6S-02

ファイル情報の可視化による特徴抽出の検討*

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

1 はじめに

コンピュータ上に存在する数多くのファイルを可視化し, ある特徴を持つファイルを視覚的に探索する手法の一つに, 目 Grep がある. コンピュータ・フォレンジックスといった, コンピュータサイエンスの分野で注目をあびているが, 可視化された画像データから経験的に特徴を読み取る必要があるため, 誰しもが容易に応用できるとは言い難い. しかし目 Grep はファイルが持つ数値データを画像化するため, 分析対象のファイルを, 可視化した画像データと紐づけて定量的に解析することが可能であると考えられる. 本研究では, ファイルのバイナリデータから特徴を抽出し, その抽出した特徴量に基づいて分割表を生成するアルゴリズムを提案する. 実験用のサンプルデータとして, 構成要素が異なる PDF ファイルを複数用意し, 提案アルゴリズムから生成された分割表を代数統計に基づく統計的検定法により解析することで, 抽出した特徴量に差異が存在するかを検証する.

2 目 Grep

目 Grep とは, コンピュータ上に存在するファイルの情報を可視化することで, ファイル解析に役立てるための技術である. 分析対象のファイルのバイナリデータを色情報として捉え, 元のデータをビットマップイメージとして画像化することにより, ファイルの可視化を実現している. ファイルの圧縮形式や種類などの情報が, 目 Grep により可視化された画像へ特徴として表れる場合があり, この技術の熟練者はその画像を目視するだけで, 元のファイルの情報や働きを認識することができるという. 実際に可視化した PDF ファイルの一部を図 1 として示す. 尚, 分析ツールとして Binary Editor(BZ) を使用した [1].

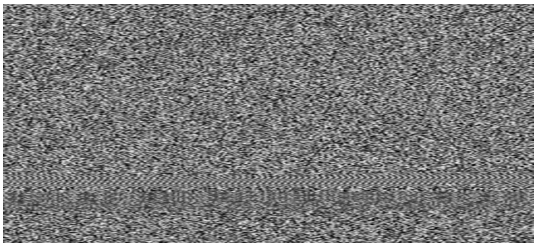

図 1: PDF ファイルの可視化

3 マルコフ基底を用いた代数統計的検定法

本研究の特徴抽出アルゴリズムから生成される分割表はスパースであるような場合が多く, 漸近理論に基づいた統計的検定法は当てはまりが悪い可能性があり, また Fisher の正確検定も計算量的に困難性である場合がある. 本研究では, マルコフ基底とマルコフ連鎖モンテカルロ法を組み合わせた代数統計に基づく統計的検定法から, 解析結果の妥当性を検討する [2]. この手法は, 対象とする分割表と周辺頻度を共有する分割表を超幾何分布に従ってランダムサンプリングすることで近似的に p 値を算出する手法であり, スパース, もしくはサンプル数が少ないような分割表の分析に威力を発揮する. マルコフ基底は周辺頻度を共有する分割表を連結にするような働きをもつベクトルの集合であり, 分割表を多項式に対応付けることで, その多項式環上のイデアルの生成系として与えられる. 本研究で分析する二元分割表のマルコフ基底の構造は比較的シンプルであり, すでにその具体形は知られている.

4 特徴抽出アルゴリズム

本特徴抽出アルゴリズムは 2 つの phase に区分される. phase.1 はファイルの種類ごとにその特徴を表現するような曲線を推定し, phase.2 は phase.1 で推定された曲線から分割表を生成する. それぞれのアルゴリズムを以下に示す.

Phase.1

*An Examination of Feature Extraction by Visualizing Data.

†Taiki Oosawa, Shizuoka Institute of Science and Technology, gs14003@ym.sist.ac.jp
‡Kousuke Ootani, Takeshi Matsuda, Shizuoka Institute of Science and Technology, tmatsuda@cs.sist.ac.jp

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 $s_{n,l}$ を集めた集合を S_l とする.
$$S_l = \{s_{1,l}, s_{2,l}, \cdots, s_{256,l}\}$$

2. 集合 S_l から集合 Y_l を以下のように与える.
$Y_l = \{y_{1,l}, y_{2,l}, \cdots, y_{16,l}\},\ y_{k,l} = \Sigma_{j=1}^{16} s_{16k-16+j,l}$

3. データの種類 i ごとに, 集合 $Y^{(i)}$ を計算する.
$Y^{(i)} = \{y_1^{(i)}, y_2^{(i)}, \cdots, y_{16}^{(i)}\},\ y_k^{(i)} = \frac{1}{10}\Sigma_{l=10i-9}^{10i} y_{k,l}$

