An Off-the-shelf Approach to Authorship Attribution

Nasir, Jamal Abdul; Görnitz, Nico; Brefeld, Ulf

Published in:
COLING 2014 - 25th International Conference on Computational Linguistics, Proceedings of COLING 2014

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

An Off-the-shelf Approach to Authorship Attribution

Jamal Abdul Nasir
Dep. of Computer Science
LUMS Lahore
Pakistan
jamaln@lums.edu.pk

Nico Görnitz
Dep. of Computer Science
TU Berlin
Germany
goernitz@tu-berlin.de

Ulf Brefeld
Dep. of Computer Science
TU Darmstadt
Germany
brefeld@cs.tu-darmstadt.de

Abstract

Authorship detection is a challenging task due to many design choices the user has to decide on. The performance highly depends on the right set of features, the amount of data, in-sample vs. out-of-sample settings, and profile- vs. instance-based approaches. So far, the variety of combinations renders off-the-shelf methods for authorship detection inappropriate. We propose a novel and generally deployable method that does not share these limitations. We treat authorship attribution as an anomaly detection problem where author regions are learned in feature space. The choice of the right feature space for a given task is identified automatically by representing the optimal solution as a linear mixture of multiple kernel functions (MKL). Our approach allows to include labelled as well as unlabelled examples to remedy the in-sample and out-of-sample problems. Empirically, we observe our proposed novel technique either to be better or on par with baseline competitors. However, our method relieves the user from critical design choices (e.g., feature set) and can therefore be used as an off-the-shelf method for authorship attribution.

1 Introduction

Automatically attributing a piece of text to its author is one of the oldest problems studied in linguistics (Mendenhall, 1887). Despite being an old problem, authorship attribution is still highly topical and todays applications range from plagiarism detection (Maurer et al., 2006), identifying the origin of anonymous harassments in emails, blogs, and chat rooms (Tan et al., 2013) to copyright and estate issues as well as resolving historical questions of disputed authorship (Mosteller and Wallace, 1964; Fung, 2003).

Intrinsically, the goal of authorship detection is to identify the characteristic traits of an author. The idea is that, these traits distinguish an author from others in terms of writing style, use of words, etc. Thus, prior work often focuses on designing and extracting features from text to capture these traits. There is a great deal of features proposed for authorship detection, including word or character n-grams (Burrows, 1987; Houvardas and Stamatatos, 2006), part-of-speech (Stamatatos et al., 2001), probabilistic context-free grammars (Raghavan et al., 2010), or linguistic features (Koppel et al., 2006). However, indicative features for one author do not necessarily help to characterise another. A major problem in authorship detection is therefore to find the right set of features for a given task at hand (Forman, 2003).

Algorithmically, a variety of different models have been studied in the context of authorship detection, ranging from probabilistic approaches (Seroussi et al., 2011) and similarity-based methods (Koppel et al., 2011) to vector space models (Fung, 2003; Li et al., 2006). The approaches either treat documents as independent (instance-based) or concatenate documents by the same author (profile-based). Intuitively, the latter is helpful if an author has a concise way of expressing herself so that the concatenated document allows to extract a statistic that is sufficient for capturing her style. On the other hand, instance-based approaches are better suited for expressive authors and have advantages in sparse data scenarios.

Another aspect in authorship attribution is the application scenario of the final model. In transductive (in-sample) settings, the unlabelled documents of interest are already included in the training process

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 895–904, Dublin, Ireland, August 23-29 2014.

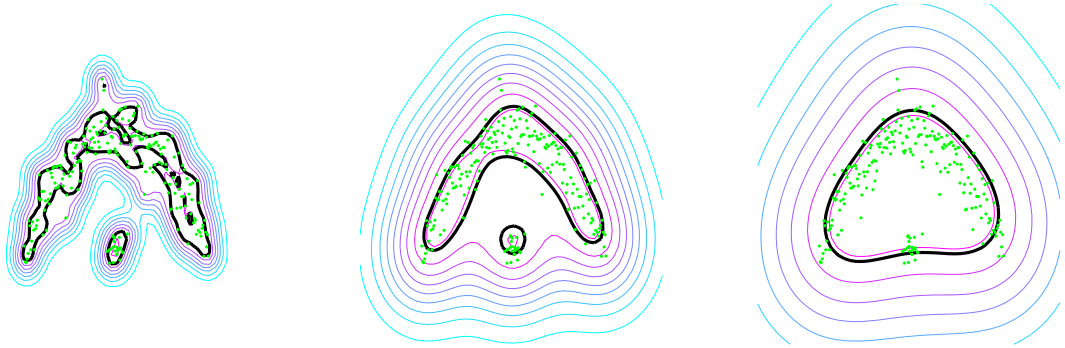

Figure 1: Three solutions of an anomaly detection problem where data is represented by RBF kernels with different band-width parameters. Combining anomaly detection with multiple kernel learning allows to include all three kernels simultaneously in the optimisation and to find the optimal linear mixture of the three (or more) kernels together automatically for a given task.

and the model does not necessarily perform well on new and unseen texts. By contrast, inductive (out-of-sample) scenarios generally allow to learn models that can be applied to any future text but require larger training samples to achieve accurate performances.

