位相的データ解析による手書き文字の筆跡鑑定

岩頭 由樹[1]　佐藤 好久[2]

概要：現在収集できるようになったデータの量や種類は，情報技術の発展により，大きく増加している．それに伴い，たくさんのデータを分別するデータ解析技術の発展も求められている．例えば，画像に書かれた文字をデータ解析によって認識して，その画像に何が書かれているか抽出することである．本研究の目的は，ある文字が書かれた画像に TDA(Topological Data Analysis) 技術を使って得られた特徴を，誰が書いた文字かを，正しく分類できるか確かめるものである．手書きの文字が，TDA とパターン認識によって筆跡がどう解析されるかを示していった．本研究の手法は，先行研究の TDA を用いない手法に比べて識別率が高いことが確認された．

キーワード：情報数学，画像分類，機械学習

Handwriting analysis of handwritten letters by topological data analysis

Abstract: The amount and types of data that can now be collected have increased significantly with the development of information technology. This has also required the development of data analysis technology to sort through the large amount of data. For example, it is required to recognize characters written on an image and extract what is written on the image by data analysis. The purpose of this research is to confirm the features obtained by using TDA (Topological Data Analysis) technology on an image with certain characters can be correctly classified as to who wrote the characters. We show how handwriting letters can be analyzed by TDA and pattern recognition. It is confirmed that the method proposed in this study has a higher discrimination rate than the methods used in previous studies without TDA.

Keywords: Information Mathematics, Image Classification, Machine Learning

1. はじめに

現在収集できるようになったデータの数や種類は，情報技術の発展とともに，大きく増加している．それに伴い，たくさんのデータを分別するデータ解析技術の発展も求められている．例えば，画像に書かれた文字をデータ解析によって認識して，画像に何が書かれているか抽出することである．TDA(Topological Data Analysis) 技術とは，近年データ解析技術に対する新しい手法として注目されている．あるデータが持つ「形」を認識し，その形がもつ情報を位相的な観点から解析する手法である．従来の方法では，データを分析する際，統計的モデルという「形状」を仮定して分析するため，統計的モデルに当てはめることができないデータは解析がうまくいかないことがあるが，TDA は統計的モデルという「形状」に依存しないため，様々なデータの「形状」を解析できるとされている．

筆跡鑑定の先行研究 [5] では，手書き文字の濃度ヒストグラム，画素値，文字の占有率 (筆跡の太さ，大きさ) という特徴値を検出した後，機械学習 (ニューラルネットワーク) により研究を行っている．また，TDA を利用した文字認識の先行研究 [4] では，画像データからから得られたパーシステント図をベクトル化し，機械学習 (サポートベクターマシンと平均最近傍法) を用いて文字画像の識別を行っている．そして前述の 2 つの研究より，1 文字の手書き画像データから TDA によって得られたパーシステント図をベクトル化し，機械学習 (サポートベクターマシン，平均最近傍法，ニューラルネットワーク) を用いて筆跡鑑定

[1] 九州工業大学大学院情報工学府情報創成工学専攻
Department of Creative Informatics,
Graduate school of Computer Science and Systems Engineering,Kyushu Institute of Technology
[2] 九州工業大学大学院情報工学研究院　知能情報工学研究系
Faculty of Computer Science and Systems Engineering,
Kyushu Institute of Technology

の研究 [6] を行い，TDA を用いた筆跡鑑定が使わないときより識別率が高かったことがわかった．今回は，過去の研究 [6] から発展させ，複数文字列の手書き画像で筆跡鑑定を行うほか，1 文字の筆跡鑑定で学習データの量を変えた場合はどう結果が変化するかの確認も行う．

本研究の目的は，TDA 解析により筆跡鑑定を行い，複数文字列の手書き画像は誰が書いたものかをどれだけ正しく識別されるかを示すことで，筆跡鑑定における TDA の働きを確かめることにある．また，文字画像データの筆跡鑑定の精度をより良くする可能性を検討するためでもある．

2. 白黒二値画像のパーシステント図

文字画像データの解析のために，パーシステント図を利用する．そのためにパーシステントホモロジー群について示していく．

2.1 準備

ここからは，位相的データ解析のパーシステントホモロジー群のための事前知識を示していく．

有限個の点 $P = \{x_i \in \mathbb{R}^N | i = 1, \ldots, m\}$ からなるデータを考える．この P に対し，抽象的単体複体を考える．有限集合 V と V の部分集合の有限個の族 $\mathcal{K}$ が，(1) $v \in V \Longrightarrow \{v\} \in \mathcal{K}$，(2) $\sigma \in \mathcal{K}, \tau \subset \sigma \Longrightarrow \tau \in \mathcal{K}$ を満たすときの $(V, \mathcal{K})$ を，**抽象的単体複体**，または簡単に**抽象的複体**という．位相的データ解析では，有限個のデータ (点) を「図形」として幾何的に表現するため，抽象的単体複体を利用する．その代表的なものに，**チェック複体**，**ヴィートリス・リップス複体**があり，それぞれ $\mathcal{C}(P, r)$，$\mathcal{R}(P, r)$ と書く．その抽象的単体複体は，それぞれ次のような性質をもつ．$0 \leq \forall s, t \leq k$ とする．

$$\{x_{i_0}, \ldots, x_{i_k}\} \in \mathcal{C}(P, r) \Leftrightarrow \bigcap_{j=0}^{k} B_r(x_{i_j}) \neq \emptyset$$

$$\{i_0, \cdots, i_k\} \in \mathcal{R}(P, r) \Leftrightarrow B_r(x_{i_s}) \cap B_r(x_{i_t}) \neq \emptyset$$

