パーシステンス図のボトルネック距離を利用したテキストデータの文書分類

藤吉　陽成[1]　佐藤　好久[2]

概要：情報技術の発展に伴い、収集可能なデータの数や種類が急激に増加している。そして、膨大なデータから有益な情報を抽出するデータ解析技術には更なる発展が求められている。位相的データ解析（TDA）は、データが持つ「形」を位相的な観点から解析する手法である。本研究では、TDA を利用することで文書から特徴抽出をし、文書分類における TDA の有効性を検証したい。６つの話題テーマからなる対話データをデータセットとし、話題テーマが未知の対話データの話題を正しく予測することを目指している。研究手順として、文書内の文間距離を測り、パーシステンス図を作成し、パーシステンス図間のボトルネック距離を利用して文書分類を実装している。

キーワード：情報数学，分類学習，文脈／談話処理

Document Classification of Text Data Using Bottleneck Distance of Persistence Diagrams

YOUSEI FUJIYOSHI[1]　YOSHIHISA SATO[2]

Abstract: With the advancement of information technology, the amount and variety of data that can be collected have been rapidly increasing. Consequently, data analysis techniques capable of extracting useful information from large-scale data are required to advance.Topological Data Analysis (TDA) is a methodology that analyzes the “shape” of data from a topological perspective.In this study, we investigate the effectiveness of TDA for document classification by extracting features from textual data using topological methods.We use a dialogue dataset consisting of six different topic themes and aim to correctly predict the topic of unseen dialogue data. As a research procedure, we measure distances between sentences within a document, construct persistence diagrams, and perform document classification using the bottleneck distance between persistence diagrams.

Keywords: Information Mathematics, Classification Learning, Context ／ Discourse Processing

1. はじめに

情報技術の発展に伴い、収集可能なデータの数や種類が急激に増加している。そして、膨大なデータから有益な情報を抽出するデータ解析技術には更なる発展が求められている。本論文では、データ解析の１つの手法である位相的データ解析 (TDA : Topological Data Analysis) に着目し、その有用性を確かめ、新たに応用できないかを考えたい。

従来のデータ解析では、解析するデータを既知の分布モデルに当てはめて解析を行うため、分布モデルに当てはめることのできないデータに対しては解析が難しいとされてきた。それに対して、TDA はデータの分布モデルを必要としないため、あらゆるデータに対して分析を行うことができるとされている。

先行研究 [1] では、英文の Webtext とエッセイをデータセットとして、TDA を利用して、機械学習により人間の書いた文書と AI の書いた文書の分類を試みている。先行研究 [6] では、日本語の小論文をデータセットとして、TDA

[1] 九州工業大学大学院 情報工学府 情報創成工学専攻
[2] 九州工業大学大学院 情報工学研究院 知能情報工学系

を利用して、機械学習により人間の書いた小論文と AI の書いた小論文の分類を試みている。本研究では、日本語の 2 人対話データをデータセットとして、TDA を利用して、6 種類の対話話題の分類を試みている。そのために、パーシステンスダイアグラムをベクトル化したものを入力とした場合と、ボトルネック距離を入力とした場合で、k-NN 法を実装し、TDA を利用した文書分類が日本語においても可能であるか、また、文書分類においてパーシステンスダイアグラムをどのように扱うのが有効であるかを検証したい。