In this paper, we propose a general machine learning-based approach to authorship detection. Our approach remedies the above mentioned problems by fusing existing techniques: (i) We cast authorship attribution as an anomaly detection problem where one model is learned for every author. The idea is to identify a concise region in feature space that contains (most of) the documents of the author of interest while other documents are considered outliers. Thus, the model can be viewed as a profile-based approach in feature space while the data is treated on an instance-based level. (ii) We remedy the in-sample / out-of-sample problem by providing a semi-supervised extension of the commonly unsupervised outlier detection framework. By doing so, we may include authorship labels for the already known documents and leave the disputed ones unlabelled. (iii) Finally, we devise our model consequentially as a member of the multiple kernel learning family to automatically include a mathematically well founded feature selection framework that renders the method generally applicable. The optimal solution is given by a (possibly sparse) linear mixture of kernel functions.

Empirically, we observe that our approach consistently outperforms baseline competitors or confirms common knowledge with respect to the authorship of disputed articles. The main advantage of the method however lies in its simplicity. Practitioners do not need to take critical design choices in terms of which features to use and which not. By contrast, all features (kernels) can be used in the optimisation and the method itself finds the optimal combination for the problem at hand.

The remainder of this paper is structured as follows. Section 2 reviews related work. Our main contribution is presented in Section 3. We report on empirical results in Section 4 and Section 5 concludes.

2 Related Work

Authorship attribution using linguistic and stylistic features has a long tradition and can be dated back to the nineteenth century. As a first attempt, Mendenhall (Mendenhall, 1887) uses features based on word lengths to characterise the plays of Shakespeare. Later in the first half of the 20th century, different textual statistics, such as Zipf's distribution (Zipf, 1932) and Yule's k-statistic (Yule, 1944) have been proposed to quantify textual style. A study conducted by (Mosteller and Wallace, 1964) is one of the most influential modern work in authorship attribution. They use a Bayesian approach to analyse frequencies of a small set of function words for *The Federalist Papers*, a series of 85 political essays written by John Jay, Alexander Hamilton, and James Madison. Until the late 1990s, research in *stylometry* has been dominated by feature engineering to quantify writing style (Holmes, 1998) and about 1,000 different measures have been proposed (Rudman, 1997).

Document representation is essential for author attribution tasks. Features aim to capture characteristic traits of authors that persist across topics. Traditional stylometric features include function and high-frequency words, hapax legomena, Yules k-statistic, syllable distributions, sentence length, word length and word frequencies, vocabulary richness functions as well as syntactic features. Many studies combine features of different types using multivariate analyses. Some researchers use punctuation symbols while others experiment with n-grams (Diederich et al., 2003). Grammatical style markers with natural language processing techniques are also used to extract features from the documents.

Also in terms of technical approaches, authorship attribution has been studied with a wide range of different approaches. The deployed techniques can be broadly divided into three categories: machine learning (Diederich et al., 2003), multivariate/cluster analysis (Khmelev, 2000), and natural language processing (Stamatatos et al., 2000). Principal components analysis (PCA) is one of the widely used techinques for authorship studies, for instance, (Holmes and Crofts, 2010) apply PCA to identify the authorship of unknown articles that have been attributed to Stephen Crane. In addition, machine learning-based approaches, including neural networks (Neme et al., 2011) and support vector machines (SVMs) (Diederich et al., 2003), are frequently used to discriminate documents by different authors. An excellent survey on the diversity of approaches for authorship detection is provided by (Stamatatos, 2009).

Density level set estimation, also known as one-class learning (Tax and Duin, 1999; Schölkopf et al., 1999), is the problem of learning a representation of a single class of interest, rejecting data points that deviate from the learned model of normality. Thus, it has been proven very successful in anomaly detection scenarios such as network intrusion detection (Görnitz et al., 2009). Various extensions have been proposed, i.e. to incorporate prior and additional knowledge (Görnitz et al., 2013; Blanchard et al., 2010) in a semi-supervised fashion (Chapelle et al., 2006) and to learn linear combinations of kernels (Kloft et al., 2011; Rakotomamonjy et al., 2008) which is especially usefull whenever the right choice of representation is unknown.