ここで，P に対して，各点 x_i を中心とした半径 $r(> 0)$ の球 $B_r(x_i) = \{x \in \mathbb{R}^N | \|x - x_i\| \leq r\}$ を配置する．ここで $\|x\|$ はユークリッドノルムを表す．この 2 つの抽象的単体複体は，半径 r が $0 < r < r'$ のとき，$\mathcal{C}(P, r) \subset \mathcal{C}(P, r')$，また $\mathcal{R}(P, r) \subset \mathcal{R}(P, r')$ が成立する．これより，正の数からなる増大列 $r_0 < r_1 < \cdots < r_T$ が与えられたとき，$\mathcal{C}(P, r_0) \subset \mathcal{C}(P, r_1) \subset \cdots \subset \mathcal{C}(P, r_T)$，また $\mathcal{R}(P, r_0) \subset \mathcal{R}(P, r_1) \subset \cdots \subset \mathcal{R}(P, r_T)$ となる．この増大列を**フィルトレーション**と呼ぶ．

また，抽象的単体複体の「時系列」K^t $(t = 0, 1, 2, \ldots)$ が $\mathbb{K} : K^0 \subset K^1 \subset K^2 \subset \cdots \subset K^t \subset \cdots$ を満たすフィルトレーション $\mathbb{K} = \{K^t | t = 0, 1, 2, \ldots\}$ が，$K^j = K^\Theta$ を満たすとき，フィルトレーション $\mathbb{K}$ は**有限型**であるという．Θ の最小値を**飽和時刻**という．

2.2 パーシステントホモロジー群と白黒二値画像のパーシステント図

有限個の単体からなる抽象的単体複体の多面体と同相な位相空間 X の環 $\mathbb{Z}_2$ 係数 q 次元ホモロジー群 $H_q(X)$ は，有限生成 $\mathbb{Z}_2$ 加群として与えられる．直感的に，$H_q(X)$ の生成元は X の q 次元の穴を表し，その個数は X 内の本質的な q 次元の穴の個数を表す．例えば，$q = 0, 1$ はそれぞれ連結成分，輪を表す．

加群の構造定理により，係数環 $\mathbb{Z}_2$ による K^t のホモロジー群 $H_q(K^t; \mathbb{Z}_2)$ は，いくつかの因子 $\mathbb{Z}_2$ による直和分解

$$H_q(K^t) \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \cdots \mathbb{Z}_2 \qquad (*)$$

を持つ．

一方，$t < s$ より，$K^t \subset K^s$ である．包含写像 $i^{ts} : K^t \hookrightarrow K^s$ が誘導する準同型写像 $\varphi : H_q(K^t) \to H_q(K^s)$ を通じて，直和分解 $(*)$ の各生成元の時間変化を見ることができる．例えば，直和分解 $(*)$ の第 i 番目の因子の時間変化 $t_1 < t_2 < \cdots < t_n < \cdots$ に対応する部分を取り出すと，

$$\cdots \to 0 \to \cdots 0 \xrightarrow{t_1} \mathbb{Z}_2 \xrightarrow{t_2} \mathbb{Z}_2 \to \cdots \to \mathbb{Z}_2 \xrightarrow{t_n} 0 \to \cdots$$

のようになる．このとき，対応する生成元は時刻 $t = t_1$ で発生し，時刻 $t = t_n$ で消滅したと考えられる．

このようなホモロジー群の時系列を多項式環 $\mathbb{Z}_2[x]$ による加群としてとらえたものがフィルトレーション $\mathbb{K}$ の q 次元のパーシステントホモロジー群 $PH_q(\mathbb{K})$ である．このとき，$PH_q(\mathbb{K})$ は，次数付き $\mathbb{Z}_2[x]$ 加群として，次のような一意的な分解をもつ．

$$PH_q(\mathbb{K}) \simeq \bigoplus_{i=1}^{p} I(b_i, d_i)$$

$b_i, d_i \in \mathbb{R}$ かつ $b_i < d_i$ である．ただし，

$$I(b_i, d_i) = (x^{b_i})/(x^{d_i})$$

である．ここで，(x^k) は x^k で生成される $\mathbb{Z}_2[x]$ の単項イデアルである．

各 $I(b_i, d_i)$ の生成元は，$t = b_i$ で発生し，$t < d_i$ まで持続し，$t = d_i$ で消滅する q 次元の穴を表現している．b_i，d_i，$d_i - b_i$ をそれぞれ**発生時刻**，**消滅時刻**，**持続時間**という．このとき，各生成元の発生消滅時刻対 (b_i, d_i) からなる多重集合

$$D_q(\mathbb{K}) = \{(b_i, d_i) | i = 1, \ldots, p\}$$

を $\mathbb{K}$ の q 次元パーシステント図という．持続時間の長い発生時刻と消滅時刻のペア $(b_i, d_i) \in D_q(\mathbb{K})$ は $\mathbb{K}$ の信頼できるトポロジー構造として評価し，反対に，持続時間が短いものはノイズになりやすい．

