Une méthode non supervisée pour la vérification d'auteur à base d'un modèle gaussien multivarié

Mohamed Amine Boukhaled

LIP6 (Laboratoire d'Informatique de Paris 6), Université Pierre et Marie Curie and CNRS, ACASA Team. / Labex OBVIL.
4, place Jussieu, 75252-PARIS Cedex 05 (France)
mohamed.boukhaled@lip6.fr

RÉSUMÉ. *Dans cet article, nous présentons une première étude sur l'utilisation d'une méthode de détection des cas aberrants à base de distance pour la tâche de vérification de l'auteur. Nous avons considéré une méthode non supervisée basée sur un modèle gaussien multivarié. Pour évaluer l'efficacité de la méthode proposée, nous avons mené une expérimentation sur un corpus de textes littéraires français classiques. Nos résultats préliminaires montrent que la méthode proposée peut réaliser une haute performance de vérification qui peut atteindre un score de F_1 de 83%, supérieure à celle de la méthode de référence.*

ABSTRACT. *In this paper, we present a first study on using a distance-based outlier detection method for the authorship verification task. We have considered an unsupervised method based on a multivariate Gaussian model. To evaluate the effectiveness of the proposed method, we conducted experiments on a classic French corpus. Our preliminary results show that the proposed method can achieve a high verification performance that can reach an F_1 score of 83% outperforming the baseline.*

MOTS-CLÉS : *vérification non supervisée de l'auteur, détection des cas aberrants, modèle Gaussien multivarié.*

KEYWORDS: *unsupervised authorship verification, outlier detection, multivariate Gaussian model.*

Ce travail a été réalisé dans le cadre d'une thèse dirigée par Jean-Gabriel Ganascia, professeur à l'Université Pierre et Marie Curie UPMC, directeur de l'équipe ACASA du LIP.

1. Introduction

La vérification de l'auteur (vérification de paternité du texte) est la tâche qui consiste à déterminer si un texte donné est écrit par un auteur candidat ou non. La vérification de l'auteur est un cas particulier du problème de l'attribution de l'auteur, qui peut être adressée à son tour comme une tâche de discrimination multi-classes ou comme une tâche de catégorisation de texte (Sebastiani 2002). Cependant, dans le cas de vérification de l'auteur, on nous donne des échantillons de textes écrits par un et un seul auteur et on nous demande d'évaluer si un autre texte donné, non auparavant vu, est écrit par cet auteur ou non (Koppel et al. 2009). Etant un problème de catégorisation, cette modification de la formulation initiale du problème d'attribution d'auteur rend la tâche de vérification de paternité beaucoup plus difficile. Cela est principalement dû au fait que la construction d'un modèle caractérisant un seul auteur est beaucoup plus difficile que la construction d'un modèle distinguant entre deux auteurs différents (Koppel & Schler 2004).
La vérification de l'auteur comporte deux étapes principales consécutives, une étape d'indexation et une étape de validation de paternité :
1. La première étape d'indexation est fondée sur les marqueurs de style. Elle est effectuée d'abord sur le texte en utilisant des techniques de traitement du langage naturel comme l'étiquetage syntaxique ou l'analyse morphologique.
2. La seconde étape de validation de paternité est appliquée en utilisant les marqueurs indexés.
De nombreux marqueurs de style ont été utilisées pour caractériser les styles d'écriture : la longueur des phrases et la richesse du vocabulaire (Yule 1944), les mots-outils (Holmes et al. 2001; Zhao & Zobel 2005), les signes de ponctuation (Baayen et al. 2002), les étiquettes syntaxiques (Kukushkina et al. 2001), ou les marqueurs basés sur les caractères (Kešelj et al. 2003). Il y a un accord entre les chercheurs dans le domaine que les mots-outils sont les marqueurs le plus fiable du style d'un auteur (Stamatatos 2009).

L'étape de vérification peut être traitée comme un problème de classification binaire (« écrit par l'auteur » comme label positif versus « pas été écrit par l'auteur » comme label négatif). Cependant, cette formulation du problème présente un inconvénient majeur. En fait, dans le cas de la classification binaire, il faut collecter une quantité raisonnable de textes représentatifs de l'ensemble de la classe "pas écrite par l'auteur", ce qui est difficile, voire impossible. D'un autre point de vue, l'étape de vérification peut être également adressée comme un problème de classification à une classe. Dans ce cas, on n'a plus besoin d'avoir des exemples négatifs.