3 Methodology

In this section, we cast semi-supervised anomaly detection as an instance of multiple kernel learning. The rational for this idea is shown in Figure 1. The figure shows three solutions for an anomaly detection task. The data is represented by RBF kernels with different band-width parameters. As shown in the figure, the choice of the band width parameter is crucial and leads to very different solutions. Usually, kernel parameters are thus included in the model selection and their optimisation is often time consuming. Fusing the anomaly detection with multiple kernel learning allows to include all three kernels simultaneously in the optimisation and to find the best linear mixture of the three (or more) kernels together with model parameters for the data at hand.

We briefly review anomaly detection and semi-supervised anomaly detection in Sections 3.1 and 3.2, respectively, and present our main contribution, multiple kernel learning for anomaly detection, in Section 3.3.

3.1 Preliminaries

Anomaly detection is often cast as an unsupervised one-class problem where the goal is to find a hyperplane that separates the majority of the input examples from the origin with maximum margin, so that, by definition, points not exceeding the hyperplane are considered outliers. Analogously, we aim to learn a separating hyperplane for articles by an author of interest, such that documents not exceeding the learned hyperplane are written by other authors.

Given a n articles $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of possibly different authors, a straight forward optimisation problem that identifies the hyperplane in terms of its normal vector $\mathbf{w}$ and threshold ρ is known as one-class support vector machine (Tax and Duin, 1999; Schölkopf et al., 1999) and given by

$$\begin{aligned} \min_{\mathbf{w},\rho,\boldsymbol{\xi}\geq 0} \quad & \frac{1}{2}\| \mathbf{w} \|_2^2 - \rho + \eta_u \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad \forall_{i=1}^{n} : \quad & \langle \mathbf{w}, \phi(\boldsymbol{x}_i) \rangle \geq \rho - \xi_i. \end{aligned}$$

The hyperplane is realised by the decision function

$$f(\boldsymbol{x}) = \langle \mathbf{w}, \phi(\boldsymbol{x}) \rangle - \rho$$

and new articles are credited to the author if $f(\boldsymbol{x}) > 0$ and are considered work by someone else if $f(\boldsymbol{x}) \leq 0$. The threshold ρ can be interpreted as a measure of expressiveness of an author. E.g., authors who have a very clear and concise style realise smaller thresholds than expressive authors that may adopt to different writing styles.

3.2 Semi-supervised Anomaly Detection

Using only unlabelled data is usually leading to inaccurate models in the presence of only a few data points. We therefore extend the problem setting to include m labeled examples $(\boldsymbol{x}_{n+1}, y_{n+1}), \ldots, (\boldsymbol{x}_{n+m}, y_{n+m})$ in addition to the n unlabelled ones. Labels $y_i \in \{+1, -1\}$ are considered binary, that is in case $y_i = +1$, document $\boldsymbol{x}_i$ belongs to the author of interest. To combine sums and hence, improve readability, we introduce labels $y_i = +1 \; \forall \, i = 1, \ldots, n$ for all unlabelled examples and an indicator function $\mathbb{I}_c \equiv [c > n]$ to mask labeled examples; the function $\mathbb{I}_c$ simply returns 1 if $c > n$ and 0 otherwise.

A semi-supervised generalisation of the hypersphere model of the previous section is the convex semi-supervised anomaly detection (SSAD) (Görnitz et al., 2013) which uses an L_2-regularizer together with the hinge-loss. Let γ be the margin for the labeled examples and κ, η_u, and η_l trade-off parameters, the optimisation problem becomes

$$\begin{aligned} \min_{\mathbf{w}, \rho, \gamma \geq 0, \boldsymbol{\xi} \geq 0} \quad & \frac{1}{2} \| \mathbf{w} \|_2^2 - \rho - \kappa\gamma + \sum_{i=1}^{n+m} (\mathbb{I}_i \eta_l + (1 - \mathbb{I}_i)\eta_u)\xi_i \\ \text{s.t.} \quad & \forall_{i=1}^{n+m} : \; y_i \langle \mathbf{w}, \phi(\boldsymbol{x}_i) \rangle \geq y_i \rho + \mathbb{I}_i \gamma - \xi_i. \end{aligned}$$

The solution $\boldsymbol{w}$ admits a dual representation and can be written as

$$\boldsymbol{w} = \sum_{i=1}^{n+m} \alpha_i y_i \phi(\boldsymbol{x}_i)$$

and hence, the decision function depends only on inner products of the input examples wich paves the way for kernel functions $K_\phi(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle$ (see (Müller et al., 2001) for an introduction to kernels). It holds

$$f(\boldsymbol{x}) = \sum_{i=1}^{n+m} \alpha_i y_i \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}) \rangle - \rho = \sum_{i=1}^{n+m} \alpha_i y_i K_\phi(\boldsymbol{x}_i, \boldsymbol{x}) - \rho.$$

We omit the subscript ϕ in the remainder to not clutter notation unnecessarily.

