単独剽窃検知
—所与の一文書だけから剽窃箇所を推定する–

溝江 将　　田中久美子
東京大学工学部計数工学科　東京大学大学院情報理工学系研究科

1　はじめに

情報技術の発展により、インターネット上からの情報の検索、取得が容易になった。それに伴い、取得した情報を出典を明示せずに不正に利用すること、すなわち、剽窃行為が学術レポート等で大きな問題となっている。

一般的に、剽窃を検出する際には、「種々のデータベースに対して検索を行う」ということが行われている。この方法で剽窃の証拠を見つけるには、クエリを作るためにまず剽窃の疑われる部分を発見しなければならない。この疑わしい部分を選別する作業は、多くの場合人手で行われている。この時人間が行っていることは、調査対象として与えられたただ一つの文書からその文書中に存在する剽窃箇所を推定することであり、本稿ではこれを単独剽窃検知と呼ぶ。

本稿では、まず既存の単独剽窃検知を行う手法を紹介し、次に one-class SVM を用いて単独剽窃検知を行う手法と、その手法を用いて既存の手法を改良した手法を提案する。最後にこれら三つの手法を比較し、その得失を論じる。

2　今までの単独剽窃検知の手法

単独剽窃検知 (intrinsic plagiarism detection) という問題は、Eissen et al. [2, 3] により始めて導入された。この問題は、 Stein et al. [9] の一部を要約して、以下のように定義できる。

> 単独剽窃検知は、通常より広い意味での著者確認問題と見ることができる。すなわち、
> 1. ただ一つの文書が与えられ、
> 2. 文書中で想定される著者以外の者が書いたことが疑われる部分を見つけ出す。
>
> という問題である。

この問題を計算機を用いて解決することは、 Stein et al. [10, 9]、Seaward et al. [7]、Stamatatos [8]、Zechner et al. [11] などで研究されている。

ここでは特に、最も新しく、上記の手法の中でも比較的性能の高い Stein et al. [9] の手法の詳細に触れる。

3　Stein の手法

3.1　手法の詳細

剽窃部分の占める割合が 50% 未満である文書 d が与えられた時、これを 5000 字ずつの部分文書にわけ、これらを $\{s_1, s_2, \ldots, s_n\}$ とする。それぞれの部分文書に対応する特徴量ベクトルを $\mathbf{s}_k = (x_1, x_2, \ldots, x_m)$ とする。

ある部分文書 s について、それが剽窃を含む部分である事象を S^o、それが剽窃を含まない部分である事象を S^t とする。このとき、s が剽窃を含む部分であるかについての仮説を最大事後確率推定で決める。すなわち、s に関する仮説 H を

$$H = \underset{S \in \{S^t, S^o\}}{\operatorname{argmax}}\ p(S) \prod_{i=1}^{m} \frac{p(x_i|S)}{p(x_i)}, \tag{1}$$

として決定する。ただし、各々の特徴量 x_i の条件付き確率密度関数 $p(x_i|S)$ を

$$p(x_i|S^o) = \begin{cases} \frac{1}{6\hat{\sigma}_i} & |x_i - \hat{\mu}_i| < 3\hat{\sigma}_i \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

$$p(x_i|S^t) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_i} \exp\left(-\frac{(x_i - \hat{\mu}_i)^2}{2\hat{\sigma}_i^2}\right), \tag{3}$$

とする。ここにおいて、$\hat{\mu}_i, \hat{\sigma}_i$ はそれぞれ推定された平均値、標準偏差である。

注意すべきこととして、式 (1) において、$\hat{\sigma}_i < |x_i - \hat{\mu}_i| < 2\hat{\sigma}_i$ となる特徴量についての積はとらない。

また、この手法の追試の際には、式 (1) において、推定された分散 $\hat{\sigma}_i = 0$ となった特徴量についても積を取らなかった。加えて、式 (2) において、$|x_i - \hat{\mu}_i| \geq 3\hat{\sigma}_i$ であっても、$p(x_i|S^o) = \frac{1}{6\hat{\sigma}_i}$ とした。これらは、尤度の計算を容易にするためである。

3.2 Stein の手法の問題点

Stein の手法の問題点としては、剽窃を含む部分文書の特徴量が一様分布する、という仮定をしていることが挙げられる。例えば、文書の想定される著者が他の一つの文書から剽窃を行ったとする。この時、一般の著者のスタイルが正規分布にしたがうという仮定の下では、剽窃箇所の特徴量は正規分布に従い、一様分布しない。

