Probabilistic Anomaly Detection Method for Authorship Verification

Mohamed Amine Boukhaled, Jean-Gabriel Ganascia

Probabilistic Anomaly Detection Method for Authorship Verification

Mohamed Amine Boukhaled, Jean-Gabriel Ganascia

LIP6 (Laboratoire d'Informatique de Paris 6), Université Pierre et Marie Curie and CNRS (UMR7606), ACASA Team, 4, place Jussieu,
75252-PARIS Cedex 05 (France) ,
{mohamed.boukhaled, jean-gabriel.ganascia}@lip6.fr

Abstract. Authorship verification is the task of determining if a given text is written by a candidate author or not. In this paper, we present a first study on using an anomaly detection method for the authorship verification task. We have considered a weakly supervised probabilistic model based on a multivariate Gaussian distribution. To evaluate the effectiveness of the proposed method, we conducted experiments on a classic French corpus. Our preliminary results show that the probabilistic method can achieve a high verification performance that can reach an F_1 score of 85%. Thus, this method can be very valuable for authorship verification.

Keywords: Authorship verification; anomaly detection; multivariate Gaussian distribution

1 Introduction

Authorship verification is a special case of the authorship attribution problem. The authorship attribution problem can be generally formulated as follows: given a set of candidate authors for whom samples of written text are available, the task is to assign a text of unknown authorship to one of these candidate authors [17]. This task has been addressed mainly as a problem of multi-class discrimination, or as a text categorization task [16]. In the authorship verification problem, though, we are given samples of texts written by a single author and are asked to assess if a given different text is written by this author or not [13]. As a categorization problem, modifying the original attribution problem in this way makes the task of authorship verification significantly more difficult partly because building a characterising model of one author is much harder than building a distinguishing model between two authors [12].

Authorship verification has two key steps: an indexing step based on style markers is performed on the text using some natural language processing techniques such as

tagging, parsing, and morphological analysis; then an identification step is applied using the indexed markers to verify the validity of the authorship. Many style markers have been used to characterise writing styles, from early studies based on sentence length and vocabulary richness [19] to more recent and relevant work based on function words [9], [20], punctuation marks [2], part-of-speech (POS) tags [14], parse trees [6] and character-based features [11]. There is an agreement among researchers that function words are the most reliable indicator of authorship [17].

The verification step can be addressed as a one-class problem (written-by-the-author) or as a binary classification problem (written-by-the-author as positive vs not-written-by-the-author as negative). However, both of these formulations of the problem have drawbacks: In the case of binary classification, one should collect a reasonable amount of representative texts of the entire "not-written-by-the-author" class, which is difficult, if not impossible. In the case of one-class classification, one does not take advantage from negative examples that we do not actually lack for them even though they are not representative of the entire class.

In this paper, we address the authorship verification problem as an anomaly detection problem where texts written by the candidate author are seen as normal data while texts not written by that author are seen anomalous data. We propose a probabilistic anomaly detection method that can benefit from negative examples for the authorship verification process.

We first give an overview of the anomaly detection problem in section 2 and then describe our method in section 3. We than experimentally validate the proposed method in section 4 using a classic French corpus. Finally we use this method to settle a literary mystery case.

2 Anomaly Detection

Anomaly detection is a challenging task which consists of identifying patterns in data that do not conform to expected (normal) behaviour. These non-conforming patterns are called anomalies or outliers [3]. Anomaly detection has been successfully used in many applications such as fault detection, radar target detection and hand written digit recognition [15].

This technique has also been used to deal with textual data for various purposes such as detecting novel topics, events, or news stories in a collection of documents or news articles [3]. Anomaly detection is based on the idea that one can never train a classification algorithm on all the possible classes that the system is likely to encounter in real application. Anomaly detection is also suitable for situations in which the class imbalance problem can affect the accuracy of classification (see Fig. 1) [18].

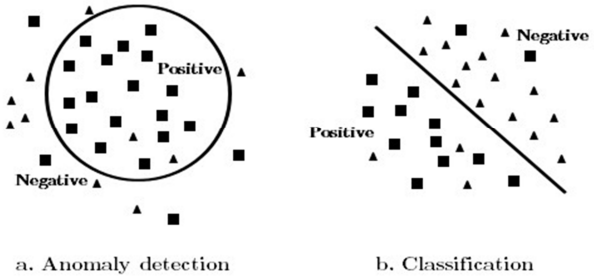

Fig. 1. The anomaly detection and the classification learning schemas

Many anomaly detection techniques fall under the statistical approach of modelling data based on its statistical properties and using this information to estimate whether a test sample comes from the same distribution or not [15]. Another common method for anomaly detection is the one-class SVM that determines a hyper sphere enclosing the normal data [8]. In this contribution, we describe and use a probabilistic anomaly detection method for authorship verification that straightforwardly follows the definition given above. The method is discussed in the next section.