3.3 Multiple Kernel Learning for Anomaly Detection

Learning linear combinations of multiple kernels is an appealing strategy when the right choice of representations is unknown. We therefore generalise the semi-supervised anomaly detection of the previous section as a member of the multiple kernel learning framework (Lanckriet et al., 2004). Thus, we aim to learn a weighted combination of T kernels with mixing coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_T)$:

$$\begin{aligned} K_{\text{MKL}}(\boldsymbol{x}, \boldsymbol{x}') &:= \sum_{t=1}^{T} \beta_{\text{t}} \, K_t(\boldsymbol{x}, \boldsymbol{x}') = \sum_{t=1}^{T} \beta_{\text{t}} \langle \phi_t(\boldsymbol{x}), \phi_t(\boldsymbol{x}') \rangle \\ &= \sum_{t=1}^{T} \langle \sqrt{\beta_{\text{t}}} \phi_t(\boldsymbol{x}), \sqrt{\beta_{\text{t}}} \phi_t(\boldsymbol{x}') \rangle. \end{aligned}$$

In general, properties of the mixing coefficients include (i) non-negativity, hence $\beta_{\mathrm{t}} \geq 0$ and (ii) normalisation $\| \boldsymbol{\beta} \|_p = 1$. Recent work (Kloft et al., 2011) suggests to use the more general p-norm instead of a common 1-norm (Lanckriet et al., 2004; Bach et al., 2004; Rakotomamonjy et al., 2008). The latter usually leads to sparse mixing coefficients which is not appealing in every situation whereas p-norm with $1 \leq p \leq \inf$ admits sparsity adjustments for the problem at hand and thus adds flexibility. Incorporating multiple feature representations in our model introduced in Section 3.1 leads to

$$f_{\mathrm{MKL}}(\boldsymbol{x}) = \sum_{t=1}^{T} \langle \hat{\mathbf{w}}_t, \sqrt{\beta_{\mathrm{t}}} \phi_t(\boldsymbol{x}) \rangle - \rho = \sum_{t=1}^{T} \sqrt{\beta_{\mathrm{t}}} \langle \hat{\mathbf{w}}_t, \phi_t(\boldsymbol{x}) \rangle - \rho. \tag{1}$$

Due to technical reasons, i.e. to preserve convexity, we substitute the model parameters $\mathbf{w}_t = \sqrt{\beta_t} \hat{\mathbf{w}}_t$ and arrive at the revised primal MKL-SSAD optimisation problem:

$$\begin{aligned} \min_{\{\mathbf{w}_t\}, \rho, \gamma \geq 0, \boldsymbol{\xi} \geq 0, \boldsymbol{\beta} \geq 0} \quad & \frac{\lambda}{2} \| \boldsymbol{\beta} \|_p^2 + \frac{1}{2} \sum_{t=1}^{T} \frac{1}{\beta_{\mathrm{t}}} \| \mathbf{w}_t \|_2^2 - \rho - \kappa\gamma + \sum_{i=1}^{n+m} (\mathbb{I}_i \eta_l + (1 - \mathbb{I}_i) \eta_u) \xi_i \\ \text{s.t.} \quad & \forall_{i=1}^{n+m} : \; y_i \sum_{t=1}^{T} \langle \mathbf{w}_t, \phi_t(\boldsymbol{x}_i) \rangle \geq y_i \rho + \mathbb{I}_i \gamma - \xi_i. \end{aligned} \tag{2}$$

(Kloft et al., 2011) prove the equivalence of Tikhonov and Ivanov regularisation which allows to move the regulariser on the mixing coefficients in the objective function. We will exploit this relation on various occasions in this section. Deriving the Lagrange dual problem, we arrive at the intermediate saddle point problem