Dans cet article, nous abordons le problème de la vérification de l'auteur comme un problème de détection des cas aberrants où les textes écrits par l'auteur candidat sont considérées comme des cas normaux alors que les textes qui ne sont pas écrits par cet auteur sont considérés comme des cas aberrants ou des anomalies. Nous proposons d'utiliser une méthode non supervisée de détection des cas aberrants à base de distance pour traiter cette question.

Nous commençons d'abord l'article par un bref aperçu sur le problème de détection des cas aberrants dans la section 2, puis nous décrivons notre méthode dans la section 3. Une validation expérimentale de la méthode proposée est présentée dans la section 4. Une conclusion dans la section 5 clôture l'article.

2. Détection des cas aberrants

La détection des cas aberrants est une tâche difficile qui consiste à analyser les tendances dans un ensemble de données afin d'identifier les instances qui ne sont pas conformes au comportement attendu (normal). Ces instances de données sont appelées des anomalies ou des cas aberrants (Chandola et al. 2009). La détection des cas aberrants a été utilisée avec succès dans de nombreuses applications telles que la détection de fraudes, la détection de cibles radar et la reconnaissance de chiffres et de lettres manuscrites (Markou & Singh 2003) .

Cette technique a été également utilisée pour traiter des données textuelles à des fins diverses telles que la détection de nouveaux sujets ou des événements dans une collection d'articles de presse (Chandola et al. 2009). La détection des cas aberrants est basée sur l'idée que l'on ne peut jamais entrainer un algorithme de classification sur toutes les classes possibles que le système est susceptible de rencontrer dans une application réelle. Elle est aussi très adaptée aux situations où le déséquilibre entre les différentes classes peut affecter la précision de la classification (Wressnegger et al. 2013). La figure 1 illustre le différence entre l'apprentissage pour la classification et l'apprentissage pour la détection des cas aberrants.

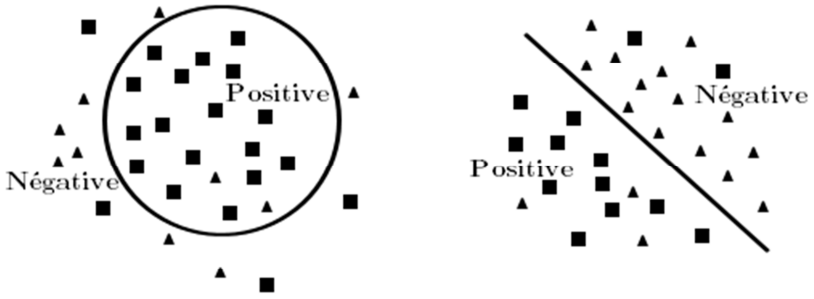

Figure 1 : La détection d'anomalie vs l'apprentissage de classification

La plupart des techniques de détection des cas aberrants relèvent de l'approche statistique de modélisation des données basée sur les propriétés statistiques puis

l'utilisation du modèle produit pour estimer si un échantillon de test est généré à partir de la même distribution ou pas (Markou & Singh 2003).
Une autre méthode courante pour la détection d'anomalies est le classifieur SVM à une classe qui détermine une hyper sphère renfermant les données normales (Heller et al. 2003).

Dans cette contribution, nous décrivons et utilisons une méthode probabiliste non supervisée de détection des cas aberrants pour la vérification de l'auteur. Cette méthode est décrite dans la section suivante.

