD basé sur la divergence de Kullback-Leibler de Zhao & Zobel (2007)) peut être modifié pour tenir compte de ces deux sources de variations, soit les profils d'auteurs et la liste des attributs sélectionnés. Une manière naturelle de les représenter consiste à générer un ensemble de séparateurs selon une stratégie de *bagging* (*bootstrap aggregating*).

Dans la suite de cet article, nous présenterons un survol des connaissances en attribution d'auteur (section 2). La troisième section expose les grandes lignes des corpus utilisés dans nos expériences. La quatrième section décrit la méthode de classification KLD. La cinquième section présente la génération de deux ensembles de séparateurs basés sur le modèle KLD. Enfin, la sixième résume l'évaluation des divers modèles en se basant sur nos deux corpus.

2. État des connaissances

Afin de résoudre la question de l'attribution d'auteur, trois grandes familles d'approche ont été proposées (Juola, 2006). En premier lieu, on a cherché à définir des mesures stylométriques supposées invariantes par auteur (Holmes, 1998). Dans cette perspective, on a retenu la longueur moyenne des mots ou des phrases, le nombre moyen de syllabes par mots, voire la taille du vocabulaire V (notée |V|) par

rapport à la longueur du document. Comme variantes, on a proposé le rapport entre le nombre de *hapax legomena* (mots apparaissant une seule fois) (notée V_1) et la taille du vocabulaire (soit $|V_1| / |V|$), ou le rapport entre le nombre de mots apparaissant deux fois (noté $|V_2|$) et la taille du vocabulaire (Sichel, 1975). Ces mesures possèdent l'inconvénient d'être difficiles à interpréter face à des textes de tailles différentes. Certes des solutions correctives ont été proposées (Smith & Kelly, 2002) mais l'instabilité de ces mesures (Baayen, 2008) rend leur usage peu probant. De plus, le thème, l'époque (Juola, 2003) ou le genre (poésie, pièce de théâtre, roman, texte en vers ou en prose) influencent de telles mesures.

Une deuxième famille d'approches se fonde sur le vocabulaire. Dans cette perspective, Mosteller & Wallace (1964) proposent de sélectionner de manière semi-automatique les vocables les plus pertinents (l'étude finale retiendra 35 vocables). Ces travaux mettent en lumière l'importance des mots fréquents et, en particulier, des mots fonctionnels (déterminants, prépositions, conjonctions, pronoms et quelques adverbes et auxiliaires). Par exemple, le premier auteur (Hamilton) utilise 36 fois le mot *while* mais jamais *whilst*, tandis que le second auteur potentiel (Madison) n'écrit jamais *while* mais écrit douze fois *whist*. Toutefois, on ne dispose pas d'une définition précise pour déterminer ce qu'est un mot fonctionnel (ou mot-outil). De plus, Damerau (1975) souligne que certains vocables sont plus adéquats pour discriminer les différences de style entre auteurs.

En poursuivant cette voie, Burrows (2002) propose de sélectionner les mots en se basant sur la fréquence d'occurrence. Ainsi la liste des attributs à retenir comprendra les 50 à 150 vocables les plus fréquents, ensemble comprenant une forte proportion de mots fonctionnels. Ce seuil sera repoussé à 800 (Hoover, 2004) puis à 4 000 (Hoover, 2007) avec l'inclusion de mots lexicaux fréquents (noms, adjectifs, adverbes et verbes).

Les études menées par Zhao & Zobel (2005, 2007) proposent de définir *a priori* les vocables à retenir. Dans ce cas, on retient essentiellement les mots fonctionnels en ignorant les mots lexicaux qui reflètent plus les thèmes traités. Pour la langue anglaise, ces auteurs suggèrent une liste de 363 formes, un ensemble correspondant au contenu d'une liste de mots-outils d'un moteur de recherche. Dans une perspective similaire, Hughes *et al*. (2012) proposent de retenir 307 mots (fonctionnels) afin de décrire les styles de différents auteurs couvrant une période d'environ 350 ans. Cette étude démontre que le style des auteurs du XVIIIe (ou ceux du XIXe siècle) était assez similaire. L'époque possède donc une influence marquée sur le style. Toutefois, cette tendance vers un style lié à une période donnée tend à disparaître au cours du XXe siècle.