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \min_{\{\mathbf{w}_t\}, \boldsymbol{\beta} \geq 0} \; & \frac{\lambda}{2} \| \boldsymbol{\beta} \|_p^2 + \frac{1}{2} \sum_{t=1}^{T} \frac{1}{\beta_{\mathrm{t}}} \| \mathbf{w}_t \|_2^2 - \sum_{i=1}^{n+m} \alpha_i y_i \sum_{t} \langle \mathbf{w}_t, \phi_t(\boldsymbol{x}_i) \rangle \\ \text{s.t.} \quad & \kappa \leq \sum_{i=1}^{n+m} \mathbb{I}_i \alpha_i \quad \text{and} \quad 0 \leq \alpha_i \leq \mathbb{I}_i \eta_l + (1 - \mathbb{I}_i) \eta_u \quad \forall\, i \end{aligned}$$

We are solving the optimisation problem in a block-coordinate descent fashion by alternating between $\mathbf{w}$ and $\boldsymbol{\beta}$. This enables us to compute the latter analytically assuming fixed variables $\mathbf{w}$ and setting the partial derivative to zero:

$$\lambda \, {\beta_{\mathrm{t}}}^{p-1} \| \boldsymbol{\beta} \|_p^{2-p} - \frac{\| \mathbf{w}_t \|_2^2}{{\beta_{\mathrm{t}}}^2} = 0.$$

Therefore, given $\Upsilon \geq 0$ we get

$$\beta_{\mathrm{t}} = \Upsilon \| \mathbf{w}_t \|_2^{\frac{2}{p+1}}.$$

Furthermore, it holds that at any optimal point $\| \boldsymbol{\beta} \|_p = 1$ and solving for Υ gives $\Upsilon = 1/(\sum_{t=1}^{T} \| \mathbf{w}_t \|_2^{\frac{2p}{p+1}})^{\frac{1}{p}}$. Putting things together, gives the analytical update rule

$$\beta_{\mathrm{t}} = \frac{\| \mathbf{w}_t \|_2^{\frac{2}{p+1}}}{(\sum_{t=1}^{T} \| \mathbf{w}_t \|_2^{\frac{2p}{p+1}})^{\frac{1}{p}}} \tag{3}$$

which, since only norms are involved, ensures non-negativity for the mixing coefficients. Substituting $\mathbf{w}_t$ using the representer theorem $\mathbf{w}_t = \beta_{\mathrm{t}} \sum_{i=1}^{n+m} \alpha_i \boldsymbol{y}_i \phi_t(\boldsymbol{x}_i)$ yields the final optimisation problem for

Algorithm 1 Proposed optimization algorithm for MKL-SSAD (2)

Require: $\boldsymbol{x}, \boldsymbol{y}, \eta_u, \eta_l, \kappa$ & $p-$ norm
  Initialize kernel mixture coefficients such that $\| \boldsymbol{\beta}^{z=0} \|_p = 1$
  while Until Convergence **do**
    Step 1: solve the convex SSAD problem as stated in Eqn. (4)
    $\boldsymbol{\alpha}^{z+1} = \text{argmax}_{\boldsymbol{\alpha}: 0 \leq \alpha_i \leq \mathbb{I}_i \eta_l + (1-\mathbb{I}_i)\eta_u} J(\boldsymbol{\alpha}, \boldsymbol{\beta}^z) \quad \text{s.t.} \quad \kappa \leq \sum_{i=1}^{n+m} \mathbb{I}_i \alpha_i$
    Step 2: optimize the weights according to Eqn. (3)
    $\boldsymbol{\beta}^{z+1} = \text{argmin}_{\boldsymbol{\beta} \geq 0} J(\boldsymbol{\alpha}^{z+1}, \boldsymbol{\beta}) \quad \text{s.t.} \quad \| \boldsymbol{\beta} \|_p^2 \leq 1$
    z=z+1
  end while
  return Trained parameter vector $\boldsymbol{\alpha}^\star$, weights $\boldsymbol{\beta}^*$

MKL-SSAD:

$$\max_{\boldsymbol{\alpha}} \min_{\boldsymbol{\beta}: \| \boldsymbol{\beta} \|_p^2 \leq 1} \overbrace{-\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \sum_{t=1}^{T} \beta_t K_t(\boldsymbol{x}_i, \boldsymbol{x}_j)}^{=: J(\boldsymbol{\alpha}, \boldsymbol{\beta})} \tag{4}$$
$$\text{s.t.} \quad \kappa \leq \sum_{i=1}^{n+m} \mathbb{I}_i \alpha_i \quad \text{and} \quad 0 \leq \alpha_i \leq \mathbb{I}_i \eta_l + (1 - \mathbb{I}_i)\eta_u \quad \forall\, i$$

As a block-coordinate descent method, we can iteratively alternate between the two optimisation blocks and every limit point of Algorithm 1 is a globally optimal point (cmp. also (Kloft et al., 2011)). Algorithm 1 summarises the proposed optimisation procedure.