3 Proposed Method

In our method, we address the authorship verification problem as an anomaly detection problem where texts written by a given author X are seen as normal data, while texts not written by that author X are seen anomalous data. We use a probabilistic anomaly detection method that can benefit from anomalous examples for the authorship verification process based on a multivariate Gaussian modelling. Given the fact that unsupervised anomaly detection approaches often fail to match the required detection rates in many tasks and there exists a need for labelled data to guide the model generation [7], our proposed methods is weakly supervised in the sense that it takes into consideration a small amount of representative anomalous data for the model generation.

The approach to anomalous text detection is to train a multivariate Gaussian distribution model on the style markers extracted from sample of text written by an author X. Every newly arriving text (data instance) that we went to verify as written by X or not is contrasted with the probabilistic model of normality, and a normality probability is computed. The probability describes the likelihood of the new text to have been written by X compared to the average data instances seen during the training. If the probability does not surpass a predefined threshold α, the instance is considered an anomaly and the text is considered not to have been written by the author X. To define the probability threshold, we cross-validate over a data set containing both anomalous

and non-anomalous data and we set the threshold to the value that maximizes the authorship verification performance on this data set. The method can be formulated into three steps as follow: Let x_i be a n-dimensional vector representing the text i $(i = 1, \dots, m)$.

1. Train a Multivariate Gaussian distribution model $M(x)$ on the normal data. This is done by estimating the two distribution parameters: the multivariate location μ and the covariance matrix Σ :

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

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} \left(x^{(i)} - \mu\right)\left(x^{(i)} - \mu\right)^T \tag{2}$$

2. Given a new instance x, compute the probability $p(x)$:

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right) \tag{3}$$

3. Predict the anomaly $(y = 1)$ of the instance x given the probability threshold $\propto$:

$$y = \begin{cases} 1 \; if \; p(x) < \propto \\ 0 \; if \; p(x) \geq \propto \end{cases} \tag{4}$$

The nature of the style markers used as attributes to describe and to get an n-dimensional vector representing the text is very important and determines the applicability of our method. In fact, the nature of these attributes should respect the Gaussian assumption made to train the multivariate Gaussian model. For our experiment, we chose to test this method on two types of style markers separately. Each text in our data set is mapped onto a vector of the frequency of the most frequent function words and a vector of the frequency of POS-tags.

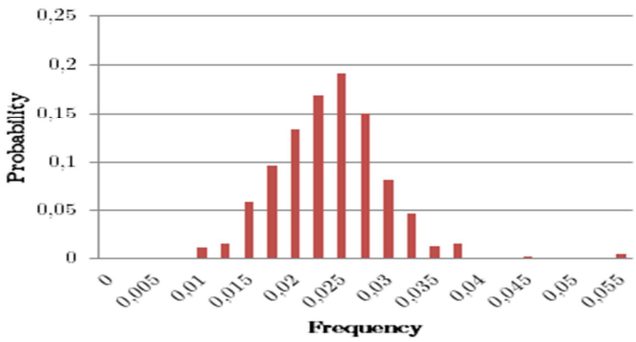

Fig. 2. The probability of frequency of the French function word "de" has a Gaussian behavior

There are two main reasons for using the frequency of function words as attributes. First, because of their high frequency in a written text, function words are very likely to have a Gaussian behaviour (see Figure 2). Secondary, function words, unlike content words, are difficult to consciously control, thus they are more independent from the topic or the genre of the text [4]. In fact, Koppel and Schler found that all the work of distinguishing the styles of different authors is accomplished with a small set of features containing frequent function words [12]. Based on that information and to get a right balance between the features-set size and the dataset size, we limit our study to the most 30th frequent function words. The part-of-speech-based markers are also shown to be very effective because they partly share the advantages of function words.

4 Experimental Validation

4.1 Data Set

To test the effectiveness of our method, we used novels written by: Balzac, Dumas and France. This choice was motivated by our special interest in studying the classic French literature of the 19th century, and by the availability of electronic texts from these authors on the project Gutenberg website1 and in the Gallica electronic library2. Our choice of authors was also affected by the fact that we wished to get a challenging problem since these three authors are knows to have relatively comparable syntactic styles. More information about the data set used for the experimentation is summarized in Table 1.