2. パーシステンスダイアグラム

まず、2 次元データに対するパーシステンス図について説明する。パーシステンスホモロジー群は、図形の連結成分や輪っか、空洞といった構造に着目することで、データの「形」を情報として抽出する手法である。図 1 のような点群を考える。この点群に対してフィルトレーションを用いて、各点における円の半径を徐々に大きくする。2 つが交わったところを線分、3 つが交わったところを中身の詰まった三角形とする。時刻 $t = 0$ において 4 つの点が存在する。時刻 $t = 1$ において円同士が交わり、点 b,c,d が線分となり、点は消滅していることが分かる。つまり点 b,c,d は、発生時刻 (birth) が $t = 0$ であり、消滅時刻 (death) が $t = 1$ と表すことができる。また、時刻 $t = 2$ において点 a,b,c と点 b,c,d がそれぞれ交わり、穴を生成している。点 b,c,d の穴は時刻 $t = 3$ において穴が潰れて消滅し、点 a,b,c の穴も時刻 $t = 4$ において消滅している。つまり点 b,c,d による穴は発生時刻 (birth) が $t = 2$ であり、消滅時刻 (death) が $t = 3$ と表すことができる。さらに、点 a,b,c による穴は発生時刻 (birth) が $t = 2$ であり、消滅時刻 (death) が $t = 4$ と表すことができる。このような (birth,death) の時間対を集めたものをパーシステンスダイアグラムとよぶ。0 次元パーシステンスダイアグラムは点について、1 次元パーシステンスダイアグラムは三角形などの閉じた折れ線について考える。また、パーシステンスダイアグラムをもとに、横軸を発生時刻 (birth)、縦軸を消滅時刻 (death) とし、時間対の集まりを平面上にプロットした図をパーシステンス図とよぶ。高次元データに対するパーシステンス図も同様に定義される。

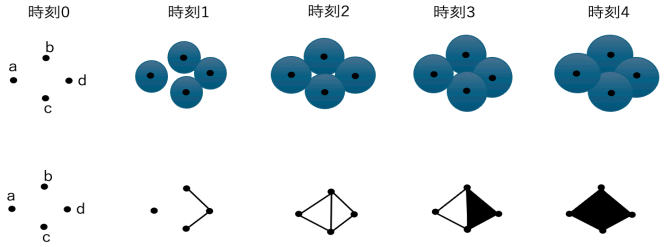

図 1 パーシステントホモロジー群

3. 自然言語処理

TDA は高次元の数値データを扱う研究において多くの成功を収めてきた。しかし、テキストデータでは、そのまま TDA を適用することはできない。TDA は点群データの形状の位相的特徴を抽出するために用いられるため、自然言語に適用するためには、テキストデータを幾何的オブジェクトに変換しなければならない。自然言語処理の多くで最初のステップとなるのは、自然言語を数値表現に変換することである。この数値表現では、多くの場合、単語の「意味」や「関連性」をカプセル化することを目的とした実数値ベクトルの形をとる。さらに、これらのベクトル埋め込みでは、似た意味を持つ単語が適切な空間において「近く」に配置されるように設計されることが多い。その結果、これらのベクトル埋め込みは本質的に幾何学的な性質を持ち、TDA を適用して分析することが可能になる。本研究では、文書を文単位で分割し、各文をベクトル化している。また、TF-IDF、Sentence-BERT、Word2Vec の 3 つのベクトル化手法を用い、それぞれについて実験を行った。

3.1 角距離

本研究では、文書内の文をベクトル化し、ベクトル化した文間の距離を角距離によって測っている。角距離は 2 つのベクトル $\boldsymbol{v}, \boldsymbol{u}$ に対して以下のように定義される。

$$d(\boldsymbol{v}, \boldsymbol{u}) = \cos^{-1}\left(\frac{\boldsymbol{v} \cdot \boldsymbol{u}}{\|\boldsymbol{u}\| \, \|\boldsymbol{v}\|}\right)$$

この角距離は、単にベクトル $\boldsymbol{v}$ とベクトル $\boldsymbol{u}$ の角度を測るものであり、$\mathbb{R}^n$ 内のベクトル間の類似性を捉えるのに適している。

