6S-01

ファイル情報の可視化による分類法の検討*

大谷 康介† 大沢 泰貴‡ 松田 健‡
†静岡理工科大学総合情報学部 ‡静岡理工科大学大学院システム工学専攻

1 はじめに

コンピュータ・フォレンジクスはコンピュータ上のデータを解析し, 不正アクセス等の犯罪の法的な証拠性を明らかにする等に使われている技術だが, データ量が膨大になると同時に解析にかかるコストも膨大となってしまう. そのコストを軽減するために応用が期待されている技術として目 Grep がある. 目 Grep はファイルのバイナリデータを可視化して目視のみで特徴を捉えることができるため, その特徴を抽出し解析に応用可能であれば, 解析にかかるコストを軽減できると考えられる. 本研究では, 可視化された PDF ファイルのバイナリデータから特徴を抽出し, 複数の種類のファイルに自動分類するアルゴリズムを提案する. 実際に複数の構成要素が異なるテスト用データを用意し, 提案アルゴリズムにより分類可能であるかについて検証する.

2 バイナリデータによる可視化

コンピュータ上のファイルのバイナリデータを色づけして可視化することにより, 目視で特徴を捉える事を可能とする技術を目 Grep と呼ぶ. 目 Grep で捉える事ができる特徴例としては, ファイルの文字コードや圧縮形式, 種類等があげられる. 実際に BZ という Binary Editor を用いて画像化した PDF ファイルの一部を図 1 から図 5 に示す.[1]. 5 つの図から明らかになるように PDF ファイルでも中身の構成要素によって異なるパターンに可視化されることがわかる. 本研究では,pdf ファイルを収集・生成し, 図 1 から図 5 のようにそれらのファイルを (1) 文章のみ,(2) 文章+画像,(3) 画像のみ,(4)web サイトへリンクを含むもの,(5) スキャンの 5 種類のファイルに分類し, これらのデータの特徴抽出を行うことでファイルの分類が可能であるかどうか検討を試みる.

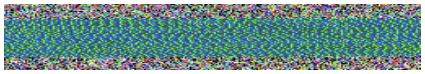

図 1: 文章のみ ($j = 1$)

図 2: 文章+画像 ($j = 2$)

図 3: 画像のみ ($j = 3$)

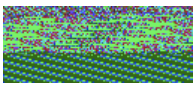

図 4: リンクあり ($j = 4$)

図 5: スキャン ($j = 5$)

3 分類アルゴリズム

本稿では, 前章で分類した 5 種類のファイルの構成要素からそれぞれのファイルの特徴を抽出し, その特徴を用いてファイルの自動分類を行うためのアルゴリズムについて考える. まず学習用データとなるファイルの特徴を抽出するため, ファイルの情報を 16 進数に変換して特徴ベクトルを作成し, 特徴ベクトルを表現するような曲線を推定する. ファイル分類のフェーズでは, 未知データが推定した曲線にどの程度当てはまるのかということを調べて分類を行う. 本研究の提案アルゴリズムを以下に示す.

1. l 番目のデータを 16 進バイナリとし, 数 $n(1 \leq n \leq 256)$ の頻度 $s_{n,l}$ を要素としたベクトルを $\mathbf{s}_l$ とする.
$$\mathbf{s}_l = (s_{1,l}, s_{2,l}, \cdots, s_{256,l})$$
2. ベクトル $\mathbf{s}_l$ からベクトル $\mathbf{y}_l$ を以下のように与える.
$$\mathbf{y}_l = \{y_{1,l}, y_{2,l}, \cdots, y_{16,l}\},\ y_{k,l} = \Sigma_{t=1}^{16} s_{16k-16+t,l}$$
3. データの種類 i ごとに, ベクトル $\mathbf{y}^{(i)}$ を計算する.
$$\mathbf{y}^{(i)} = \{y_1^{(i)}, y_2^{(i)}, \cdots, y_{16}^{(i)}\},\ y_k^{(i)} = \frac{1}{10}\Sigma_{l=10i-9}^{10i} y_{k,l}$$
4. 関数 $F(x_k) = y_k^{(i)}(x_k = -\pi + \frac{2\pi}{15}(k-1))$ を以下の関数 $f(x)$ で近似する.
$$f(x_k) = \frac{a_0}{2} + \Sigma_{m=1}^{4}(a_m \cos mx_k + b_m \sin mx_k)$$
$$a_0 = \frac{1}{8}\Sigma_{k=1}^{16} F^{(i)}(x_k),\ a_{m'} = \frac{1}{8}\Sigma_{k=1}^{16} F(x_k)\cos m'x_k$$
$$b_{m'} = \frac{1}{8}\Sigma_{k=1}^{16} F(x_k)\sin m'x_k$$

*A Study of Classification Method by Visualizing Data.
†Kousuke Ootani, Shizuoka Institute of Science and Technology
‡Taiki Oosawa, gs14003@ym.sist.ac.jp, Takeshi Matsuda, Shizuoka Institute of Science and Technology, tmatsuda@cs.sist.ac.jp

5. どのクラス ($j = 1,2,3,4,5$) に属するかは未知であるテストデータ $((\hat{x})_k, (\hat{y})_k) = (k, (\hat{y})_k) \quad (k = 1,2,\cdots,16)$ を以下で判別する.