Comme troisième famille d'approches, nous pouvons signaler le recours à des techniques d'apprentissage automatique (*machine learning*) que l'on retrouve également dans le cadre de la catégorisation automatique (Sebastiani, 2002), (Stamatatos, 2009). Dans cette perspective, nous devons d'abord sélectionner les termes possédant le meilleur pouvoir discriminant, puis entraîner un séparateur.

Comme modèle de classification, Zheng *et al*. (2006) ont comparé les performances obtenues par des réseaux de neurones, des machines à vecteurs de support (SVM) et des arbres de décision. Cette étude indique que les deux premières approches offrent de meilleurs résultats que les arbres de décision. Zhao & Zobel (2005) indiquent qu'une approche fondée sur les plus proches voisins (*k*-NN) permet d'obtenir une meilleure performance que des séparateurs basés sur le modèle naïve Bayes ou des arbres de décision. Les modèles à thèmes (*topic models*) ont également été proposés (Savoy, 2013) avec des performances similaires à un modèle naïve Bayes. Pour d'autres auteurs, ces modèles à thèmes peuvent fournir des attributs complémentaires à ceux fournis par le vocabulaire (Pearl & Steyvers, 2012).

3. Corpus d'évaluation

Comme en recherche d'information, l'évaluation empirique tient une place importante en catégorisation de textes. Grâce à des corpus de tests, nous pouvons évaluer et comparer diverses représentations et modèles de classification. Cependant, les études en attribution d'auteur disposent d'un nombre relativement restreint de corpus. De plus, les études tendent souvent à se focaliser sur une seule œuvre ou sur un nombre restreint de documents. Le nombre d'auteurs possibles demeure aussi limité car il s'avère difficile de trouver un nombre important de candidats potentiels respectant des contraintes multiples (même période et langue, cultures proches, thèmes similaires, et volume d'apprentissage important).

Désirant fonder nos conclusions sur une base assez large et au moyen d'une collection stable et facilement accessible, nous avons sélectionné un sous-ensemble de la collection CLEF - 2003 (Peters *et al*., 2004). Cette partie comprend les articles publiés durant l'année 1995 dans le journal *Glasgow Herald*. Si le corpus complet compte 56 472 documents, nous ne connaissons le ou les auteur(s) que pour 28 687 d'entre eux. De ce dernier sous-ensemble, nous avons sélectionné les articles rédigés par un seul auteur et écarté les journalistes ayant écrit peu d'articles durant l'année 1995. Finalement, nous avons regroupé les articles ayant une thématique similaire afin de rendre la tâche de catégorisation plus ardue. Dans le cadre de cette étude, nous avons retenu les rubriques sportive et politique.

Dans les tableaux 1a et 1b, nous avons indiqué le nom des journalistes, le thème principal, puis le nombre d'articles rédigés, ainsi que la longueur moyenne (en nombre de mots) par auteur. Si le nombre de documents et le nombre d'articles par auteur sont plus importants dans la collection « Sports », le corpus « Politique » présente une plus grande variabilité dans la répartition du nombre d'articles par journaliste. De même, la longueur moyenne varie peu chez les chroniqueurs sportifs en comparaison des auteurs de la catégorie politique.

	Nom	Thème	Nombre d'articles	Longueur moyenne
1	Douglas Derek	sports	411	808
2	Gallacher Ken	sports	409	727
3	Gillon Doug	sports	369	713
4	Paul Ian	sports	419	842
5	Traynor James	sports	340	983
	Total		1 948	

Tableau 1a. Répartition des 1 948 articles dans le domaine « Sports » par journaliste (*Glasgow Herald*)

	Nom	Thème	Nombre d'articles	Longueur moyenne
1	Johnstone Anne	politique	73	1 258
2	Shields Tom	politique	174	1 001
3	Smith Graeme	politique	330	520
4	Trotter Stuart	politique	337	666
5	Wishart Ruth	politique	73	1 137
	Total		987	

Tableau 1b. Répartition des 987 articles dans le corpus « Politique » par auteur (*Glasgow Herald*)

Si l'on classe par fréquence d'apparition les vocables du corpus des articles de sports, on constate que la forme la plus fréquente est la virgule, suivi du déterminant *the*, du point, puis des mots *to*, *and*, *a* et *of*. Sur le corpus de politique, le classement est similaire avec le déterminant *the*, puis la virgule, le point, et les vocables *of*, *to*, *a* et *and*.