また、剽窃が多い文書については、推定された平均 $\hat{\mu}_i$ が、非剽窃部分の真の平均 μ_i から大きく外れることが予想される。すなわち、Stein の手法は剽窃が多い文書に対して適切な判断を行うことができないと考えられる。

4 提案手法

提案手法では、複数の one-class SVM に投票をさせることで剽窃箇所を発見することを考える。one-class SVM を使用する動機としては、この問題が 1 クラス分類問題であることが挙げられる。すなわち、この問題は、剽窃した部分を含むある文書について、その文書の想定される著者の書いた部分とその他の著者の書いた部分を分ける問題である、ということである。

以下では、まず one-class SVM の紹介をし、次に提案手法の手順について述べる。最後に、提案手法を用いて Stein の手法を改良する方法について述べる。

4.1 one-class SVM

one-class SVM は Schölkopf et al. [6] により開発された手法であり、与えられた訓練データの密度が濃い部分を検出する。

具体的には、以下のようになる。与えられた訓練データ集合を $T = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l\}$ とする。特徴写像 $\Phi : T \to F$ を、与えられたカーネル $k(\mathbf{x}, \mathbf{y})$ によって内積が計算される特徴空間 F へとデータを写す写像とする。すなわち、

$$k(\mathbf{x}, \mathbf{y}) = (\Phi(\mathbf{x}), \Phi(\mathbf{y})).$$

与えられた訓練データを原点から分離する平面を求めるため、以下の二次計画問題を解く。

$$\min_{w \in F, \boldsymbol{\xi} \in \mathbb{R}, \rho \in \mathbb{R}} \frac{1}{2}\|w\|^2 + \frac{1}{\nu l}\sum_i \xi_i - \rho, \tag{4}$$

$$\text{subject to } (w \cdot \Phi(\mathbf{x}_i)) \geq \rho - \xi_i,\ \xi_i \geq 0. \tag{5}$$

ここで、$\nu \in (0, 1)$ はパラメータであり、ξ_i はスラック変数である。ν の値を 0 に近づけるほど、ペナルティである $\sum_i \xi_i$ の重みが大きくなり、ペナルティの増大を回避するハードマージンの問題になっていく。

4.2 手順

Stein の手法と同様に $d, s_k, \mathbf{s}_k$ を定義する。また、$T_d = \{\mathbf{s}_k\}_{k=1}^n$ とする。

1. まず、T_d からブートストラップ法により、C 個の訓練データ $T_1, T_2, \ldots, T_C$ を作る。
2. 次に、それぞれの T_i を用い、パラメータを $\nu = \nu_0$ に固定して one-class SVM を C 個訓練する。
3. それぞれの one-class SVM が、T_d 内のデータ $\mathbf{s}_k$ に対して、それぞれが剽窃部分であるかの判定を行う。すなわち、おのおのの SVM は、自身が非剽窃部分と判定したデータに対して 1 票を投ずる。
4. 投票結果を確認し、あるしきい値 v_0 票以下しか獲得獲得できなかった部分文書を剽窃部分と判断する。

剽窃を含むと判断するしきい値 v_0 は、提案手法を適用して出る投票の結果の度数分布から以下のように決める。

1. 文書 d に対し提案手法を適用し、$\{s_1, \ldots, s_n\}$ に対する投票結果として、$V = \{v_1, \ldots, v_n\}$ を得る。
2. v_i の度数分布から、p パーセンタイルの点の値を返す関数 q_p を用いて、v_0 を $v_0 = q_p(V)$ とする。

p の値としては、$0, 10, \ldots, 50$ などが考えられる。ただし、q_0 は、最小値を返すものとする。

提案手法の拠り所は、剽窃部分が d に占める割合が半分以下である、という仮定である。この仮定により、ブートストラッピングをして投票することで、剽窃を含まない部分に対応するベクトルが正しく判定されると考えた。

4.3 提案手法を用いた Stein の手法の改良

第三の手法として、上記二つの手法を組み合わせた以下のようなものを考える。

1. Stein の手法を用いて、剽窃を含む部分とそうでない部分とに分ける。
2. 1 で剽窃を含まないと判断された部分を対象に、one-class SVM + 投票の手法を適用する。
3. 1、2 で剽窃を含むと判断された部分を全体の結果として返す。

これは、 Stein の手法を適用した結果、剽窃を含まないと判断された部分が、実際には剽窃箇所を含んでいることから、その剽窃箇所を one-class SVM を用いて取り出すことで、適合率、再現率の上昇を図ろうとするものである。