For each of the three authors mentioned above, we collected 4 novels, so that the total number of novels was 12. The next step was to divide these novels into smaller pieces of texts in order to have enough data instances (artificial documents) to train and test the probabilistic model. Researchers working on authorship attribution in literary texts have used different dividing strategies. For example, Hoover [10] decided to take just the first 10,000 words of each novel as a single text, while Argamon and Levitan [1] treated each chapter of each book as a separate text. In our experiment, we chose simply to chunk each novel into approximately equal parts of 2000 words, which is below the threshold proposed by Eder [5] specifying the smallest reasonable text size to achieve good attribution. This increases the degree of the difficulty of the task.

[1] http://www.gutenberg.org/
[2] http://gallica.bnf.fr/

Table 1. Data set used in our experiment

Author Name	# of texts
Balzac, Honoré de	126
Dumas, Alexandre	190
France, Anatole	128

4.2 Verification Protocol

In our experiment, the corpus was POS tagged and function words were extracted. Each text is then represented by two vectors $R_n = \{r_1, r_2, \dots, r_n\}$, one for the normalized frequencies of occurrence of the top 30 function words in the corpus, and another for the normalized frequencies of occurrence of POS-tags. The normalization of the vectors of frequency representing a given text was done according to the size of the text. Then, for each author, we used 75% of the data generated by texts written by this author to estimate the parameters of the model representing this author, and 20% of the data from each author for testing it. The remaining 5% data was merged with 5% of the data (anomalous data) generated by each one of the other authors and was used to estimate the probability threshold á. To get a reasonable estimate of the expected generalization performance, we used a resampling with replacement method. The training and testing process was done 10 times. The overall authorship verification performance is taken as the average performance over these 10 runs. For evaluating the verification performance, we used the standard measures, calculating precision (P), recall (R), and $F_{\hat{a}}$ where:

$$F_{\hat{a}} = \frac{(1+\hat{a}^2)RP}{(\hat{a}^2 R)+P} \tag{5}$$

We consider precision and recall to have the same value, so we set â equal to 1.

4.3 Baselines

To evaluate the effectiveness of the proposed method we used one-class SVM and binary SVMs classifier using RBF kernel (best performing). The one-class SVM was trained and tested on the same data used to train and test the multivariate Gaussian model respectively. The binary SVM classifier was trained on both the data used to train the probabilistic model and the data used to estimate the probability threshold, and it was tested on the same data as our probabilistic model. The overall baseline classification performance is taken as the average performance over the 10 runs.

4.4 Results

The results of measuring the verification performance for the two different style markers presented in our experimental validation are summarized in Table 2 for func-

tion words and in Table 3 for POS tags. These results show in general the superiority of the proposed method over the baselines in terms of F_1 score and recall. These results also show in general a better performance when using frequent function words than POS-tag for both the proposed method and the baselines.
Our study here indicates that the multivariate Gaussian model for anomaly detection combined with features based on frequent function words can achieve a high verification performance (e.g., F1 = 0.85). By contrast, the one-class SVM performs particularly poorly on this task. The binary SVM achieved relatively good results but doesn't outperform the probabilistic model; this shows that the authorship verification problem should not be handled as a binary class problem unless a sufficient amount of representative negative data is present to avoid the class imbalance problem.

Table 2. Results of the authorship verification using frequent function words

Method	**P**	**R**	**F_1**
One-class SVMs	0,34	0,50	0,40
Binary SVMs	**0,86**	0,75	0,80
Multivariate Gaussian Model	0,82	**0,88**	**0,85**

Table 3. Results of the authorship verification using frequent POS-tags

Method	**P**	**R**	**F_1**
One-class SVMs	0,51	0,45	0,48
Binary SVMs	**0,81**	0,58	0,67
Multivariate Gaussian Model	0,69	**0,89**	**0,77**

Finally, these results are in line with previous work that claimed that semi-supervised anomaly detection approaches, originating from a supervised classifier, are inappropriate and hardly detect new and unknown anomalies, and that semi-supervised anomaly detection needs to be grounded in the unsupervised learning paradigm [7].