4 Empirical Results

In this section, we empirically evaluate the benefit of fusing semi-supervised anomaly detection with multiple kernel learning. We experiment on two data sets, the Reuters-50-50 corpus in Section 4.1 and the Federalist Papers in Section 4.2

4.1 Reuters 50-50

We use a subset of the Reuters 50-50 data set[1] to evaluate the performance of the aforementioned approaches. The reduced data contains 1000 articles written by 10 authors, Aaron Pressman, Alan Crosby, Alexander Smith, Benjamin Kang Lim, Bernard Hickey, Brad Dorfman, Darren Schuettler, David Lawder, Edna Fernandes, and Eric Auchard.

We deploy the following four kernels to represent documents: the first kernel is made of 484 function words taken from (Koppel and Schler, 2003), the second contains part-of-speech (POS) tags, the third is assembled by 3-letter suffixes, the last one simply a bag-of-words (BOW) kernel. We split the data into training (90%) and test (10%) sets and conduct a 10-fold cross-validation on the training set for model selection. The best performing models are then evaluated on the test set. In this set of experiments, we use a transductive setting where all training instances are labeled and only holdout and test articles are unlabelled. We compare the performance of our approach with different p-norms to the SSAD which uses one kernel at a time. For our MKL-based approach we use p-norms in the set $p \in \{1, 1.7783, 3.1623, 5.6234, 10\}$.

The results in terms of averaged micro- and macro-F_1 measures are shown in Table 1. MKL consistently outperforms the single-kernel baseline for all p-norms. That is, instead of extensively experimenting with SSAD and different kernel functions and parameter selections, a single run with our MKL already leads to better performances in both metrics. The rightmost column in the table shows the result for SSAD using a sum of the input kernels. Apparently, the performance is worse than using a bag-of-words kernel alone. This result underlines the necessity of effective feature selection techniques for

[1] https://archive.ics.uci.edu/ml/datasets/Reuter_50_50

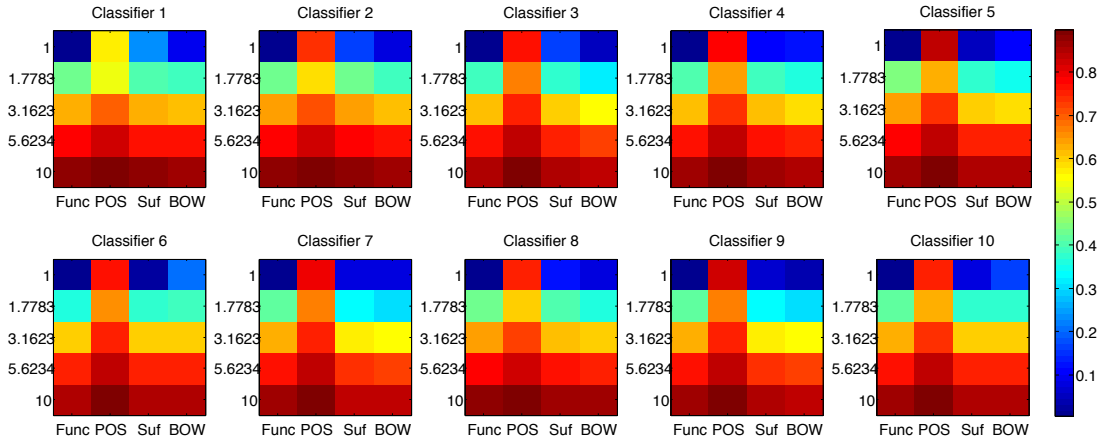

Figure 2: Kernel mixture coefficients for the 10 classes

authorship attribution. Note that our method can actually be viewed as an ensemble method that combines several models as shown in Equation (1). However, compared to traditional ensembles, our method uses a convex combination and hence returns the optimal ensemble given the data.

Table 1: F-scores for the subset of Reuters 50-50

	p-norm MKL					SSAD				
	1	1.7783	3.1623	5.6234	10	func-w	POS	Suffix-3	BOW	$\sum$
F_{micro}	73.46	73.08	73.84	73.89	74.23	63.08	54.62	70.01	72.85	61.76
F_{macro}	79.23	78.86	79.63	79.76	80.07	68.66	58.03	74.01	78.09	70.93

Figure 2 visualises the resulting mixing coefficients for the 10 authors/classifiers. While the models are very similar at first sight, small deviations indicate differences in the style of the authors. Consider for instance the top-left matrix. The contribution of the part-of-speech tag kernel (second column) to the final solution is less than for the other authors. By contrast, the importance of the Suffix-3 kernel has (slightly) more impact than for the remaining authors. This result shows that author-dependent mixtures are found that help to capture characteristic traits of the respective writing styles.