ここから，白黒二値画像のパーシステント図について記す．

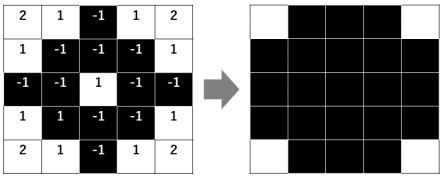

図 1　白黒二値画像についての時間変化

図 2　手書きの数字の 8

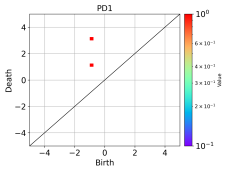

図 3　数字の 8 に対する結果

マンハッタン距離を導入し，その増減に伴う形の変化を調べていく．マンハッタン距離は，あるピクセルから別のピクセルまでに到達するために横切る辺の数の最小値で与えられる．

二値画像は黒の領域と白の領域に分けられ，計算の際はどちらの領域の表面までの距離を示すか選択する．選択した領域の外側に太らせる操作なら横切る辺の数は距離として正の数で与えられ，内側に細らせる操作なら，距離は負の値で表される．

図 1 の各ピクセルには，黒の領域までの距離が記載してある．図 1 のフィルトレーションは，単体複体によるものではないが，図 1 の左から黒の領域を 1 ピクセル太らせる操作を行うと，図 1 の右のようになる．この操作は，マンハッタン距離が 1 の白ピクセルを黒ピクセルにすることである．図 1 においては，白の領域を囲む黒の領域のわっかを太らせて，図 1 左の中心の白の領域 (穴) を消滅させている．

図 1 の左を $t=0$，右を $t=1$ とすると，$t=0$ 時点の中心にある黒い部分のわっかの存続期間は 1 で，1 次元のパーシステント図は $\{(0,1)\}$ で与えられる．

また，白黒二値画像の文字データとその 1 次元のパーシステント図を示す．白黒二値画像は，[2] の MNIST データを PNG 画像化したもののうちの 1 枚を，黒文字になるように白黒を反転させたものを用いた．

図 2 に示すような数字の画像データに対し，白黒二値画像化して，その黒い部分に注目する．それに存在する穴がいつ発生し，いつ消失するかを確認する．その結果として図 3 が得られる．この図 3 で確認できる点は，画像の文字が持つ穴の数に対応している．この場合，赤色の点が 2 個存在し，図 2 に書かれた数字の 8 の 2 個の穴は，図 2 時点よりも黒の領域が膨張してから消滅することがわかる．

3. 機械学習

ここからは，筆跡鑑定に用いるパターン認識のための機械学習の手法について示す．

3.1 サポートベクターマシン

サポートベクターマシンとは，主に二値分類に用いられる．多クラス分類への応用も可能である．[3], [8] を参考にした．

データは特徴ベクトル $\boldsymbol{x} = {}^t(x_1, x_2, \cdots, x_p)$ と目的変数 y からなる $(\boldsymbol{x}, y)$ からなり，その学習データ $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots (\boldsymbol{x}_m, y_m)$ が得られているとする．目的変数 y は，本研究においては手書き数字の認識で分類のアルゴリズムを使うのため，クラスのラベルとして使われる．

分類のアルゴリズムは，ラベルの種類の多さに応じて，2 値分類，多クラス分類に分けられる．前者は 2 つのクラスに分類する判別のことを指し，後者は 3 つ以上のクラスタリングに分類する判別のことを指す．多クラス分類では，1 つのクラスと残りのクラスに分類することでクラスを決めていく 1 対他方式，全体から 2 つのペアを作って分類し，すべての組み合わせに対して，多数決を行う 1 対 1 方式がある．これらは 2 値分類を複数回行うことで得られる．

学習データから，その多くに対して $h(\boldsymbol{x}_i) = y_i$ となるような判別器 $h : \mathbb{R}^p \to \{1, -1\}$ を学習し，新たな入力 $\boldsymbol{x}$ のラベルを $h(\boldsymbol{x})$ で予測することになる．

判別器について書く．符号関数 sign を

$$\mathrm{sign}(z) := \begin{cases} 1 & (z \geq 0) \\ -1 & (z < 0) \end{cases}$$

と定める．このとき，関数 $f : \mathbb{R}^p \to \mathbb{R}$ で，多くのデータに対して $\mathrm{sign}(f(\boldsymbol{x}_i)) = y_i$ を満たすもの見つけられたとき関数 f を判別関数といい，判別式 h を $h(\boldsymbol{x}) = \mathrm{sign}(f(\boldsymbol{x}))$ と定められる．

p 次元空間 $\mathbb{R}^p$ における 1 つの超平面

$$w_1 x_1 + w_2 x_2 + \cdots + w_p x_p + b = {}^t\boldsymbol{w}\boldsymbol{x} + b = 0$$

$({}^t\boldsymbol{w} = {}^t(w_1, w_2, \cdots, w_p))$ を境界面にすることにより，学習データ $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots (\boldsymbol{x}_m, y_m)$ が 2 種類のラベルをもつグループに完全に分離される状況を考える．この場合，判別関数 f を

$$f(x) := w_1 x_1 + w_2 x_2 + \cdots + w_p x_p + b = {}^t\boldsymbol{w}\boldsymbol{x} + b$$

と定義する．

サポートベクターマシンでは，この判別関数を与える超平面を「マージンが最大」となるようにする．このような超平面をマージン最大化超平面という．マージンとは，各学習データと分離超平面との距離の最小値のことである．

マージン最大化超平面は，マージン最大化問題 $y_i({}^t\boldsymbol{w}\boldsymbol{x}_i + b) > 0 \quad (i = 1, 2, \ldots, m)$ となるような