6. $X_{\hat{j}}$ が最小のとき, テストデータはクラス $\hat{j}(\hat{j} \in 1,2,3,4,5)$ に属するものと判断する.

$$X_j = \sum_{k=4}^{13} [f_j(x_k) - (\hat{y})_k]^2$$

7. $X'_{\hat{j}}$ が m 以上であるとき , テストデータはクラス $\hat{j}$ に属するものと判断する.

$$X'_{\hat{j}} = \sum_{k=e_1}^{e_r} L^{(\hat{j})}(l,k)$$

$$L^{(\hat{j})}(l,k) = \begin{cases} 1 & fD^- < y_{k,l} < fD^+ \\ 0 & otherwise \end{cases}$$

ただし、$fD^- = f^{(\hat{j})}(x_k) - 2 \times D^{(\hat{j})}(k), fD^+ = f^{(\hat{j})}(x_k) + 2 \times D^{(\hat{j})}(k), \quad D^{(\hat{j})}(k) = \sqrt{\frac{1}{10}\Sigma_{l=10\hat{j}-9}^{10\hat{j}}(y_{k,l} - y_k^{(\hat{j})})^2}$ である.

4 実験と考察

この章では，前章で提案した pdf ファイルの分類アルゴリズムを，2 章で定義した 5 つのどのファイルか未知である 50 個 (5 種類のファイルを 10 個ずつ) の pdf ファイルに適用し，特徴抽出や分類アルゴリズムの精度について考察を行う．表 1 は，用意したテスト用の未知データ 50 個に対してアルゴリズムを適用したときの結果である.

表 1: アルゴリズム 6 を適用した結果

判別/正解	1	2	3	4	5
1	9	10	4	7	0
2	0	0	0	0	0
3	1	0	5	3	0
4	0	0	0	0	0
5	0	0	1	0	10

表は、横を正解、縦を判別として該当するデータの数を示している．すなわち対角要素が正解のデータ数となっている．文書データ, スキャンデータは良く判別できており、文章+画像データは文章データと誤判別している．画像データはおよそ半分は文章データと誤判別しているが、半分は画像データと正しく判別している．リンクを含んでいるデータはすべて誤判別であった.

次の図 6 に示すとおり，例えば画像のみで構成されているデータでは，特徴空間上でも 2 通りのパターンの特徴の傾向を持つことがわかり，このようなデータが誤分類を導いているものと考えられる．しかしながら，

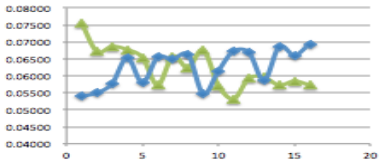

図 6: 2 つの画像データの複合グラフ

スキャンデータとその他のクラスでは直感的にも特徴区間に異なるパターンが見られるため，学習用データに対して，スキャンデータとその他の種類のファイルのデータの二値分類を行う目的で $e_1 = 1, e_2 = 4, e_3 = 8$ として提案アルゴリズムを適用したところ，表 2 のような結果が得られた.

表 2: アルゴリズム 7 を適用した結果

m	スキャン以外のデータ	スキャンのデータ
0	38	0
1	2	2
2	0	6
3	0	2

代数統計に基づく統計的検定 (分割表に対する独立性検定) を行ったところ，p 値は $p = 0.000089$ となり，文章もしくはスキャンデータから構成される PDF ファイルの特徴量には, 統計的に有意な差がみられることも確認することができた．したがって，pdf ファイルをスキャンデータで構成されているものと，それ以外のデータで構成されているものに分類するという場合に限定すれば，提案手法は有効的であると言える.

5 まとめ

本研究では, PDF ファイルのバイナリデータに色づけすることで可視化したデータの特徴を抽出し，PDF ファイルを構成する要素で PDF をいくつかの種類のファイルに分類する手法について検討した．PDF ファイルを 5 種類のクラスに分割してそれぞれのクラスに含まれるデータを調べた結果，それぞれのクラスの中でいくつかの特徴パターンが存在することが分かった．今回の検討結果からより正確にファイルを分類する手法を確立して PDF 以外のファイル分類も可能にし，コンピュータフォレンジックの分野への応用を試みることが今後の課題である.

参考文献

[1] Binary Editor BZ, http://www.vcraft.jp/soft/bz.html, (2004)

[2] Persi Diaconis and Bernd Sturmfels, “Algebraic algorithms for sampling from conditional distributions”, The Annals of Statistics, Vol. 26, No. 1, pp363-397, (1998)