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.
$$f(x_k) = \frac{a_0}{2} + \Sigma_{m=1}^{4}(a_m \cos mx_k + b_m \sin mx_k)$$
$$a_0 = \frac{1}{8}\Sigma_{k=1}^{16} F^{(i)}(x_k),\ a_{m'} = \frac{1}{8}\Sigma_{k=1}^{16} F(x_k)\cos m'x_k$$
$$b_{m'} = \frac{1}{8}\Sigma_{k=1}^{16} F(x_k)\sin m'x_k$$

Phase.2

1. データ種類 i' に関して推定した曲線を $f^{(i')}(x_k)$ とし, i' と k に関して以下の関数を与える.
$$D^{(i')}(k) = \sqrt{\frac{1}{10}\Sigma_{l=10i'-9}^{10i'}(y_{k,l} - y_k^{(i')})^2}$$

2. 以下の判別関数を与える. $fD^- = f^{(i')}(x_k) - D^{(i')}(k), fD^+ = f^{(i')}(x_k) + D^{(i')}(k)$ として,
$$L^{(i')}(l,k) = \begin{cases} 1 & fD^- < y_{k,l} < fD^+ \\ 0 & otherwise \end{cases}$$

3. 判別関数から, i' に関して以下のような (i,k) 成分を持つ分割表を生成する.
$$(i,k) = \Sigma_{l=10i-9}^{10i} L^{(i')}(l,k)$$

5 実験と考察

本章では, 次に示す PDF ファイルを特徴抽出アルゴリズムに適用し, 生成された分割表の一部を代数統計に基づいた統計的検定法で解析することで, 異なる要素で構成されるファイル間に有意な差異を持つ特徴が存在するかどうかを検証する. $i = 1$:文章, $i = 2$:文章+画像, $i = 3$:画像, $i = 4$:URL リンク, $i = 5$:スキャン で構成される PDF ファイルを各 10 個用意した. $i = 1$ のファイルサンプルから phase.1 によって推定される曲線 $f^{(1)}(x_k)$ と, その頻度の集合 $Y^{(1)}$ を同一の xy 平面上にプロットした画像を図 2 に示す.

また $i' = 1$ として phase.2 から生成される分割表を, 表 1 に示す. 表 1 は, 文章のみを構成要素とする PDF ファイルのバイナリに出現する数の頻度分布を phase.1 で曲線近似し, 各 k の値に対して推定された曲線上の点

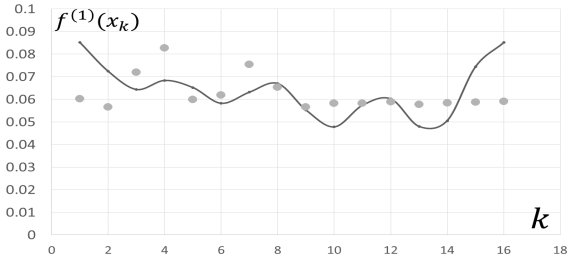

図 2: 曲線 $f^{(1)}(x_k)$ と頻度集合 $Y^{(1)}$

表 1: $i' = 1$ とした分割表

i/k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0	0	7	7	0	2	5	3	5	2	8	7	2	3	0	0
2	0	0	9	8	0	4	8	0	4	1	8	9	1	1	0	1
3	0	0	10	10	1	1	10	1	6	0	3	6	0	0	0	0
4	0	0	10	9	1	1	7	2	5	1	9	9	1	1	1	0
5	5	2	10	10	0	0	10	0	0	0	0	10	7	0	0	0

からある範囲に存在する点の数を, データの種類ごとにまとめた分割表である. 表 1 に示した $i = 1, 5$ の要素から構成される 16×2 分割表に関して, 代数統計に基づく統計的検定を行った. 結果として 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た.

6 まとめ

本研究では, コンピュータ上に存在するファイルのバイナリデータから特徴を抽出し, その特徴量に基づいてデータサンプルから分割表を生成するアルゴリズムを提案した. 5 種類の構成要素を持つ PDF ファイルをそれぞれ用意し, 提案したアルゴリズムから分割表を生成した. 生成した分割表の一部のデータを代数統計に基づく統計的検定法により解析したところ, 対象としたデータから抽出された特徴量に有意な差がみられた. 様々な形式のデータファイルに本手法を適応することで, その有用性を検証す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)