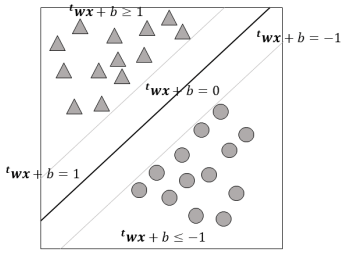

図 4 サポートベクターマシンによる識別境界の例

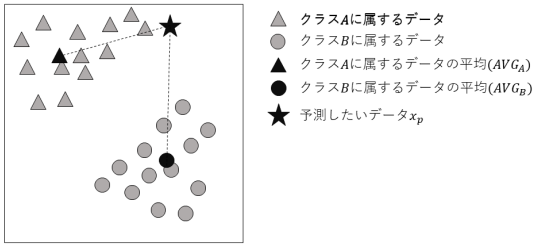

図 5 2 次元特徴ベクトルによる平均最近傍法の例

$$\max_{\boldsymbol{w},b}\min_{i=1,2,\cdots,m}\frac{|{}^t\boldsymbol{w}\boldsymbol{x}_i+b|}{\|\boldsymbol{w}\|}$$

の最適解として得られる．すなわち, $y_i({}^t\boldsymbol{w}\boldsymbol{x}_i+b)>0 \quad (i=1,2,\ldots,m)$ となるような

$$\max_{\boldsymbol{w}\in\mathbb{R}^p,b\in\mathbb{R}}\min_{i=1,2,\cdots,m}\frac{1}{2}\|\boldsymbol{w}\|^2$$

の最適解が求めたいマージン最大超平面を与える．これを**サポートベクターマシン** (SVM) という．

この SVM の定義により，図 4 が表せる．

3.2 平均最近傍法

多次元の特徴ベクトル $\boldsymbol{x}=\begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix}$ と目的変数 $\boldsymbol{y}=(y_1,\cdots,y_l)$ が与えられ，学習データ $(x_{11},\cdots,x_{1n},y_1),\cdots,(x_{m1},\cdots,x_{mn},y_l)$ が得られているとする．このとき，各種類の特徴ベクトルの個数を $N_{y_1},N_{y_2},\ldots,N_l$ としたとき，各種類の目的変数ごとの特徴ベクトルの平均 AVG_{y_i} を

$$\begin{aligned} AVG_{y_i} &\coloneqq (\overline{x_{i1}},\ldots,\overline{x_{in}},y_i) \\ &= \left(\frac{1}{N_{y_i}}\sum^{N_{y_i}}x_{i1},\cdots,\frac{1}{N_{y_i}}\sum^{N_{y_i}}x_{in},y_i\right) \end{aligned}$$

と定義できる．$X_1,\cdots,X_n$ のデータを加算するのは，目的変数 y_i に分類されたものだけである．

このとき，分類したいデータ $\boldsymbol{x}_p=(x_{p1},\ldots,x_{pn})$ に対して，$\min_{i=1,\cdots}\{(\overline{x_{i1}}-x_{p1})^2+\cdots+(\overline{x_{in}}-x_{pn})^2\}$ となるような i の目的変数 y_i を x_p に与える．これを**平均最近傍法**という．

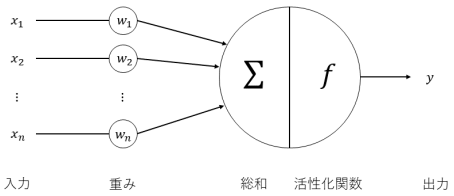

図 6 ニューロンのモデル

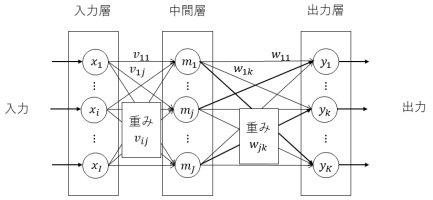

図 7 ニューラルネットワークのモデル

例として，特徴ベクトルが 2 次元の場合のときのイメージ図として図 5 を示す．図 5 の場合，予測したいデータ x_p はクラス A に属するデータの平均値に近いため，目的変数として A を与える．

3.3 ニューラルネットワーク

ニューラルネットワーク (NN) とは，参考文献 [3], [9] より，モデル化したニューロンを複数つなげたネットワーク構造において，信号伝達を繰り返して情報処理を行うアルゴリズムである．ニューラルネットワークには学習機能が備わっており判断規則を構築する．

ニューロンのモデルは，図 6 のように表せる．$x_1,x_2,\ldots,x_n$ は入力，それぞれに対応した $w_1,w_2,\ldots,w_n$ は重み，ニューロンは入力の重み付き総和 $\sum$ と活性化関数 f で発火する．活性化関数は非線形関数である．本研究においては，活性化関数は **ReLU 関数**を使用した．ReLU 関数は，以下のように定義できる．

$$f(x)=\begin{cases} 0 & (x<0) \\ x & (x\geq 0) \end{cases}$$

他のニューロンから伝播される入力情報の総和は，これにより，値の変換をされた後に出力される．入力の総和から閾値を引いた値が正なら 1，負なら 0 を出力するなど，二値の値を出力する．閾値を a とすると，出力値 y は $y=f\left(\sum_{i=1}^{n}x_iw_i-a\right)$ になる．この式において，$-a=x_0w_0$ とし, x_0 は常に 1 の入力があるユニットとする．これは, ニューロンの中のバイアスとして利用する．$-a=b$ として，b をバイアスとすると，$y=f\left(\sum_{i=1}^{n}x_iw_i+b\right)$ になる．