4. Méthode d'attribution KLD

Comme méthode d'attribution d'un article à son auteur, nous avons retenu l'approche proposée par Zhao & Zobel (2005, 2007). Ces derniers suggèrent de mesurer la distance entre le profil d'un auteur A_j (concaténation de tous ses écrits) et un texte requête (noté Q) en utilisant la divergence Kullback-Leibler (KLD) (nommée aussi entropie relative (Maning & Schütze, 1999)). Cette mesure est exprimée dans l'équation 1 dans laquelle $Prob_q[t_k]$ et $Prob_{aj}(t_k)$ indiquent la probabilité d'occurrence du vocable ou terme t_k dans la requête ou dans le profil du j^e auteur (A_j). Lors du calcul, nous imposons que $0 \cdot \log_2[0/p] = 0$, et $p \cdot \log_2[p/0] = \infty$.

$$KLD(Q \,\|\, A_j) = \sum_{k=1}^{m} \mathrm{Prob}_q[t_k] \cdot \log_2 \left[\frac{\mathrm{Prob}_q[t_k]}{\mathrm{Prob}_{aj}[t_k]} \right] \quad (1)$$

Lorsque deux distributions sont identiques, la valeur KLD est nulle. Dans tous les autres cas, la valeur retournée est positive et d'autant plus importante que la distance entre les distributions dérivées du document Q et du profil A_j est élevée.

Pour estimer les probabilités sous-jacentes, nous pouvons suivre le principe du maximum de vraisemblance en estimant que Prob[t_k] = tfa_k/n, avec tfa_k indiquant la fréquence d'occurrence du terme et n la taille du profil ou du document concerné. Cette estimation peut être lissée afin d'éliminer la présence de probabilités nulles (Manning & Schütze, 1999). Dans nos évaluations, nous avons adopté l'approche de Lidstone en estimant les probabilités par (tfa_k+λ) / (n+λ·|V|), avec |V| indiquant la taille du vocabulaire retenu. Nous avons fixé la valeur du paramètre λ à 0,01 car cette dernière retourne la meilleure performance.

5. Méthode d'ensemble

Afin d'améliorer la performance d'un séparateur, nous pouvons opter pour une sélection plus adéquate des attributs, voire augmenter leur nombre. Un tel accroissement n'est pas forcément corrélé avec une augmentation de la performance moyenne, en attribution d'auteur pour le moins (Savoy, 2012). Comme alternative, on peut changer de modèle de classification. Selon des études récentes, la tendance vers des modèles plus complexes ne s'accompagnerait pas toujours d'une augmentation sensible de la performance (Hand, 2006).

Afin d'améliorer la performance, nous pouvons nous fonder sur l'idée que deux avis valent mieux qu'un. En suivant cette idée, nous proposons d'examiner le résultat obtenu par plusieurs séparateurs et de retourner la catégorie obtenue par la majorité (*model ensembles*). Avec ce principe nous obtenons une décision plus robuste par lissage des fluctuations liées à des variations individuelles.

Pour diversifier les décisions, nous pourrions générer des séparateurs basés sur des modèles de classification distincts (par exemple, naïve Bayes, SVM, arbre de décision, etc.). Par contre, maintenir un tel ensemble disparate s'avère complexe tant au niveau logiciel que dans la gestion des phases d'apprentissage et de validation.

Comme autre source de variation, nous pouvons fonder nos décisions sur des profils d'auteur distincts. Une telle variation peut s'opérer en faisant varier la liste des œuvres retenues pour un auteur donné. Ainsi, sur l'inventaire exhaustif des textes écrits par un journaliste donné, nous pouvons générer une autre liste par ré-échantillonnage aléatoire (*bootstrap*). Par exemple, si nous disposons de l'ensemble des documents {D_1, D_2, D_3, D_4, D_5}, un profil peut être obtenu par l'échantillon {D_3, D_2, D_3, D_1, D_1} et un autre profil par {D_2, D_5, D_2, D_3, D_5}.