3. Méthode proposée

Dans la méthode proposée, nous abordons la tâche de la vérification de l'auteur comme problème de détection de cas aberrants. Les textes écrits par un auteur donné $\boldsymbol{X}$ sont considérées comme des cas normaux, alors que les textes qui ne sont pas écrits par cet auteur $\boldsymbol{X}$ sont considérés comme des anomalies ou des cas aberrants (texte anormal). Compte tenu du fait que les méthodes supervisées de détection de cas aberrants n'arrivent souvent pas à atteindre des taux de détection acceptable dans de nombreuses tâches, et qu'il existe toujours un besoin couteux de collection de données étiquetées pour guider la génération de modèle notamment pour les données représentant les anomalies (Görnitz et al. 2014), nous utilisons une méthode probabiliste de détection non supervisée des cas aberrants basée sur une modélisation gaussienne multivarié.
Notre approche de la détection de texte anormal consiste à construire un modèle gaussien multivarié $\boldsymbol{n}-$dimensionnel sur les marqueurs de style extraits à partir d'un échantillon de texte écrit par un auteur $\boldsymbol{X}$. Ensuite, chaque texte nouvellement arrivé, dont nous voudrions vérifier s'il est écrit par $\boldsymbol{X}$ ou pas, est mis en contraste avec le modèle probabiliste de normalité, et une distance évaluant l'homogénéité de ce texte avec le modèle est calculée. En effet, pour les données $\boldsymbol{n}-$dimensionnelles gaussiennement distribuées les valeurs des distances sont approximativement chi-carré distribués avec $\boldsymbol{n}$ degrés de liberté. Dans ce cas, les cas aberrants peuvent être simplement définis comme observations ayant une grande distance de Mahalanobis (Filzmoser 2004). Ainsi, la distance calculée décrit la probabilité que le nouveau texte soit écrit par l'auteur $\boldsymbol{X}$. Si la distance dépasse un certain seuil prédéfini $\boldsymbol{\alpha}$, l'instance est considérée comme une anomalie et le texte est considéré comme n'ayant pas été écrit par l'auteur $\boldsymbol{X}$ ou pas. Comme seuil, le quantile de la distribution chi-carré (par exemple. 97,5% quantile) peut être envisagé. Cette méthode a déjà été utilisée avec succès dans d'autres applications (Rousseeuw & Van Zomeren 1990).

Le procédé peut être formulé en trois étapes comme suit: Soit $\boldsymbol{x_i}$ un vecteur $\boldsymbol{n}-$dimensionnel quantifiant des marqueurs de style représentant le $\boldsymbol{i}-$éme texte $(\boldsymbol{i} = \boldsymbol{1},\ldots,\boldsymbol{m})$.

1. Construire un modèle multivarié $\boldsymbol{M}(\boldsymbol{x})$ sur les données normales en estimant les deux paramètres du modèle : le vecteur de location $\boldsymbol{\mu}$ et la matrice de covariance $\boldsymbol{\Sigma}$:

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \mu\right)\left(x^{(i)} - \mu\right)^{T}$$

2. Étant donné une nouvelle instance $\boldsymbol{x}$, calculer la distance de Mahalanobis par rapport aux paramètres du modèle $\boldsymbol{D_M}(\boldsymbol{x})$:

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

3. Prédire l'anomalie ($\boldsymbol{y = 1}$) de l'instance x étant donné le seuil de distance $\boldsymbol{\alpha}$:

$$y = \begin{cases} 0 & \text{if } D_M(x) < \alpha \\ 1 & \text{if } D_M(x) \geq \alpha \end{cases}$$

Pour l'expérimentation, deux seuils différents ont été considérés $\boldsymbol{\alpha_1}$ et $\boldsymbol{\alpha_2}$ pour les quantiles 95% et 97,5% respectivement de la distribution Chi-carré.
La nature des marqueurs de style utilisés comme attributs pour décrire les vecteurs $\boldsymbol{n}-$dimensionnels représentant les textes est très importante et détermine considérablement l'applicabilité et l'efficacité de notre procédé.

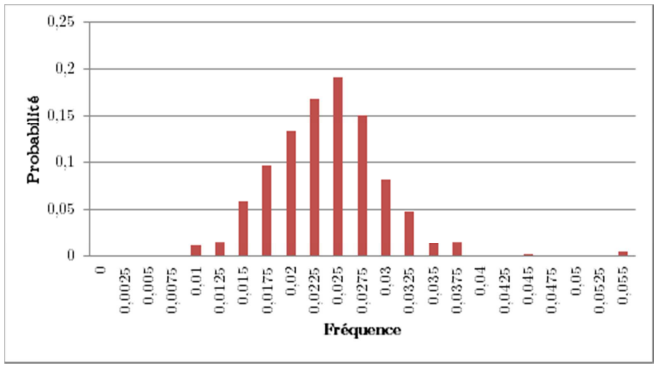

Figure 2 : Le comportement Gaussian de la probabilité de la fréquence du mot-outil français "La"