図 7 のように，ニューラルネットワークは表すことができる． x_i は入力層の i 番目の要素，m_j は中間層の j 番

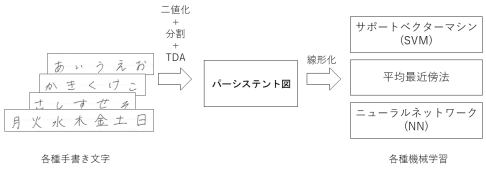

図 8　TDA を使った筆跡鑑定の概略図

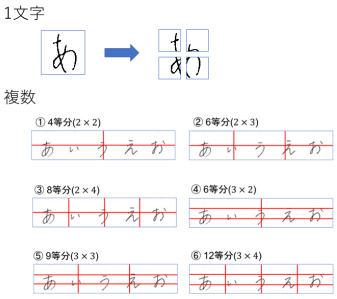

図 9　手書き画像の分割処理

目のニューロンの出力，y_k は出力層の k 番目のニューロンの出力，v_ij は入力層の i 番目から中間層の j 番目のニューロンとの間の結合重み，w_jk は中間層の j 番目から出力層の k 番目のニューロンとの間の結合重みを表す．また，入力層から出力層へ向かう演算は順方向演算といい，出力層から入力層へ向かう演算は逆方向演算という．

4.　実験手順

本研究においての実験では，TDA 解析結果のパーシステント図をベクトル化して，3 種類の機械学習の手法 (SVM, 平均最近傍法，NN) で筆跡鑑定を行った．また，TDA を用いずに，NN を用いた筆跡鑑定も行った．

本実験の概略図を図 8 に示す．手書き文字のデータは，独自に収集した，4 人分のものを用いる．

「あいうえおかきくけこさしすせそ月火水木金土日」を，1 文字につき 7 回書かせた画像を白黒二値画像にする．1 文字なら文字の枠に合わせて切り出す．複数文字列なら，ひらがななら「あいうえお」,「かきくけこ」,「さしすせそ」と，漢字なら「月火水木金土日」と切り出したものを複数文字列の手書き画像の入力データとして使用する．

TDA 解析を行う際には，図 9 のように，1 文字の画像を 4 等分し，複数文字列の画像を 6 通りの分割をした上で行う．図 9 における①の 2 × 2 の 4 等分，②の 2 × 3 の 6 等分，③の 2 × 4 の 8 等分，④を 3 × 2 の 6 等分，⑤の 3 × 3 の 9 等分，⑥の 3 × 4 の 12 等分の分割方法を使う．この分割したものを，各部分ごとにパーシステントホモロジー群，パーシステント図を算出する．パーシステント図を一定数に分割することによって，ベクトル化を行う．

図 10 ではデータ処理の流れを表している．左から，4 等分に分割した複数文字列の手書き画像，その 4 等分した画像のうちの 1 つの画像における 0 次元のパーシステント図，

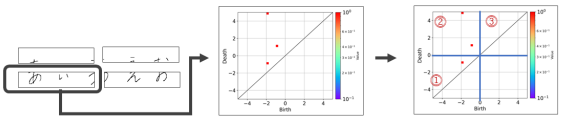

図 10　分割した画像のパーシステント図算出とベクトル化

0 次元のパーシステント図を格子状に分け，対角線を通る部分から上の部分をベクトルの要素としたものである．今回は 3 次元のベクトルへと線形化を行っている．各部分でベクトル化した 4 つのデータを，1 文字分に統合し，合計 12 次元のベクトルデータにする．図 10 では図 9 の複数の場合の①の 4 等分で分割したうえでベクトル化したため，合計 12 次元のベクトルデータになっている．図 9 の②は 18 次元，③は 24 次元，④は 18 次元，⑤は 27 次元，⑥は 36 次元のベクトルのデータになる．

また，0 次元のパーシステント図からベクトル化したものと 1 次元のパーシステント図からベクトル化したものを合わせたベクトルデータの次元は，合わせる前の 2 倍になる．この場合，図 9 における①なら合計 24 次元，②は 36 次元，③は 48 次元，④は 36 次元，⑤は 54 次元，⑥は 72 次元のベクトルのデータになる．

5 章 5 節までは複数文字列，6 節は 1 文字の筆跡鑑定を行う．

5.　実験結果

以下の実験結果において，機械学習によるパターン認識では，5 節までは 4 人分の，7 回書いてもらったうちの 4 回分を学習データ，残りの 3 回分をテストデータとして行う．つまり，各文字で 16 枚の文字画像データを学習データ，12 枚の文字画像データをテストデータとする．

5.1　TDA とサポートベクターマシンによる筆跡鑑定

機械学習として，サポートベクターマシン (SVM) を選択した場合，図 8 の機械学習の欄をサポートベクターマシンに指定した概略図の流れで筆跡鑑定を行う．

表 1　TDA と SVM による手書きの複数文字列画像の識別

識別率 (約〜%)	画像の分割法					
文字列	①	②	③	④	⑤	⑥
あいうえお	92	92	92	83	92	92
かきくけこ	67	67	75	83	83	83
さしすせそ	67	83	75	100	92	100
月火水木金土日	92	92	100	83	83	92