Cette idée de variation se fonde également sur l'instabilité possible des écrits eux-mêmes. En effet, un document peut disposer de plusieurs versions contenant diverses modifications. A titre d'exemple, on peut mentionner les *Federalist Papers* (corpus de l'étude de Mosteller & Wallace (1964)) pour lesquels différentes versions coexistent (Rudman, 2012). Ces variations peuvent correspondre à des modifications mineures (changement sur un mot), mais parfois des paragraphes entiers diffèrent d'une édition à l'autre.

La variabilité peut également porter sur la liste des termes sélectionnés pour représenter les documents. Dans ce cas, Zhao & Zobel (2007) proposent une liste de 363 mots tandis que Hughes *et al*. (2012) suggèrent une énumération quelque peu différente comprenant 307 vocables fonctionnels. Dans les années 70, Damerau (1975) avait proposé un ensemble de 287 mots, tandis que Mosteller & Wallace (1964) se limitaient à retenir 35 formes. Comme ces listes ne sont pas des sous-ensembles les unes des autres, leur union permet de générer un ensemble de 584 vocables. Depuis cet ensemble, nous pouvons également générer différents échantillons avec remplacement (*bootstrap*) et ainsi former différents séparateurs.

En résumé, nous proposons de recourir au même modèle (KLD) sur des ensembles d'apprentissage distincts (*bagging*). Ces derniers sont obtenus par variation aléatoire des profils d'auteur ou de la liste des attributs (*features*). Lors du ré-échantillonnage aléatoire, certains éléments (articles ou vocables) apparaîtront plusieurs fois tandis que d'autres seront absents. En effet, si nous disposons de n éléments, la probabilité de ne pas tirer un objet précis est de $(1-1/n)$. En répétant ce tirage n fois, la probabilité que cet élément ne soit pas présent dans l'échantillon final s'élève à $(1-1/n)^n$. Lorsque n tend vers l'infini, cette probabilité tend vers $1/e = 0{,}368$. Par exemple, sur une liste de 500 termes, une version ré-échantillonnée disposera, en moyenne, de $500 \cdot 0{,}632 = 316$ termes distincts. Créés depuis des listes d'attributs distincts, les classsifieurs peuvent fournir des réponses différentes.

Comme chacun de nos classsifieurs donne une prédiction (ou un auteur probable dans notre étude), le résultat final sera l'auteur ayant été prédit par le plus grand nombre des k classifieurs (classe majoritaire). En cas d'égalité, on choisi de manière aléatoire. Cette manière de procéder possède l'avantage de la simplicité. Comme variante, nous pourrions considérer toutes les catégories en fonction de la distance avec le texte requête. Pour un ensemble de k séparateurs, on somme les k distances de la divergence de Kullback-Leibler et on attribue le texte à l'auteur ayant la plus faible de ces sommes.

Cette technique d'apprentissage par ensemble permet de simplifier la gestion logicielle (traitement parallèle) et devrait fournir une amélioration de la performance moyenne, en particulier lorsque le modèle de classification est sensible à des variations du jeu d'apprentissage. " Bagging … is a simple but highly effective ensemble method " (Flach, 2012, p. 331) ou " Bagging generates a diverse ensemble of classifiers by introducing randomness into the learning algorithm's input, often with excellent results. " (Witten *et al*., 2011, p. 356)

Dans les études antérieures à cette démarche, nous retrouvons très souvent les arbres de décision comme modèle de classification (Breiman, 1996), (Dietterich, 2006). Dans notre cas, le modèle choisi est différent. De plus, le ré-échantillonnage aléatoire s'opérait fréquemment sur les instances et non sur les attributs (*features*) (une exception à cette règle est l'étude (Ho, 1998)) ou sur les profils d'auteur, composante propre à la question de l'attribution d'auteur.

6. Evaluation

Afin d'évaluer la performance d'un modèle d'attribution d'auteur, nous devons former un ensemble de textes pour l'apprentissage et un autre distinct pour mesurer la performance. Pour respecter cette contrainte, nous pouvons adopter la validation croisée comme stratégie d'évaluation (Hastie *et al*. 2009). Dans le cas présent, nous avons opté pour l'approche *leaving-one-out*. Dans ce cas, toutes les instances, sauf une, seront utilisées pour l'entraînement et la dernière pour le test. Par itération successive, chaque article à tour de rôle va servir à mesurer la performance, le solde formant l'ensemble d'entraînement.