4. 位相的データ解析による文書分類

今研究で用いるデータセット [7] は、6 つの話題テーマについて日本語で 2 人が対話した対話データがテキスト化されたものである。話題の種類としては、01. 食べること、02. ファッション、03. 旅行、04. スポーツ、05. マンガ・ゲーム、06. 家事になっている。実験方法は、まず、日本語の 2 人対話データに対して、3 つの手法でパーシステンスダイアグラムを求める。1 つは、対話データを TF-IDF、Sentence-BERT、Word2Vec の 3 つの手法でベクトル化し、パーシステンスダイアグラムを求める。次に距離を利用したパーシステンスダイアグラムを 2 つ求める。1 つは、対話データを単語の有無で 0,1 の 2 値で表現し、ハミング距離によって距離を測りパーシステンスダイアグラムを求める。もう 1 つは、対話データを単語の出現回数で表現し、マンハッタン距離によって距離を測りパーシステンスダイアグラムを求める。パーシステンスダイアグラムを求めるときは、SIF と SIFTS というアルゴリズによって、0 次元と 1 次元を求める。それぞれの手法において、求めたパーシステンスダイアグラム間のボトルネック距離を入力とした場合と、求めたパーシステンスダイアグラムをメッシュ化によりベクトル化したものを入力した場合の 2 通りで k-NN 法を実装し、文書分類を実装する。

4.1 フィルトレーション

4.1.1 Similarity Filtration (SIF)

SIF では、点集合 $\{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}\}$ に対して、与えられた距離行列を用いてヴィートリス・リップス複体を構築し、パーシステンスホモロジー群を計算する。このアルゴリズムでは、スケールパラメータ ε（直径または時間パラメータともよばれる）を段階的に増加させながら、トロポジー的特徴を抽出する。スケールパラメータ ε が小さい場合は角距離が小さな点同士、つまりは、類似度の高いテキスト単位同士が結びつく。スケールパラメータ ε が大きくなるにつれて、類似度の低いテキスト単位同士も結びついていく。また、SIF ではテキスト単位の順序については考慮されていない。

4.1.2 Similarity Filtration with Time Skeleton (SIFTS)

SIFTS は、SIF を改良したものであり、連続するテキスト単位間に順序関係を導入している。テキスト内の順序関係を反映するために次のような修正を加えている。

・テキスト単位 $\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}$ に対して隣接する単位間の距離を明示的に 0 と定義する（$d(\boldsymbol{x_i}, \boldsymbol{x_{i+1}}) = 0$）。これにより、テキスト内の順序関係が保持される。

上記の修正により、距離関数 d は厳密な距離ではなく、擬似距離となる。

SIFTS ではすべての連続する点が 1 次元単体（辺）として常に接続されている。これにより、すべてのテキスト単位が連続的に接続されるため、フィルトレーションのすべての段階において、0 次元ホモロジー群は常に単一の連結成分になる。すなわち、SIF では複数のクラスタが形成される可能性があるのに対して、SIFTS では常に全体が 1 つの連結体として表現される。

4.2 文書からパーシステンス図を作成

4.2.1 手法１：ベクトル化によるパーシステンス図

手法 1 の手順を以下に示す。

(1) 対話のテキストデータに対して不要文字やストップワードの削除等の前処理を行う
(2) 文書を文単位に分割し、各文をベクトル化する
(3) ベクトル化した文間の角距離を測る
(4) ベクトルをもとに point cloud を作成する
(5) フィルトレーションを作成する
(6) パーシステンス図を求める

上記の手法をベクトル化手法 TF-IDF、Sentence-BERT、Word2Vec に対して行う。

4.2.2 手法 2,3：距離を利用したパーシステンス図

手法 2,3 の手順を以下に示す。

(1) 対話のテキストデータに対して不要文字やストップワードの削除等の前処理を行う
(2) 文書を文単位に分割する
(3) 文間の距離（手法 2：ハミング距離、手法 3：マンハッタン距離）を測る
(4) 距離行列を作成する
(5) フィルトレーションを作成する
(6) パーシステンス図を求める