5 実験

以下で、 Stein の手法と提案手法の比較を行う。まず、実験で用いた特徴量とツール、データについて説明し、その後に実験結果を載せる。

5.1 実験で用いた特徴量とツール、データ

まず、特徴量について説明する。特徴量として、以下のものを用いた。

- 品詞 3-グラム
- 文字 3-グラム
- 単語の平均長
- 単語頻度
- 単語頻度クラスの平均
- 読みやすさの指標 (Flesch-Kincaid など)

特徴量は、 Stein et al. [9] の中で、著者を判別するのに効果的な特徴量として挙げられているものを参考に選択した。

次に実験で用いたツールについて説明する。one-class SVM として、LIBSVM [1] を用いた。また、品詞タグ付けに TreeTagger [5] を用いた。

最後に、実験で用いたデータについて説明する。剽窃の含まれるコーパスとして、PAN Plagiarism Corpus 2010 [4] を用いた。このコーパスは、剽窃の含まれる文書を人工的に作ったものであり、各々の文書にどれだけの剽窃箇所が含まれているか、などのデータも含まれている。

実験に用いる文書集合として、このコーパスの $0 < \theta < 0.5$ の部分をそのまま使用するつもりであった。しかし、この文書集合は、文書中に剽窃部分の占める割合 を θ としたとき、θ が小さいほど文書数が多くなるというアンバランスなものであった。そこで、精確な適合率、再現率を計算するためにコーパスの文書を以下のように処理して、どの θ に対しても、その θ を持つ文書の数ができるだけ等しくなるようにした。

1. まず、長さ 500000 字以上で、剽窃部分の占める割合 θ が $0 < \theta < 0.65$ である文書を抜き出した。
2. 1 で得られた文書の剽窃を含む部分あるいは剽窃を含まない部分を取り除くことで、$0 < \theta < 0.5$ であり、且つ長さが 50000 文字以上であるような文書を作る。

結果として、4755 文書ができた。出来上がった文書集合の θ に対する文書数の度数分布を図 1 に示す。

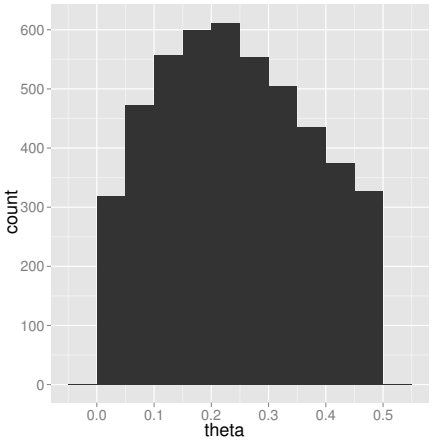

図 1 θ に対する文書数の度数分布。θ の範囲を幅 0.05 毎に区切り、頻度を取った。

5.2 実験結果

この文書集合に対し、上記の 3 つの手法を適用した結果を図 2、3 に示す。ただし、one-class SVM を用いた手法については、SVM の数 $C = 30$ である。また、パラメータ ν と判断のしきい値を決める関数 q_p は、最も性能がよいと思われる値を用いており、図にはその時の結果のみを載せてある。

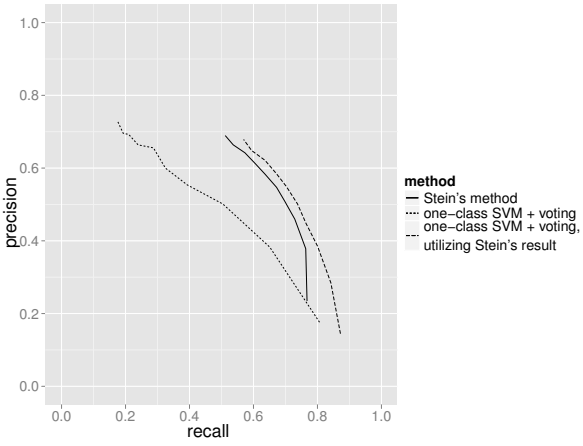

図 2 それぞれの手法についての適合率-再現率曲線。one-class SVM を用いる手法については、SVM の個数 $C = 30$、$\nu = 0.5$、そして、剽窃と判断するしきい値 v_0 は度数分布の最小値である。(すなわち、q_p の $p = 0$ である。) 曲線は θ の変化によるものである。