6.1 Détermination d'une performance de référence

Dans une première série d'évaluation, nous nous sommes intéressés à la performance moyenne obtenue par les différentes listes de vocables définis comme attributs. Dans la première colonne du tableau 2, nous avons indiqué le nom des différentes listes ainsi que leur taille. Par exemple, Zhao & Zobel (2007) proposent 363 vocables, Hugues *et al*. (2012) une énumération de 307 mots tandis que Damerau (1975) suggère 287 mots. En examinant nos deux corpus, nous retrouvons un nombre plus faible de vocables présents. Par exemple, le corpus « Sports » possède seulement 302 termes apparaissant dans la liste de Zhao. Ces différences s'expliquent par quelques entrées peu fréquentes (*howbeit*, *whereafter, whereupon*), ou à celles correspondant au comportement attendu lors de la segmentation (*doesn, weren*) ou à un choix plus arbitraire (*indicate*, *missing*, *specifying*, *seemed*). Enfin, en dernière ligne, nous avons formé une liste contenant l'union des quatre autres listes auxquelles on a ajouté neuf symboles de ponctuation (soit . , : ; - ! ? (').

	Sports		Politique	
Liste	Taille	Performance	Taille	Performance
Mosteller (35)	27 mots	50,5 % †	26 mots	50,4 % †
Damerau (287)	266 mots	78,5 % †	265 mots	82,7 % †
Hugues *et al*. (307)	265 mots	79,0 % †	262 mots	79,1 % †
Zhao & Zobel (363)	302 mots	**83,5** %	294 mots	81,0 % †
Union (584)	491 mots	83,1 %	483 mots	**86,5** %

Tableau 2. *Évaluation des diverses listes de termes selon nos deux corpus et avec l'approche basée sur la divergence de Kullback-Leibler (Zhao & Zobel, 2007)*

En appliquant notre stratégie d'évaluation et sur la base des 35 termes sélectionnés par Mosteller & Wallace (1964), le taux de bonnes prédictions (*micro-average*) de l'approche KLD correspond à 50,5 % avec la collection « Sports » ou à 50,4 % avec le corpus « Politique ». En augmentant le nombre de termes, on constate un accroissement du taux de bonnes prédictions (ou taux de réussite). Comme l'indique les valeurs du tableau 2, l'union de toutes ces listes avec l'inclusion de symboles de ponctuation tend à apporter la meilleure performance. Dans la suite de nos propos, nous allons retenir comme mesure de performance

uniquement ce taux de bonnes prédictions. Afin de savoir si une différence de performance entre deux approches s'avère statistiquement significative, nous avons opté pour le test du signe (Conover, 1971), (Yang & Liu, 1999) (test bilatéral) avec un seuil de signification $\alpha = 1$ %. En appliquant ce test, l'hypothèse H_0 admet que les deux modèles possèdent des niveaux de performance similaire. Dans le tableau 2, nous avons retenu la dernière ligne comme modèle de référence et les différences de performance statistiquement significatives sont indiquées par une croix '†'. Comme on le constate, les performances obtenues avec les listes de Mosteller & Wallace, Damerau ou Hughes sont toujours significativement différentes de la dernière ligne.

6.2 Evaluation d'un ensemble de séparateurs

Dans une deuxième série d'expériences, nous avons comparé le modèle correspondant à l'union de toutes les listes avec des stratégies basées sur un ensemble de séparateurs (*bagging*). Pour former cet ensemble composé de *k* classifieurs, nous faisons varier le nombre de vocables retenus (sur un maximum de 584) par ré-échantillonnage aléatoire. Dans la colonne « Attributs uniq. », nous éliminons les duplicatas de la liste des attributs générés par ré-échantillonnage. Dans la colonne « Attributs », la répétition d'un attribut accroît son importance dans le calcul de la similarité KLD avec les profils d'auteur (voir équation 1). Ainsi, l'impact d'un terme sera doublé si on le répète une fois, ou triplé en présence de trois occurrences dans la liste des attributs retenus.