実際に求めたパーシステンス図のいくつかを例として図 2 から図 9 として以下に示す。

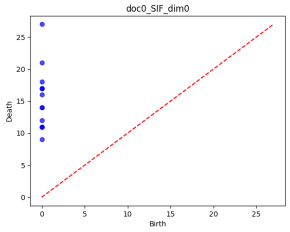

図 2　手法 1: パーシステンス図の例

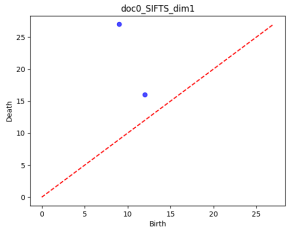

図 3　手法 1: パーシステンス図の例

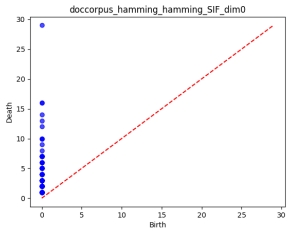

図 4　手法 2: パーシステンス図の例

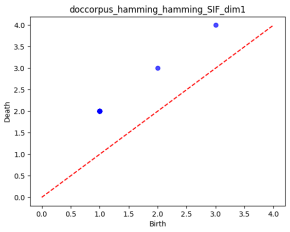

図 5　手法 2: パーシステンス図の例

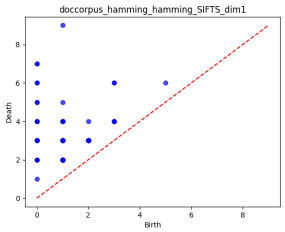

図 6　手法 2: パーシステンス図の例

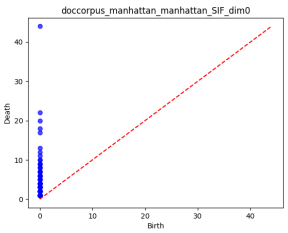

図 7　手法 3: パーシステンス図の例

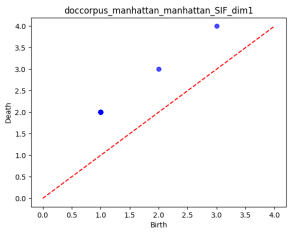

図 8　手法 3: パーシステンス図の例

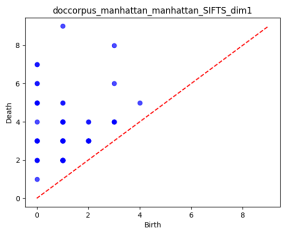

図 9　手法 3: パーシステンス図の例

4.3 メッシュ化とボトルネック距離と k-NN 法

4.3.1 メッシュ化

文書から作成したパーシステンスダイアグラムをベクトル化する手法として、メッシュ化を今研究では採用する。メッシュ化とは、パーシステンスダイアグラムを格子 (メッシュ) に分割し、各格子内に存在する点の個数を特徴量としてベクトル化する手法である。今研究では図のように、15 個の格子に分割するため、パーシステンスダイアグラムを 15 次元のベクトルに変換する。手法 1,2,3 で求めたパーシステンス図に対してベクトル化を行う。

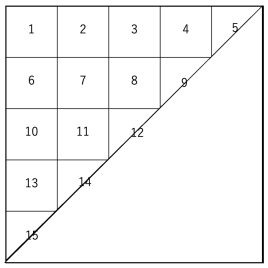

図 10 パーシステンスダイアグラムのベクトル化

4.3.2 ボトルネック距離

パーシステンス図の全体集合に対して距離関数を定義することができる。この距離関数を用いることで、2 つの異なるデータに TDA を行って得られたパーシステンスダイアグラムの比較が可能になる。この距離関数として、今回はボトルネック距離を採用する。

この距離が小さい場合もととなった 2 つのパーシステンス図は似ていると考える。手法 1,2,3 で求めたパーシステンス図に対してボトルネック距離を測る。

4.3.3 k-NN 法

線形判別関数による判別方法以外の判別分析を非線形判別分析という。非線形判別分析では、1 次関数以外の判別関数を用いる方法、距離に基づいた判別方法、多数決による判別方法、ベイズ判別法など多くの方法がある。