まず、図 2 を観察して気づくこととして、以下のこ

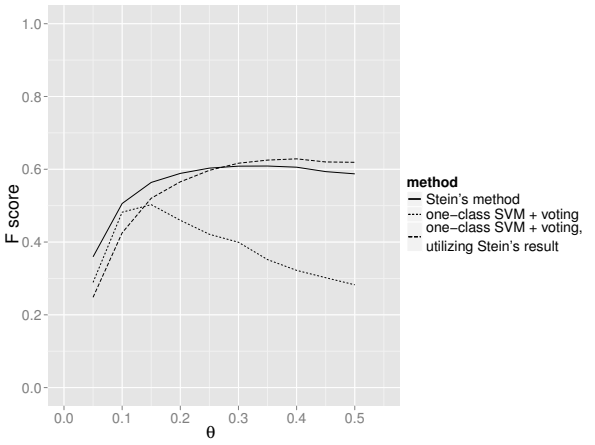

図 3　それぞれの手法についての F 値と剽窃部分の文書に占める割合 θ の関係。one-class SVM を用いる手法については、SVM の個数 $C = 30$、$\nu = 0.5$、そして、剽窃と判断するしきい値 v_0 は度数分布の最小値である。(すなわち、q_p の $p = 0$ である。)

とが挙げられる。

- one-class SVM のみを用いた手法は、Stein の手法にくらべ、適合率が高い。これは、投票により、短い剽窃箇所がきちんと外れ値として判断されているためと考えられる。一方で長い剽窃箇所があると、その剽窃箇所からできるデータ点の集合は、特徴空間で密な部分を構成する。結果として、one-class SVM では外れ値として判断されず、それが低い再現率の原因となっていると考えられる。
- Stein の手法と Stein の手法の結果を利用した提案手法の性能を比べると、Stein の手法の方が、適合率は高く、再現率は低い。これは、Stein の手法を適用して剽窃を含まないと判断された文書集合が、実際には剽窃を含まないか、剽窃を含んでいたとしてもその量はとても少ないということであると考えられる。

次に図 3 を観察する。F 値で比べると、剽窃が多い文書では、Stein の手法と提案手法を組み合わせたものの性能が良いことが分かる。これは、第 3 節で述べたとおり、Stein の手法の性能が悪化し、その性能低下を提案手法が補った、と考えられる。しかし、全体としては、既存手法の性能が安定しており、一概にはどれがよいとは言えない。

6　おわりに

本稿では、既存の単独剽窃検知の手法の追試と新たな手法の実験を行った。結果的に、既存の手法の性能は、文書中に剽窃の占める割合によらず安定していることが確認された。また、少しでも剽窃箇所を見つけることが重要である、すなわち、適合率が重要であると考えれば、剽窃の占める割合が大きいときに、提案手法は既存手法に勝っている。しかしながら、本稿で提示したどの手法についても、その性能は満足のいくものではない。

単独剽窃検知という問題は、剽窃検出の自動化において避けては通れない問題であり、より一層性能の良い手法が望まれる。

参考文献

[1] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[2] S. Eissen and B. Stein. Intrinsic plagiarism detection. *Advances in Information Retrieval*, pp. 565–569, 2006.

[3] S. Meyer zu Eissen, B. Stein, and M. Kulig. Plagiarism detection without reference collections. *Advances in data analysis*, pp. 359–366, 2007.

[4] Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. An Evaluation Framework for Plagiarism Detection. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, August 2010. Association for Computational Linguistics.

[5] Helmut Schmid. *TreeTagger: a language independent part-of-speech tagger*, 1994. Software available at `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`.

[6] B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, Vol. 13, No. 7, pp. 1443–1471, 2001.

[7] L. Seaward and S. Matwin. Intrinsic Plagiarism Detection using Complexity Analysis. In *3rd PAN WORKSHOP. UNCOVERING PLAGIARISM, AUTHORSHIP AND SOCIAL SOFTWARE MISUSE*, pp. 56–61, 2009.

[8] E. Stamatatos. Intrinsic Plagiarism Detection Using Character n-gram Profiles. In *3rd PAN WORKSHOP. UNCOVERING PLAGIARISM, AUTHORSHIP AND SOCIAL SOFTWARE MISUSE*, Vol. 2, p. 38, 2009.

[9] B. Stein, N. Lipka, and P. Prettenhofer. Intrinsic plagiarism analysis. *Language Resources and Evaluation*, pp. 1–20, 2010.

[10] B. Stein and S.M. zu Eissen. Intrinsic plagiarism analysis with meta learning. In *Proceedings of the SIGIR Workshop on Plagiarism Analysis, Authorship Attribution, and Near-Duplicate Detection*, pp. 45–50. Citeseer, 2007.

[11] M. Zechner, M. Muhr, R. Kern, and M. Granitzer. External and intrinsic plagiarism detection using vector space models. In *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, pp. 47–55, 2009.