Liste	Attributs uniq.	Attributs	Profil
Union (584)	**83,1** %	83,1 %	83,1 %
$k = 5$	82,2 %	80,5 % †	82,6 %
$k = 10$	82,3 %	81,4 % †	83,3 %
$k = 20$	82,7 %	**83,8** %	**83,4** %
$k = 25$	82,6 %	83,7 %	83,2 %
$k = 50$	82,9 %	83,7 %	83,2 %
$k = 100$	82,7 %	83,3 %	83,1 %
$k = 150$	82,8 %	83,2 %	83,1 %
$k = 200$	**83,1** %	83,4 %	83,2 %
$k = 500$	82,9 %	83,3 %	83,1 %

Tableau 3. *Évaluation des variations du nombre de séparateurs sur le corpus « Sports » et avec l'approche KLD (Zhao & Zobel, 2007)*

Comme seconde source de variation, nous pouvons générer des profils d'auteurs différents par tirage aléatoire, avec remplacement, dans la liste de tous les articles écrits par chaque journaliste (colonne « Profil »). Dans ce cas, les *k* décisions seront obtenus par *k* profils d'auteur différents, chacun étant formé par un ré-échantillonnage aléatoire sur la base des articles écrits par chaque journaliste. Enfin

le tableau 3 regroupe les évaluations tirées du corpus de sports tandis que le tableau 4 présente les performances obtenues avec le corpus « Politique ».

Liste	Attributs uniq.	Attributs	Profil
Union (584)	86,5 %	86,5 %	86,5 %
$k = 5$	85,9 %	83,2 % †	86,2 %
$k = 10$	86,4 %	82,9 % †	86,4 %
$k = 20$	**87,0** %	85,6 %	86,8 %
$k = 25$	86,8 %	86,9 %	86,5 %
$k = 50$	86,6 %	**87,0** %	86,8 %
$k = 100$	86,6 %	86,3 %	**87,3** % †
$k = 150$	86,4 %	86,0 %	86,8 %
$k = 200$	86,6 %	86,5 %	86,7 %
$k = 500$	86,7 %	86,5 %	87,1 %

Tableau 4. *Évaluation des variations du nombre de classifieurs sur le corpus « Politique » et avec l'approche KLD (Zhao & Zobel, 2007)*

En comparant les résultats des diverses méthodes basés sur des ensembles, on ne distingue aucune augmentation importante de la performance moyenne. A la limite, le corpus d'articles en politique laisse percevoir un léger accroissement du taux de réussite. Lorsque l'on compare ces performances avec l'approche simple KLD (première ligne dans les tableaux 3 et 4), les différences de performances moyennes restent faibles et elles s'avèrent très souvent statistiquement non-significatives (absence du symbole '†'). Finalement entre les trois méthodes de génération d'ensemble de séparateurs, les différences demeurent marginales. Modifier les attributs ou les profils d'auteur permettent d'obtenir des taux de réussite similaires.

6.3 Analyses des résultats

Nous pouvons aussi changer un peu notre perspective d'analyse. Lorsque nous disposons de $k = 500$ séparateurs obtenus par ré-échantillonnage de la liste des attributs (colonne « Attributs » du tableau 3), le taux de réussite obtenu selon la règle de la majorité s'élève à 83,3 %. Si on analyse les performances individuelles de ces 500 classifieurs, on constate que la valeur minimale s'élève à 23,2 % et le maximum à 70,9 % (moyenne : 37,9 %, écart-type : 7,8). La règle de la majorité permet donc d'obtenir une performance supérieure à celle obtenue par le meilleur séparateur pris individuellement.

Si l'on considère les variations possibles avec le profil des auteurs (corpus « Sports », dernière ligne et colonne du tableau 3), les fluctuations de performance sont nettement plus faibles. Avec $k = 500$ séparateurs, la performance individuelle minimale s'élève à 81 % et le maximum à 83,4 % (moyenne : 82,2 %, écart-type : 0,4). La règle de la majorité permet d'obtenir un taux de réussite de 83,1 %, soit une valeur légèrement en-dessous du maximum individuel.