その実行結果は，表 1 の通りになる．表 1 の①，②，③，④，⑤，⑥は，左から図 9 の画像データの文字列の分割方法に対応している．

表 1 は，ひらがなの手書き文字列，漢字の手書き文字列を合わせて，ほとんどの場合で分割数が多いほど識別率が

高くなり，正しく筆者を識別できるようになっているとわかる．また，漢字の 2×4 の分割では，識別率が 100%とわかる．

5.2 TDA と平均最近傍法による筆跡鑑定

機械学習として，平均最近傍法を選択した場合，図 8 の機械学習の欄を平均最近傍法に指定した概略図の流れで筆跡鑑定を行う．

表 2 TDA と平均最近傍法による手書きの複数文字列画像の識別

識別率 (約〜%)	画像の分割法					
文字列	①	②	③	④	⑤	⑥
あいうえお	83	92	92	83	83	92
かきくけこ	67	67	75	83	92	83
さしすせそ	67	83	67	92	92	100
月火水木金土日	75	83	92	75	83	83

その実行結果は，表 2 の通りになる．表 2 の①，②，③，④，⑤，⑥は，左から図 9 の画像データの文字列の分割方法に対応している．

表 2 は，表 1 のときと同様に，ひらがなの手書き文字列，漢字の手書き文字列を合わせて，ほとんどの場合で分割数が多いほど識別率が高くなり，正しく筆者を識別できるようになっているとわかる．また，ひらがな「さしすせそ」の 3×4 の分割では，識別率が 100%とわかる．

また，筆跡鑑定の際に使う機械学習の比較としては，表 1 の結果は，表 2 よりも識別率が高くなったものと同じものがほとんどだが，「かきくけこ」の 3×3 のように表 2 の方が良い識別率だったものもあることがわかる．

5.3 TDA とニューラルネットワークによる筆跡鑑定

機械学習として，ニューラルネットワーク (NN) を選択した場合，図 8 の機械学習の欄をニューラルネットワークに指定した概略図の流れで筆跡鑑定を行う．

表 3 TDA と中間層のニューロン数 200 の NN の筆跡鑑定による識別率 (約〜%)

識別率 (約〜%)		画像の分割法					
	文字列	①	②	③	④	⑤	⑥
1000	あいうえお	92	92	92	92	92	100
	かきくけこ	92	58	58	75	67	75
	さしすせそ	92	92	100	100	92	100
	月火水木金土日	75	83	92	83	83	100
5000	あいうえお	92	92	92	92	92	92
	かきくけこ	92	83	67	83	83	67
	さしすせそ	92	92	100	100	92	92
	月火水木金土日	75	83	75	92	83	75
10000	あいうえお	92	92	92	92	92	100
	かきくけこ	83	83	67	92	75	83
	さしすせそ	83	92	100	100	83	92
	月火水木金土日	75	83	92	83	83	92

表 4 TDA と中間層のニューロン数 520 の NN の筆跡鑑定による識別率 (約〜%)

識別率 (約〜%)		画像の分割法					
	文字列	①	②	③	④	⑤	⑥
1000	あいうえお	92	92	92	92	92	100
	かきくけこ	75	83	58	75	75	75
	さしすせそ	92	92	100	100	92	100
	月火水木金土日	75	83	75	92	92	92
5000	あいうえお	92	83	92	92	92	100
	かきくけこ	83	83	58	83	75	75
	さしすせそ	92	92	100	100	92	100
	月火水木金土日	75	83	92	92	92	92
10000	あいうえお	92	100	92	83	92	100
	かきくけこ	92	75	58	83	67	75
	さしすせそ	92	100	100	100	92	100
	月火水木金土日	75	83	92	92	92	92

その実行結果は，表 3 と表 4 の通りになる．①，②，③，④，⑤，⑥は，左から図 9 の画像データの文字列の分割方法に対応している．中間層のニューロンの数は，200 と 520 の 2 通りに指定して行った．表 3 と表 4 にある 1000，5000，10000 は，学習の繰り返し回数である．学習の繰り返し回数は，1000 回，5000 回，10000 回の 3 通りで行った．

実行結果より，表 3 と表 4 で中間層のニューロン数を変化させても識別率に大きな変化はなく，また学習の繰り返し回数を増加させても識別率に大きな改善が見られることはないことが分かる．

表 3 と表 4 では，「あいうえお」から「月火水木金土日」までの複数文字列の手書き画像のうち，繰り返し回数によっては画像の分割数が多いほど識別率が良くなっている文字列もあれば，識別率が変わらないもの，また逆に，繰り返し回数 5000 回の「かきくけこ」のように識別率が悪くなっていったものもあることがわかる．

また，筆跡鑑定の際に使う機械学習の比較としては，表 3 と表 4 の結果は，表 1 と表 2 よりも識別率が高くなったものもあれば，低くなったものもあることが分かる．

5.4 1 次元のデータの筆跡鑑定と 0 次元と 1 次元の合計データでの筆跡鑑定の比較

次は，複数文字列の手書き文字の画像を，図 9 のように分割した後，1 次元のパーシステント図を用いてベクトルデータにしたもので筆跡鑑定，0 次元のパーシステント図のベクトル化データと 1 次元のパーシステント図のベクトル化データを合わせたベクトルデータを用いて筆跡鑑定を行う．図 8 の機械学習の欄を SVM または平均最近傍法に指定した概略図の流れで行う．

1 次元のパーシステント図を用いてベクトルデータにしたものに対する実行結果は，表 5 と表 6 の通りになる．実行結果より，こちらでは画像の分割数が多いほど識別率が良くなっているわけではないと分かる．

表 5 1 次元のパーシステント図の筆跡鑑定 (TDA+SVM)