このうち、多数決による判別方法の 1 つとして k **最近傍 (k-Nearest Neighbor) 法 (k-NN 法)** がある。この手法は、伝統的なパターン分類アルゴリズムであり、判別すべき個体の周辺の個体で最も近いものを k 個見つけ、その k 個の多数決によりどのグループに属するかを判断するものである。

例として図 11 を挙げる。

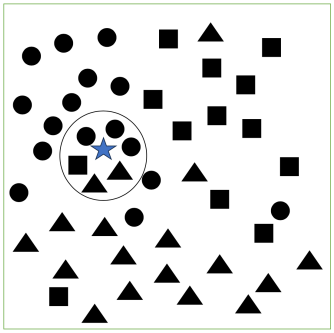

図 11 k-NN 法の例

図 2 は $k = 6$ の場合である。最初に、未知のデータである ★ の周辺で最も近い 6 個を探す。図中で円により囲まれた図形がそれにあたり、● が 3 個、▲ が 2 個、■ が 1 個である。k-NN 法によれば、未知のデータ ★ は ● のグループに属すると判断される。

本研究では、ラベルが未知のパーシステンス図とのボトルネック距離が小さいパーシステンス図を小さい順に k 個見つける場合と、メッシュ化によりベクトル化したものとのユークリッド距離が小さいパーシステンス図を小さい順に k 個見つける場合の 2 通りで、その k 個のパーシステンス図のラベルの多数決により未知のパーシステンス図のラベルを判断する。

5. 研究結果

5.1 手法 1：ベクトル化によるパーシステンス図

表 1 手法 1：メッシュ化した場合の k-NN 法の結果（平均分類精度 %）

k	**SIF,0**	**SIF,1**	**SIFTS,1**
TF-IDF			
$k = 1$	18.3	13.3	**21.7**
$k = 3$	16.7	13.3	17.5
$k = 5$	**20.0**	**15.8**	16.7
$k = 7$	19.2	10.1	15.8
Sentence-BERT			
$k = 1$	16.7	**22.5**	**21.7**
$k = 3$	**20.8**	19.2	20.0
$k = 5$	18.3	19.2	15.0
$k = 7$	**20.8**	16.7	19.2
Word2Vec			
$k = 1$	17.5	**21.7**	18.3
$k = 3$	**21.7**	18.3	**20.0**
$k = 5$	20.8	17.5	18.3
$k = 7$	20.8	16.7	19.2

表 2 手法 1：ボトルネック距離を入力とした k-NN 法の結果（平均分類精度 %）

k	**SIF,0**	**SIF,1**	**SIFTS,1**
TF-IDF			
$k=1$	**15.8**	**16.7**	14.2
$k=3$	13.3	15.0	14.2
$k=5$	13.3	13.3	**20.8**
$k=7$	12.5	15.0	18.3
Sentence-BERT			
$k=1$	**19.2**	18.3	**25.9**
$k=3$	18.3	**20.8**	18.3
$k=5$	16.7	14.2	19.2
$k=7$	16.7	15.8	18.3
Word2Vec			
$k=1$	**21.7**	13.3	13.3
$k=3$	19.2	18.3	15.0
$k=5$	19.2	**20.0**	**17.5**
$k=7$	20.8	17.5	16.7

5.2 手法 2： ハミング距離によるパーシステンス図

表 3 手法 2 メッシュ化した場合の k-NN 法の結果（平均分類精度 %）

k	**SIF,0**	**SIF,1**	**SIFTS,1**
ハミング距離			
$k=1$	**33.2**	18.3	**33.2**
$k=3$	24.2	15.8	25.0
$k=5$	27.5	19.2	22.5
$k=7$	27.5	**23.3**	20.8