Comme seconde analyse, nous souhaitons identifier les articles plus difficiles à classifier correctement. Si l'on reprend le meilleur résultat du corpus politique (tableau 4), celui-ci est obtenu par 100 séparateurs générés par variation des profils d'auteurs. La règle de la majorité permet de déterminer correctement 862 auteurs sur un total de 987 (862 / 987 = 0,87335). Pour 125 documents, le système a fourni une réponse erronée. Pour ces deux ensembles d'articles, nous avons calculé la longueur moyenne en tenant compte uniquement des vocables de la liste « Union » (soit 483 mots présents dans le corpus politique). Pour les articles correctement classifiés, cette longueur moyenne s'élève à 457,3 (écart-type : 236,4) contre une moyenne de 349,5 (écart-type : 212,8) pour les documents mal-classifiés. Si l'on admet sous H_0 que les deux moyennes sont identiques, le test de Student (Conover, 1971) (bilatéral, seuil $\alpha = 1\%$) doit être rejeté. Les deux moyennes sont statistiquement différentes. Déterminer un auteur sur un texte court s'avère plus difficile que sur la base d'un document plus long. Les mêmes conclusions peuvent être tirées sur le corpus d'articles de sport.

7. Conclusion

Dans cette étude, nous avons présenté l'attribution d'auteur comme une tâche particulière en catégorisation de textes. Dans ce cadre, un regard plus attentif sur les documents révèle que l'hypothèse d'un écrit fixe et stable n'est pas toujours vérifiée. Ainsi, on peut disposer de divers profils du même auteur en considérant des inventaires différents de ses écrits. De même, une œuvre particulière peut varier d'une édition à l'autre, impliquant une non-stabilité du profil de l'auteur correspondant. Ces variations impliquent des performances différentes pour des séparateurs entraînés sur des données distinctes.

Dans la même perspective, la liste des attributs sélectionnés pour représenter les documents et calculer la similarité entre documents et profils d'auteurs peut différer d'un modèle de classification à l'autre. A titre d'exemple, on peut citer la liste de 363 mots proposée par Zhao & Zobel (2007), tandis que Hughes *et al*. (2012) suggère une liste quelque peu distincte de 307 entrées. Evidemment, des listes d'attributs distinctes impliquent des taux de réussite différents, sans qu'une simple augmentation du nombre d'attributs apporte toujours un accroissement de la qualité du système (Savoy, 2012).

Ces sources de variation permettent de générer des séparateurs possédant des taux de réussite différents. En suivant l'adage qui prétend « que deux têtes valent mieux qu'une », nous pouvons prendre comme décision finale la catégorie sélectionnée par la majorité des séparateurs (*bagging*). Afin de pouvoir évaluer et comparer ces stratégies de classification, nous avons repris la divergence Kleiber-Leibner proposée par Zhao & Zobel (2007), méthode proposant une bonne performance de référence.

Sur la base d'un journal (*Glasgow Herald*), nous avons extrait deux corpus comprenant 1 948 articles dans le domaine du sport et 987 traitant de la politique. L'évaluation faite sur deux corpus indique que les trois stratégies d'ensemble n'apportent pas un accroissement significatif de la performance moyenne. Toutefois, la règle de la majorité permet d'offrir un taux de réussite élevé même en présence de séparateurs peu performants. Enfin, les textes les plus difficiles à catégoriser correspondent à des textes plus courts que la moyenne.

Remerciements

L'auteur remercie les trois arbitres pour leurs commentaires constructifs ayant permis d'améliorer le contenu de cette communication.

8. Bibliographie

Baayen R.H. *Analyzing Linguistic Data. A Practical Introduction to Statistics using R*. Cambridge, Cambridge University Press, Cambridge, 2008.

Breiman L. « Bagging predictors », *Machine Learning*, vol. 24, n° 2, 1996, p. 124-140.

Burrows J.F. « Delta: A measure of stylistic difference and a guide to likely authorship », *Literary and Linguistic Computing*, vol. 17, n° 3, 2002, p. 267-287.

Conover W.J. *Practical Nonparametric Statistics*, 2nd Ed., New York, John Wiley & Sons, 1971.

Damerau F.J. « The use of function word frequencies as indicators of style », *Computers and the Humanities*, vol. 9, 1975, p. 271-280.

Dietterich T.G. « An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization », *Machine Learning*, vol. 40, n° 2, 2000, p. 139-157.

Flach P. *Machine Learning. The Art and Science of Algorithms that Make Sense of Data*. Cambridge, Cambridge University Press, 2012.

Hand D.J. « Classifier technology and the illusion of progress », *Statistical Science*, vol. 21, n° 1, 2006, p. 1-14.