識別率 (約〜%)	画像の分割法					
文字列	①	②	③	④	⑤	⑥
あいうえお	30	33	33	25	25	25
かきくけこ	50	42	42	17	17	42
さしすせそ	25	8	25	33	8	33
月火水木金土日	33	33	33	25	33	33

表 6 1 次元のパーシステント図の筆跡鑑定 (TDA+平均最近傍法)

識別率 (約〜%)	画像の分割法					
文字列	①	②	③	④	⑤	⑥
あいうえお	58	58	58	25	25	25
かきくけこ	33	25	33	42	33	33
さしすせそ	8	8	8	42	17	42
月火水木金土日	25	33	33	25	33	33

また，全体的に表 1 と表 2 よりも識別率が低くなっており，1 次元のパーシステント図より 0 次元のパーシステント図を用いたほうが識別率が高いことがわかる．

表 7 0 次元と 1 次元のパーシステント図の筆跡鑑定 (TDA+SVM)

識別率 (約〜%)	画像の分割法					
文字列	①	②	③	④	⑤	⑥
あいうえお	92	92	92	83	92	92
かきくけこ	67	67	75	83	83	83
さしすせそ	67	83	75	100	92	100
月火水木金土日	92	92	100	83	83	92

表 8 0 次元と 1 次元のパーシステント図の筆跡鑑定 (TDA+平均最近傍法)

識別率 (約〜%)	画像の分割法					
文字列	①	②	③	④	⑤	⑥
あいうえお	83	92	92	83	83	92
かきくけこ	67	67	75	83	92	83
さしすせそ	67	83	67	92	92	100
月火水木金土日	75	83	92	75	83	83

0 次元のパーシステント図のベクトル化データと 1 次元のパーシステント図のベクトル化データを合わせたベクトルデータを用いた識別結果は，表 7，表 8 の通りになる．実行結果より，こちらでは画像の分割数が多いほど識別率が良くなっていると分かる．

また，表 1 と表 2 と比較すると，識別率が全て同じであるとわかる．この筆跡鑑定では，0 次元のパーシステント図をベクトル化したデータが主に参照されているとわかる．

5.5 TDA を用いずに筆跡鑑定

また，参考として，TDA を使わずに筆跡鑑定を行った結果を示す．画像の画素値を検出したデータを，ニューラルネットワークを用いて筆跡鑑定を行う．

その実行結果は，表 9 の通りになる．中間層のニューロンの数は，200 と 520 の 2 通りに指定して行った．表 9 にある 1000，5000，10000 は，学習の繰り返し回数である．学習の繰り返し回数は，1000 回，5000 回，10000 回の 3 通りで行った．

表 9 画像値だけ検出したデータで行った NN の筆跡鑑定 (約〜%)

中間層の ニューロン数	200			520		
繰り返し回数	1000	5000	10000	1000	5000	10000
あいうえお	50	25	25	25	25	25
かきくけこ	25	25	25	25	25	25
さしすせそ	25	25	25	25	25	25
月火水木金土日	25	25	25	25	25	25

表 10 濃度ヒストグラム，画像値，文字の占有率を検出したデータで行った NN の筆跡鑑定 (約〜%)

中間層の ニューロン数	200			520		
繰り返し回数	1000	5000	10000	1000	5000	10000
あいうえお	25	25	25	25	25	25
かきくけこ	75	25	25	25	50	25
さしすせそ	50	25	25	50	25	25
月火水木金土日	50	75	25	25	50	25

表 9 より，文字画像の画素値を検出したデータで筆跡鑑定を行うと，識別率が 50%を超えた文字はほとんどないことが分かる．また，表 9 は，表 3 と表 4 の結果と比較すると，すべての文字列で識別率が下がっていることが分かる．「かきくけこ」，「さしすせそ」，「月火水木金土日」のように，繰り返し回数と中間層のニューロン数を変えても識別率が 25%から動かないものがほとんどだと分かる．

また，筆跡鑑定の先行研究 [5] のように，画像の c(筆跡の太さ，大きさ) を検出したデータを，ニューラルネットワークを用いて筆跡鑑定を行う．

その実行結果は，表 10 の通りになる．中間層のニューロンの数は，200 と 520 の 2 通りに指定して行った．表 9 にある 1000，5000，10000 は，学習の繰り返し回数である．学習の繰り返し回数は，1000 回，5000 回，10000 回の 3 通りで行った．

表 10 より，文字画像の画素値，濃度ヒストグラム，文字の占有率検出したデータで筆跡鑑定を行うと，識別率が 50%を超えた文字列は表 9 のときよりは多くなっていることが分かる．また，表 10 は，表 3 と表 4 の結果と比較すると，ほとんどの文字で識別率が下がっていることが分かる．表 10 は，中間層のニューロン数 200 のときの「月火水木金土日」のように，識別率繰り返し回数 5000 回では 75%なのに回数を重ねると識別率が 25%のようにとても低くなってしまうものもあることが分かる．

5.6 学習データ数を変更して TDA と SVM で筆跡鑑定

実際の筆跡鑑定に使われるデータには限りがある．その

データのうちどのくらいの量を学習データにすると筆跡鑑定の結果が良くなるかを確かめるために，学習データ数を変更して筆跡鑑定を行う．以下のやり方で，学習データとテストデータの枚数を変更した．