表 4 手法 2 ボトルネック距離を入力とした k-NN 法の結果（平均分類精度 %）

k	**SIF,0**	**SIF,1**	**SIFTS,1**
ハミング距離			
$k=1$	**30.5**	**30.5**	**33.2**
$k=3$	23.4	23.4	16.7
$k=5$	21.7	21.7	19.2
$k=7$	18.3	20.8	20.8

5.3 手法 3：マンハッタン距離によるパーシステンス図

表 5 手法 3 メッシュ化した場合の k-NN 法の結果（平均分類精度 %）

k	**SIF,0**	**SIF,1**	**SIFTS,1**
マンハッタン距離			
$k=1$	**36.4**	20.8	**36.4**
$k=3$	26.7	21.7	25.1
$k=5$	25.1	23.4	18.3
$k=7$	25.1	**26.7**	16.7

表 6 手法 3 ボトルネック距離を入力とした k-NN 法の結果（平均分類精度 %）

k	**SIF,0**	**SIF,1**	**SIFTS,1**
マンハッタン距離			
$k=1$	**31.0**	**35.6**	**38.5**
$k=3$	18.3	28.8	20.8
$k=5$	30.5	26.7	21.7
$k=7$	25.1	25.1	22.3

6. 結果・考察

まず、手法 1 についての分析をまとめる。

メッシュ化を用いた場合、TF-IDF によるベクトル化では、SIF,0 次元で 20.0%、SIF,1 次元では 15.8%、SIFTS,1 次元で 21.7%を記録した。SIF,0 次元、SIFTS,1 次元で少しではあるがランダムと比較して性能向上が見られた。SIF,0 次元では文単位の違いを、SIFTS,1 次元では、文書構造の違いを捉えられると考えられるため、話題の変化が文や文書構造に少し変化を与える可能性が考えられる。次に、Sentence-BERT によるベクトル化では、SIF,0 次元で 20.8%、SIF,1 次元では 22.5%、SIFTS,1 次元で 21.7%を記録した。全てにおいてもランダムと比較して少しの性能向上が見られ、TF-IDF と同様に文単位での違い、文書構造の違いに加え、SIF,1 次元で捉えられると考えられる文同士の繋がりの違いを捉えた可能性が考えられる。次に、Word2Vec によるベクトル化では、SIF,0 次元で 21.7%、SIF,1 次元では 21.7%、SIFTS,1 次元で 20.0%を記録した。Sentence-BERT と同様に全てにおいてランダムと比較して少しの性能向上が見られた。ベクトル化手法の違いによる精度の変化はほとんど見られなかった。

ボトルネック距離を用いた場合、TF-IDF によるベクトル化では、SIF,0 次元で 15.8%、SIF,1 次元では 16.7%、SIFTS,1 次元で 20.8%を記録した。SIFTS,1 次元のみ少しではあるがランダムと比較して性能向上が見られた。話題の変化が文書構造に少し変化を与える可能性が考えられる。次に、Sentence-BERT によるベクトル化では、SIF,0 次元

で 19.2%、SIF,1 次元では 20.8%、SIFTS,1 次元で 25.9%を記録した。SIF,0 次元、SIF,1 次元においてもランダムと比較して少しの性能向上が見られ、SIFTS,1 次元では手法 1 における最高精度を記録した。ボトルネック距離を用いることで文書構造の違いを少し捉えやすくなる可能性が考えられる。次に、Word2Vec によるベクトル化では、SIF,0 次元で 21.7%、SIF,1 次元では 20.0%、SIFTS,1 次元で 17.5%を記録した。

手法 1 では、全体としてランダムからの精度向上は微量であった。文脈を考慮できる Sentence-BERT とボトルネック距離を組み合わせれば、SIFTS,1 次元において 25.9%を記録し、少し文書構造の違いを捉えられる可能性が考えられる。