5 A Classic French Literary Mystery: "Le Roman de Violette"

In this section, we apply our probabilistic method to settle one of the classic French literary mysteries. "Le Roman de Violette"3 is a novel published in 1883. The authorship of this novel has still not been determined. Even though the novel was edited under the name of Alexandre Dumas, some literary critics state that a serious candidate for its authorship is "La Marquise de Mannoury d'Ectot". But this hypothesis

[3] http://ero.corneille-moliere.com/?p=page52&m=ero&l=fra

cannot be definitely proved, partly because there is only one known book written by that author, which limits the quantity of text available to validate the computational authorship identification methods including our method.

We applied our proposed authorship verification method to handle this case. Since there is not enough available text written by “La Marquise de Mannoury d’Ectot” to verify whether she is the writer of “Le Roman de Violette” or not, we set Alexandre Dumas as the author candidate that we want to verify as the writer or not. We trained the probabilistic model based on frequent function words on texts written by Alexandre Dumas. The only known book written by “La Marquise de Mannoury d’Ectot” was used as the representative anomalous text to set the probability threshold. Finally, the verification test was performed on the “Roman de Violette”. The authorship probability produced by the novel using our proposed method is under the threshold needed to validate the authorship. This result suggests that the novel “Le Roman de Violette” was not written by Alexandre Dumas.

6 Conclusion

In this paper, we have presented a study on using an anomaly detection method for the authorship verification task. We have considered a weakly supervised probabilistic model based on a multivariate Gaussian distribution. To evaluate the effectiveness of the proposed method, we conducted experiments on a classic French literary corpus. Our preliminary results show that the probabilistic method can achieve a high verification performance that can reach an F1 score of 85%.

Based on the current study, we have identified several future research directions. First, we will explore incorporating the non-verification option into our probabilistic model. In fact, in the field of authorship identification, the non-attribution option is better than a false attribution. Second, this study will be expanded to include more style markers. Third, we intend to experiment with other languages and text sizes using standard corpora employed in the field at large.

Acknowledgment

This work was supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02.

References

1. Argamon, S., & Levitan, S.: Measuring the usefulness of function words for authorship attribution. *In Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing (2005)*

2. Baayen, H., van Halteren, H., Neijt, A., & Tweedie, F.: An experiment in authorship attribution. In *6th JADT* (pp. 29–37) (2002)
3. Chandola, V., Banerjee, A., & Kumar, V.:Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, *41*(3), 15 (2009)
4. Chung, C., & Pennebaker, J. W.: The psychological functions of function words. *Social Communication*, 343–359 (2007)
5. Eder, M.: Does size matter? Authorship attribution, small samples, big problem. *Literary and Linguistic Computing*, fqt066 (2013)
6. Gamon, M.: Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the 20th international conference on Computational Linguistics* (p. 611) (2004)
7. Görnitz, N., Kloft, M. M., Rieck, K., & Brefeld, U. (2014). Toward supervised anomaly detection. *arXiv Preprint arXiv:1401.6424*.
8. Heller, K., Svore, K., Keromytis, A. D., & Stolfo, S.: 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) (2003)
9. Holmes, D. I., Robertson, M., & Paez, R.. Stephen Crane and the New-York Tribune: A case study in traditional and non-traditional authorship attribution. *Computers and the Humanities*, *35*(3), 315–331 (2001)
10. Hoover, D. L.: Frequent collocations and authorial style. *Literary and Linguistic Computing*, *18*(3), 261–286 (2003)
11. Kešelj, V., Peng, F., Cercone, N., & Thomas, C.. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING* (Vol. 3, pp. 255–264) (2003)
12. Koppel, M., & Schler, J.: Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning* (p. 62) (2004)
13. Koppel, M., Schler, J., & Argamon, S.: Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, *60*(1), 9–26 (2009)
14. Kukushkina, O. V, Polikarpov, A. A., & Khmelev,V.: Using literal and grammatical statistics for authorship attribution. *Problems of Information Transmission*, *37*(2), 172–184 (2001)
15. Markou, M., & Singh, S.: Novelty detection: a review—part 1: statistical approaches. *Signal Processing*, *83*(12), 2481–2497 (2003)
16. Sebastiani, F.. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, *34*(1), 1–47 (2002)
17. Stamatatos, E.. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, *60*(3), 538–556 (2009)
18. Wressnegger, C., Schwenk, G., Arp, D., & Rieck, K.: 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) (2013)
19. Yule, G. U.: The statistical study of literary vocabulary. CUP Archive (1944)
20. Zhao, Y., & Zobel, J.: Effective and scalable authorship attribution using function words. In *Information Retrieval Technology* (pp. 174–189). Springer (2005)