En fait, la nature de ces attributs doit respecter l'hypothèse gaussienne supposée pour former le modèle gaussien multivarié. Pour notre expérience, nous avons choisi de

prendre les fréquences des mots-outils en tant que marqueurs de style. Chaque texte dans notre ensemble de données est représenté par avec un vecteur des fréquences des 30 mots-outils les plus fréquents. Il y a deux principales raisons pour utiliser la fréquence des mots-outils comme attributs. Tout d'abord, en raison de leur haute présence dans le texte écrit, les mots-outils sont très susceptibles d'avoir un comportement gaussien (Voir la figure 2 par exemple). Aussi, les mots-outils, contrairement aux mots de contenu, sont difficiles à contrôler consciemment, ils sont donc plus indépendant du sujet ou du genre du texte (Chung & Pennebaker 2007).

4. Validation

4.1. *Données*

Pour tester l'efficacité de notre méthode, nous avons utilisé des romans écrits par Balzac, Dumas et France Anatole. Ce choix est motivé par notre intérêt particulier à étudier et à appliquer cette méthode pour la littérature française classique du 19ème siècle. Plus d'informations sur l'ensemble des textes utilisées pour l'expérimentation sont présentées dans le tableau 1. Pour chacun de ces trois auteurs mentionnés ci-dessus, nous avons recueilli quatre romans. L'étape suivante a consisté à diviser ces romans en plus petit morceaux de textes afin d'avoir suffisamment d'instances de données pour former et tester le modèle probabiliste. Les chercheurs qui travaillent sur l'attribution d'auteur dans les textes littéraires ont utilisé différentes stratégies de division. Par exemple, Hoover (2003) a décidé de ne prendre que les 10 000 premiers mots de chaque roman comme un texte unique, tandis qu'Argamon et Levitan (2005) traitaient chaque chapitre de chaque livre comme un texte séparé. Dans notre expérience, nous avons simplement choisi de découper chaque roman en parties à peu près égales de 2000 mots. Ce nombre de mots par texte est en dessous du seuil proposé par Eder (2013) précisant la plus petite taille du texte raisonnable pour parvenir à une bonne attribution. Ceci augmente le degré de la difficulté de la tâche de vérification.

Table 1 : L'ensemble de données utilisé dans notre expérience

Auteurs	# de textes
Balzac, Honoré de	126
Dumas, Alexandre	190
France, Anatole	128

4.2. *Protocole de vérification*

Dans notre expérience, les mots-outil ont d'abord été extraits. Chaque texte est alors représenté par un des vecteurs $\boldsymbol{R_i} = \{\boldsymbol{r_1}, \boldsymbol{r_2}, \ldots, \boldsymbol{r_{30}}\}$ de fréquences normalisées

d'apparition des 30 mots-outils les plus fréquents dans le corpus. Ensuite, pour chaque auteur, 75% des données générées sont utilisées pour estimer les paramètres du modèle gaussien multivarié représentant cet auteur tandis que les 25% restants sont utilisées pour le tester. Pour obtenir une estimation raisonnable de la performance, ce processus d'entrainement et de test a été répété 10 fois sur un échantillonnage avec remplacement. La performance globale de vérification de l'auteur est considérée comme la performance moyenne sur ces 10 exécutions. Pour évaluer la performance de la vérification, nous avons utilisé les mesures standards: précision (***P***), le rappel (***R***), et le $\boldsymbol{F_1}$ score.
Le SVM à une classe avec un noyau RBF a été utilisé comme algorithme de référence. Il a été entrainé et testé sur les mêmes données utilisées pour former et tester le modèle proposé pour chacune des 10 exécutions.

4.3. *Résultats*

Les résultats préliminaires de la mesure de la performance de vérification dans notre validation expérimentale sont résumés dans le tableau 2. Ces résultats montrent la nette supériorité de la méthode proposée sur la méthode de référence. Notre étude montre ici que le procédé de vérification non supervisé proposé, combiné avec des marqueurs de style à base de mots-outils fréquents, peut atteindre un rendement de vérification très acceptable (F1 = 0,83).
Comme on peut s'y attendre, l'augmentation du seuil de validité de la paternité (α_2> α_1) se traduira par un rappel plus grand et une précision plus petite, mais sans effet significatif sur le score F1. En revanche, le classifieur SVM à une classe performe particulièrement mal dans cette tâche.