ⓐ 学習データが 4 回分，テストデータが 3 回分
ⓑ 学習データが 3 回分，テストデータが 3 回分
ⓒ 学習データが 5 回分，テストデータが 2 回分
ⓓ 学習データが 5 回分，テストデータが 3 回分

ⓐは 5 節までのデータの分け方と同じである．なおⓓのテストデータの数は，学習データとして使用していない 2 回分と，すでに学習データとして使用した 1 回分を合算したものである．つまりⓑは学習データが 12 枚になる．ⓒとⓓは学習データが 20 枚になり，ⓒだけテストデータが 8 枚になる．「あ」の 1 文字だけで，TDA と SVM を用いて筆跡鑑定を行う．

表 11 学習データ数を変えた場合の筆跡鑑定結果 (SVM)

識別率 (約〜%)				
文字	ⓐ	ⓑ	ⓒ	ⓓ
あ	67	67	75	75

結果は表 11 の通りである．学習データ数を少なくした場合は少なくする前と結果が変わらず，学習データ数を多くした場合は，テストデータ数が少なくても識別率が高くなった．ここから，学習データ数を多くした方が識別率が高くなることが分かった．

6. 結論

TDA を用いた筆跡鑑定のほうが，文字画像データの画素値を検出したデータでの筆跡鑑定，先行研究 [5] のように文字画像の画素値，濃度ヒストグラム，文字の占有率検出したデータで筆跡鑑定のように，TDA を使用せずに行った筆跡鑑定よりも，その文字を書いた人物が誰かの正しく識別する精度が高くなったことが分かった．TDA を用いないと，繰り返し回数によって識別率が急激に変化することがあったことも分かった．

TDA を利用した筆跡鑑定において，NN と平均最近傍法と SVM は，使用した機械学習ごとに識別率が上がった文字と下がった文字，変わらないとすべての結果がみられたため，どれが最も良い結果が得られているかは断定できないことが分かった．また，TDA を用いた筆跡鑑定に使う複数の手書き文字列の画像データは，SVM と平均最近傍法を使用した場合，分割数を多くすると識別率が高くなることが分かった．これは，TDA を用いると文字を書く位置や文字の太さも文字画像データの情報として保存されることと関係していると考えられる．また，TDA を用いた筆跡鑑定では，0 次元のパーシステント図を用いるほうが識別率が良くなることが分かる．これは，0 次元のパーシステント図は連結成分の birth と death，1 次元のパーシステント図は穴 (わっか) の birth と death を見ていることが影響していると考えられる．

また，学習データを増やした方が識別率が高くなることが分かった．

今後の課題としては，筆跡画像データの増加による識別率の変化の調査が挙げられる．また，学習データが増えすぎると過学習が起きてしまうため，それを考慮した識別率を上げるための筆跡画像データの枚数の限度の調査も挙げられる．

謝辞 本研究を進めるにあたり，研究に使用したデータを提供していただいた友人に感謝いたします．

参考文献

[1] 平岡裕章. タンパク質構造とトポロジー. 初版, 共立出版,2013,131p.
[2] Yann LeCun, Corinna Cortes, Christpher JC Burges. ”MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges”. 入手先 ⟨http://yann.lecun.com/exdb/mnist/⟩ (参照 2021-11-10).
[3] 林澪央. 位相的データ解析を用いたパターン認識の研究 〜手書き数字による実証〜. 九州工業大学, 修士論文, 2020.
[4] 岩頭由樹. 位相的データ解析を用いた手書き数字の認識. 九州工業大学, 卒業論文,2021.
[5] 高橋 真奈茄, 小出 洋. 機械学習を用いたパターン認識による筆者識別. 第 57 回 プログラミング・シンポジウム, 2016.1.8-10.
[6] 岩頭由樹. 文字画像データの位相的データ解析による筆跡鑑定への応用. 火の国情報シンポジウム 2023, 2023.3.13-14.
[7] トポロジカルデータ解析コミュニティ. ”HomCloud（基礎編）”, 東北大学. 2021-05-17. 入手先 ⟨https://www.wpi-aimr.tohoku.ac.jp/TDA/members/files/HomCloud(基礎編）.pdf⟩,(参照 2022-01-15).
[8] 金子真也. ”コンピュータ先端ガイド 2 巻 3 章 勉強会 (SVM)”,slideshare. 2017-05-23. 入手先 ⟨https://www.slideshare.net/MasayaKaneko/svm-76257267⟩ (参照 2022-01-15), p.10-11.
[9] 小林一郎. 人工知能の基礎. 初版, 株式会社 サイエンス社,2016,209p.
[10] A.Zomorodian and G.Carlsson. Computing Persistent Homology. Discrete Comput. Geom. Vol.33, 2005. p249-274.
[11] Ippei Obayashi, Yasuaki Hiraoka. Persistence Diagrams with Linear Machine Learning Models, arXiv. 2017,arXiv:1706.10082v2 [math.AT] 6 Jul 2017. 入手先 ⟨https://arxiv.org/pdf/1706.10082.pdf⟩,(参照 2022-02-03).
[12] Chazal, F., de Silva, V., and Oudot, S. Persistence stability for geometric complexes. Geometriae Dedicata, 173(1), 2014a, p193–214.
[13] Chazal, F., Glisse, M., Labruere, C., and Michel, B. Convergence rates for persistence diagram estimation in topological data analysis. In Proceedings of the 31st International Conference on Machine Learning (ICML14), 2014b. pp. 163–171.
[14] Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. Stability of persistence diagrams. Discrete Comput. Geom. 37(1): 2007. p103–120.