次に、手法 2・手法 3 についての分析をまとめる。

メッシュ化を用いた場合、ハミング距離を用いると、SIF,0 次元で 33.2%、SIF,1 次元では 23.3%、SIFTS,1 次元で 33.2%を記録した。マンハッタン距離を用いると、SIF,0 次元で 36.4%、SIF,1 次元では 26.7%、SIFTS,1 次元で 36.4%を記録した。手法 1 と比較すると全体的な性能向上が見られた。ハミング距離やマンハッタン距離を用いることで単語の有無や頻度に着目した文と文の距離を測ることができるため、それにより文の違いや文同士の繋がりの違い、文構造の違いを手法 1 と比較して強く捉えられたと考えられる。特に、文単位の違い、文書構造の違いについてはハミング距離を用いると強く捉えられた。

ボトルネック距離を用いた場合、ハミング距離を用いると、SIF,0 次元で 30.5%、SIF,1 次元では 30.5%、SIFTS,1 次元で 33.2%を記録した。マンハッタン距離を用いると、SIF,0 次元で 31.0%、SIF,1 次元では 35.6%、SIFTS,1 次元で 38.5%(本研究における最高精度) を記録した。メッシュ化と同様に手法 1 から全体的な性能向上が見られた。メッシュ化と同様に、手法 1 と比較すると全体的な性能向上が見られた。手法 2、3 が手法 1 と比較して性能向上した原因として、TF-IDF、Sentence-BERT、Word2Vec といったベクトル化によって 1 文全体を考慮してベクトル化して point cloud を作成するため、文の端々に存在する単語の有無や頻度といった細かな違いを切り捨て、全体的な意味で類似した文同士を近くに配置してしまい、TDA におけるデータの持つ「形」を生む機会を奪ってしまったことが考えられる。ハミング距離やマンハッタン距離であれば、単語の有無や頻度といった事実のみで距離を測るため、1 文を全体的な意味で考えると類似した文同士であっても 1 単語の有無や頻度によって大きく距離ができる。それにより、TDA におけるデータの持つ「形」を生み出し、その単語の有無や頻度によって生まれる「形」の差が対話データの話題変化における文章の差と相関関係があり、文章の違いを捉えることに繋がり、性能が向上したと考えられる。また、手法 2 と手法 3 を比較すると、少しではあるが全体的に手法 3 の方が高精度を記録した。単語の有無のみでなく、頻度にまで着目することで文書の違いをより捉えられる可能性が考えられる。また、メッシュ化とボトルネック距離を用いる違いとして、SIF,1 次元での性能差が挙げられる。メッシュ化では SIF,0 次元、SIFTS,1 次元と比較して、SIF,1 次元で性能が約 10%ほど低下した。メッシュ化をすることで、文同士の繋がりの違いが捉えづらくなると考えられる。

しかし、最高精度であっても 38.5%であったため、絶対的な分類精度の向上は大きな課題に挙げられる。よって、ハミング距離やマンハッタン距離といった距離は、絶対的な性能向上に課題がありつつも、TDA を用いた文書分類に有効な可能性があると考えられる。メッシュ化とボトルネック距離による精度の違いは SIF,1 次元でのみ見られた。TDA の文書分類おいて、ボトルネック距離を採用することで限定的には利点が見られたものの、最高分類精度という観点では有意差が見られなかった。