4.2 Revisiting the Federalist Papers

The Federalist Papers are a series of 85 articles and essays written during 1787–1788. They were published anonymously to persuade the citizens of the State of New York to ratify the Constitution. Later, these papers were credited to Alexander Hamilton, John Jay, and James Madison; 73 of the documents are uniquely associated with one of the three authors while the remaining 12, also known as the disputed papers, have been claimed by both, Hamilton and Madison. Three of the 73 articles are considered joint work by Hamilton and Madison. Previous studies often assign all 12 disputed papers to Madison which we assume as ground-truth in the remainder (Mosteller and Wallace, 1964; Fung, 2003).

To confirm or refuse these previous findings, we conduct an experiment using same four kernels as in the previous section, that is, a function words kernel (Koppel and Schler, 2003), a part-of-speech (POS) tag kernel, a Suffix-3 kernel, and a bag-of-words (BOW) kernel. We compare the performance of our approach (MKL) with semi-supervised anomaly detection (SSAD) (Görnitz et al., 2013), support vector data description (SVDD) (Tax and Duin, 1999), and the one-class SVM (OCSVM) (Schölkopf et al., 1999). As before, the baselines cannot use all kernels at a time and are evaluated on every kernel separately. For simplicity, we show only the MKL results for parameter $p = 2$ as all other p-norms that we tried out lead to the same result.

We randomly divide the undisputed papers into training (80%) and holdout (20%) and use the 12 disputed papers for testing. We make sure that training sets contain at least three examples of every author and two articles written jointly by Hamilton and Madison. Otherwise we discard and draw again. We repeat experiments five times with randomly drawn training and holdout sets and report on averaged accuracies for the disputed test set.

Table 2: Results for the disputed articles of the Federalist papers.

	kernel	H&M	M	J	H
MKL	(all)	0	**12**	0	0
SSAD	484fw	0	**12**	0	0
	POS	9	0	3	0
	Suffix3	0	**12**	0	0
	BoW	0	0	0	12
SVDD	484fw	12	0	0	0
	POS	12	0	0	0
	Suffix3	12	0	0	0
	BoW	12	0	0	0
OCSVM	484fw	12	0	0	0
	POS	12	0	0	0
	Suffix3	12	0	0	0
	BoW	12	0	0	0

The results are shown in Table 2. The one-class SVM and SVDD constantly credit the 12 disputed articles as joint work by Hamilton and Madison. The outcome of SSAD highly depends on the kernel function; while the part-of-speech kernel distributes the papers on Jay (3) and Hamilton and Madison (9), respectively, the bag-of-words kernel assigns all documents to Hamilton. By contrast, SVDD with function word and Suffix-3 kernels attribute the articles to Madison. The same outcome is observed for our novel MKL that also credits the 12 papers to Madison. Thus, MKL and SSAD with function words and BoW kernel confirm todays assumption that all 12 papers have been written by Madison. However, choosing SSAD as the base classifier in the absence of prior knowledge leaves much room for interpretations and the user in the need of deciding between three solutions, depending on which kernel she prefers. By using our MKL, selecting features and or kernel functions is no longer necessary as the learning algorithm itself picks the right combination of kernels for the problem at hand. Thus, the more kernels are thrown into, the richer the decision space for the MKL.

5 Conclusion

We proposed a universal method for authorship detection. Our approach built upon semi-supervised anomaly detection and generalised existing techniques to utilise multiple kernels; a requirement which is particularly beneficial for authorship attribution as features are usually tailored to tasks at hand and do not necessarily translate well to other authors. Our method is proven to converge to the optimal solution and simple to implement. Our empirical results show the robustness of our approach as it consistently outperforms baseline competitors on a subset of Reuters 50-50 or confirms common knowledge wrt the authorship of disputed articles of the Federalist Papers. The main advantage of the method however lies in its simplicity. Practitioners do not need to take critical design choices in terms of which features to use and which not. By contrast, all features (kernels) can be used in the optimisation and the method itself finds the optimal combination for the problem at hand.

Acknowledgements

Jamal Abdul Nasir is supported by a grant from the Higher Education Commission, H-9 Islamabad, Pakistan. Nico Görnitz is supported by the German Bundesministerium für Bildung und Forschung

(BMBF FKZ 01GQ0850 and 01IB001A). Ulf Brefeld is also affiliated with the German Institute for Educational Research (DIPF), Frankfurt/Main, Germany.

References

F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. 2004. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proc. of the International Conference on Machine Learning*.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. 2010. Semi-Supervised Novelty Detection. *Journal of Machine Learning Research*, pages 2973–2973–3009–3009, December.