Hastie T., Tibshirani R. & Friedman J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, 2nd Ed., New York, Springer, 2009.

Ho T. « The random subspace method for constructing decision forests », *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 20, n° 8, 1998, p. 832-844.

Holmes D.I. « The evolution of stylometry in humanities scholarship », *Literary and Linguistic Computing*, vol. 13, n° 3, 1998, p. 111-117.

Hoover D.L. « Testing Burrows's delta », *Literary and Linguistic Computing,* vol. 19, n° 4, 2004, p. 453-475.

Hoover D.L. « Corpus stylistics, stylometry, and the styles of *Henry James* », *Style*, vol. 41, n° 2, 2007, p. 160-189.

Hughes J.M., Foti N.J., Kralauer D.C. & Rockmore D.N. « Quantitative patterns of stylistic influence in the evolution of literature », *Proceedings PNAS*, 2012, p. 7682-7686.

Juola P. « The time course of language change », *Computers and the Humanities*, vol. 37, n° 1, 2003, p. 77-96.

Juola P. « Authorship attribution », *Foundations and Trends in Information Retrieval*, vol. 1, n° 3, 2006.

Love H. *Attributing Authorship: An Introduction*, Cambridge University Press, Cambridge, 2002.

Manning C.D. & Schütze H. *Foundations of Statistical Natural Language Processing*, Cambridge, The MIT Press, 1999.

Marsland S. *Machine Learning. An Algorithmic Perspective*. Boca Raton, CRC Press, 2009.

Mosteller F. & Wallace D.L. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*, Reading (MA), Addison-Wesley, 1964.

Pearl L. & Steyvers M. « Detecting authorship deception: A supervided machine learning approach using author writeprints », *Literary and Linguistic Computing*, vol. 27, n° 2, 2012, p. 183-196.

Peters C., Braschler M., Gonzalo J. & Kluck M. (Eds). *Comparative Evaluation of Multilingual Information Access Systems*. Berlin, Springer-Verlag, LNCS #3237, 2004.

Rudman J. « The twelve disputed 'Federalist' papers: A case for collaboration », *Proceedings Digital Humanities 2012*, p. 353-356.

Savoy J. « Authorship attribution based on a probabilistic topic model », *Information Processing & Management*, vol. 49, n° 1, 2013, p. 341-354.

Savoy J. « Etude comparative de stratégies de sélection de prédicteurs pour l'attribution d'auteur » *Actes 9ième COnférence en Recherche d'Information et Applications* CORIA'2012, Bordeaux, mars 2012, p. 215-228.

Sebastiani F. « Machine learning in automatic text categorization », *ACM Computing Survey*, vol. 14, n° 1, 2002, p. 1-27.

Sichel H.S. « On a distribution law for word frequencies », *Journal of the American Statistical Association*, vol. 70, n° 351, 1975, p. 542-547.

Smith J.A. & Kelly C. « Stylistic constancy and change across literary corpora: Using measures of lexical richness to date works », *Computers and Humanities*, vol. 36, 2002, p. 411-430.

Stamatatos E. « A survey of modern authorship attribution methods », *Journal American Society for Information Science and Technology*, vol. 60, n° 3, 2009, p. 433-214.

Witten I.H., Frank E. & Hall, M.A. *Data Mining. Practical Machine Learning Tools and Techniques,* Amsterdam, Morgan Kaufmann, 3rd Ed., 2011.

Yang, Y., & Liu, JX. (1999). A Re-examination of text categorization methods. *Proceedings of the ACM-SIGIR'1999*, 42-49.

Ycart B. « Alberti's letter counts », 2012, http://hal.archives-ouvertes.fr/hal-00745627 (dernière visite, le 9 novembre 2012).

Zhao Y. & Zobel J. « Effective and scalable authorship attribution using function words », *Proceedings of AIRS*, 2005, Berlin, Springer-Verlag, p. 174-189.

Zhao Y. & Zobel J. « Searching with style: Authorship attribution in classic literature », *Proceedings ACSC2007*, 2007, Ballarat, p. 59-68.

Zheng R., Li J., Chen H. & Huang Z. « A framework for authorship identification of online messages: Writing-style features and classification techniques », *Journal of the American Society for Information Science & Technology*, vol. 57, n° 3, 2006, p. 378-393.