Table 2 : Comparaison des performances moyennes de la vérification pour les trois auteurs en utilisant le SVM à une classe et la méthode proposée (« Unsupervised Outlier Detection », UOD)

Méthode	**P**	**R**	**F_1**
SVMs à une classe	0,34	0,50	0,40
UOD ($\propto_1$)	**0,82**	0,85	**0,83**
UOD ($\propto_2$)	0,79	**0,89**	**0,83**

Enfin, ces résultats sont en adéquation avec des travaux antérieurs qui ont affirmé que les méthodes de détection des cas aberrantes, originaires d'un paradigme de classification supervisé comme le SVM à une classe, sont souvent inappropriées et peinent à détecter les cas aberrants et les anomalies nouveaux et inconnus (Adomavicius & Tuzhilin 2005) .

5. Conclusion

Dans cet article, nous avons présenté une étude sur l'utilisation d'une méthode de détection des cas aberrants, en fonction de la distance, pour la tâche de vérification de l'auteur. Nous avons considéré une méthode non supervisée basée sur un modèle gaussien multivarié. Pour évaluer la performance de la méthode, nous avons effectué une expérimentation sur un corpus de littéraire française classique. Nos résultats préliminaires montrent que cette méthode probabiliste peut réaliser une haute performance de vérification qui peut atteindre un score F_1 de 83%.

Sur la base de ces résultats, nous avons identifié plusieurs futures pistes de recherche. Tout d'abord, nous allons explorer l'intégration de l'option de non-attribution dans notre méthode. En effet, dans le domaine de l'attribution de l'auteur, l'option de non-attribution est mieux qu'une fausse attribution. Deuxièmement, nous avons l'intention d'expérimenter avec d'autres marqueurs de style, d'autres langues et d'autres tailles de texte à l'aide d'autres corpus du domaine.

Remerciement
Ce travail a bénéficié d'une aide d'État gérée par l'Agence Nationale de la Recherche dans le cadre des Investissements d'Avenir portant la référence ANR-11-IDEX-0004-02

Références

Adomavicius, G. & Tuzhilin, A., 2005. *Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions*, IEEE Educational Activities Department.

Baayen, H. et al., 2002. An experiment in authorship attribution. In *6th JADT*. pp. 29–37.

Chandola, V., Banerjee, A. & Kumar, V., 2009. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3), p.15.

Chung, C. & Pennebaker, J.W., 2007. The psychological functions of function words. *Social communication*, pp.343–359.

Filzmoser, P., 2004. A multivariate outlier detection method. In *Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling*. pp. 18–22.

Görnitz, N. et al., 2014. Toward supervised anomaly detection. *arXiv preprint arXiv:1401.6424*.

Heller, K. et al., 2003. One class support vector machines for detecting anomalous windows registry accesses. In *Workshop on Data Mining for Computer Security (DMSEC), Melbourne, FL, November 19, 2003*. pp. 2–9.

Holmes, D.I., Robertson, M. & Paez, R., 2001. Stephen Crane and the New-York Tribune: A case study in traditional and non-traditional authorship attribution. *Computers and the Humanities*, 35(3), pp.315–331.

Kešelj, V. et al., 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*. pp. 255–264.

Koppel, M. & Schler, J., 2004. Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*. p. 62.

Koppel, M., Schler, J. & Argamon, S., 2009. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1), pp.9–26.

Kukushkina, O. V, Polikarpov, A.A. & Khmelev, D.V., 2001. Using literal and grammatical statistics for authorship attribution. *Problems of Information Transmission*, 37(2), pp.172–184.

Markou, M. & Singh, S., 2003. Novelty detection: a review—part 1: statistical approaches. *Signal processing*, 83(12), pp.2481–2497.

Rousseeuw, P.J. & Van Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411), pp.633–639.

Sebastiani, F., 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1), pp.1–47.

Stamatatos, E., 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3), pp.538–556.

Wressnegger, C. et al., 2013. A close look on n-grams in intrusion detection: anomaly detection vs. classification. In *Proceedings of the 2013 ACM workshop on Artificial intelligence and security*. pp. 67–76.

Yule, G.U., 1944. *The statistical study of literary vocabulary*, CUP Archive.

Zhao, Y. & Zobel, J., 2005. Effective and scalable authorship attribution using function words. In *Information Retrieval Technology*. Springer, pp. 174–189.