J. F. Burrows. 1987. *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. 2006. *Semi-supervised learning*. {MIT} Press.

J. Diederich, J. Kindermann, E. Leopold, and G. Paass. 2003. Authorship attribution with support vector machines. *Applied Intelligence*, 19(1):109–123.

G. Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305.

Glenn Fung. 2003. The disputed federalist papers: Svm feature selection via concave minimization. In *Richard Tapia Celebration of Diversity in Computing Conference*.

Nico Görnitz, Marius Kloft, and Ulf Brefeld. 2009. Active and semi-supervised data domain description. In *ECML*, pages 407–422. Springer.

Nico Görnitz, Marius Kloft, Konrad Rieck, and Ulf Brefeld. 2013. Toward Supervised Anomaly Detection. *Journal of Artificial Intelligence Research (JAIR)*, 46:235–262.

David I. Holmes and Daniel W. Crofts. 2010. The diary of a public man: a case study in traditional and non-traditional authorship attribution. *LLC*, 25(2):179–197.

David I. Holmes. 1998. The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, 13(3):111–117, September.

J. Houvardas and E. Stamatatos. 2006. N-gram feature selection for authorship identification. In *AIMSA*.

Dmitry V. Khmelev. 2000. Disputed authorship resolution through using relative empirical entropy for markov chains of letters in human language texts. *Journal of Quantitative Linguistics*, 7(3):201–207.

M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. 2011. lp-Norm Multiple Kernel Learning. *JMLR*, 12:953–997.

M. Koppel and J. Schler. 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of the International Joint Conference on Artificial Ingelligence*.

M. Koppel, N. Akiva, and I. Dagan. 2006. Feature instability as a criterion for selecting potential style markers: Special topic section on computational analysis of style. *Journal of the American Society for Information Science and Technology*, 57(11):1519–1525.

M. Koppel, J. Schler, and S. Argamon. 2011. Authorship attribution in the wild. *Language Resources & Evaluation*, 45(1):83–94.

G. Lanckriet, N. Cristianini, L. E. Ghaoui, P. Bartlett, and M. I. Jordan. 2004. Learning the kernel matrix with semi-definite programming. *JMLR*, 5:27–72.

J. Li, R. Zheng, and H. Chen. 2006. From fingerprint to writeprint. *Communications of the ACM*, 49(4):76–82.

Hermann Maurer, Frank Kappe, and Bilal Zaka. 2006. Plagiarism – a survey. *Journal of Universal Computer Science*, 12(8)8:1050–1084.

T. C. Mendenhall. 1887. The characteristic curves of composition. *Science*, ns-9(214S):237–246.

Frederick Mosteller and David L. Wallace. 1964. Inference and Disputed Authorship: The Federalist. Springer-Verlag, New York. 2nd Edition appeared in 1984 and was called *Applied Bayesian and Classical Inference*.

Klaus-Robert Müller, Sebastian Mika, Gunnar Rätsch, Koji Tsuda, and Bernhard Schölkopf. 2001. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201.

Antonio Neme, Blanca Lugo, and Alejandra Cervera. 2011. Authorship attribution as a case of anomaly detection: A neural network model. *Int. J. Hybrid Intell. Syst.*, 8(4):225–235.

S. Raghavan, A. Kovashka, and R. Mooney. 2010. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 Conference*.

Alain Rakotomamonjy, Francis R. Bach, Stephan Canu, and Yves Grandvalet. 2008. SimpleMKL. *JMLR*, 9:2491–2521.

Joseph Rudman. 1997. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31(4):351–365.

B Schölkopf, J C Platt, J Shawe-Taylor, a J Smola, and R C Williamson. 1999. Estimating the support of a high-dimensional distribution. Technical report, July.

Y. Seroussi, I. Zukerman, and F. Bohnert. 2011. Authorship attribution with latent dirichlet allocation. In *Proceedings of the 15th International Conference on Computational Natural Language Learning*.

Efstathios Stamatatos, Nikos Fakotakis, and George K. Kokkinakis. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. In *Computers and the Humanities*.

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

Enhua Tan, Lei Guo, Songqing Chen, Xiaodong Zhang, and Yihong (Eric) Zhao. 2013. Unik: Unsupervised social network spam detection. In *Proceedings of CIKM*.

D. Tax and R. Duin. 1999. Data domain description using support vectors. In *Proceedings of the European Symposium on Artificial Neural Networks*, volume 256, pages 251–256. Citeseer.

G. Udnv Yule. 1944. *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

G. K. Zipf. 1932. *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press.