次に、k-NN 法の k の値による性能向上がほとんど見られなかったことについて分析をまとめる。分類精度が高い項目において、今研究では k-NN 法の k の値が増えても性能向上が見られなかった。話題が同じ文章同士のパーシステント図間のボトルネック距離が近いことに全体的な相関が見られれば、k の値の増加は性能向上に繋がると考えられる。しかし、今研究ではほとんどの場合で $k = 1$ において最高精度を記録したため、そのような相関は見られず、最もボトルネック距離が近い文章同士の話題は同じである可能性が少し高いが、同じ話題のパーシステント図全体を見るとボトルネック距離はばらけていると考えられる。以下に、表 7 として話題 1(食べること) についての全体との分散比、表 8 として話題 3(旅行) についての全体との分散比、表 9 として TF-IDF, ボトルネック距離についての話題毎の全体との分散比、表 10 としてマンハッタン距離, ボトルネック距離についての話題毎の全体との分散比を示す。話題 1(食べること) については、所々で分散が小さくなっている項目が見られたが、平均すると話題 1 の分散は全体の分散と変わらない。この傾向は、話題 4(スポーツ) においても見られた。他の話題については、どの項目でも話題毎の分散と全体の分散は変わらなかった。したがって、ベクトル化や距離、パーシステンスダイアグラムの扱いによらず、k-NN 法にかけるデータは一様に混ざりあっていることを意味し、その結果として、$k = 1$ のときに高い精度を記録し、k の値が増えても精度は高まらなかったと思われる。

表 7 全体分散に対する話題 1(食べること) 分散の比

手法	SIF,0	SIF,1	SIFTS,1
TF-IDF			
ボトルネック距離	0.89	0.85	0.93
メッシュ化	0.91	1.29	0.61
Sentence-BERT			
ボトルネック距離	0.93	0.94	0.96
メッシュ化	0.78	0.62	0.63
Word2Vec			
ボトルネック距離	0.95	0.83	0.90
メッシュ化	1.07	0.99	0.70
ハミング距離			
ボトルネック距離	0.92	0.84	0.48
メッシュ化	0.92	0.84	0.48
マンハッタン距離			
ボトルネック距離	0.59	0.98	0.52
メッシュ化	0.59	0.98	0.52

表 8 全体分散に対する話題 3(旅行) 分散の比

手法	SIF,0	SIF,1	SIFTS,1
TF-IDF			
ボトルネック距離	1.15	0.79	1.09
メッシュ化	1.10	1.13	1.62
Sentence-BERT			
ボトルネック距離	1.00	1.34	1.05
メッシュ化	1.36	1.52	1.52
Word2Vec			
ボトルネック距離	1.03	0.91	1.18
メッシュ化	1.30	1.16	1.37
ハミング距離			
ボトルネック距離	1.31	1.03	1.49
メッシュ化	1.31	1.03	1.49
マンハッタン距離			
ボトルネック距離	1.13	0.94	1.47
メッシュ化	1.13	0.94	1.47

表 9 全体分散に対する TF-IDF, ボトルネック距離の分散の比

手法	SIF,0	SIF,1	SIFTS,1
TF-IDF			
01. 食べること	0.89	0.85	0.93
02. ファッション	0.92	1.01	0.97
03. 旅行	1.14	0.79	1.09
04. スポーツ	1.04	1.16	1.00
05. 漫画・ゲーム	1.20	1.12	0.93
06. 家事	1.06	1.00	1.10

表 10 全体分散に対するマンハッタン距離, ボトルネック距離の分散の比

手法	SIF,0	SIF,1	SIFTS,1
マンハッタン距離			
01. 食べること	0.59	0.98	0.52
02. ファッション	0.77	1.13	1.15
03. 旅行	1.13	0.94	1.47
04. スポーツ	0.85	0.94	1.26
05. 漫画・ゲーム	1.27	0.84	0.57
06. 家事	0.92	0.84	0.85

参考文献

[1] Løvlie Bendik,
Text Classification via Topological Data Analysis,
https://github.com/bendilo/TDA_text_classification.

[2] X. Zhu,
Persistent homology: An introduction and a new text representation for natural language processing, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, 2013, https://www.ijcai.org/Proceedings/13/Papers/288.pdf.

[3] 平岡裕章, 「タンパク質構造とトポロジー」, 初版, 共立出版, 2013.

[4] 東北大学自然言語処理研究グループ, 鈴木研究室, https://github.com/cl-tohoku/bert-japanese

[5] 東北大学自然言語処理研究グループ, 乾研究室, https://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/

[6] 藤吉陽成, パーシステント図のボトルネック距離を利用した線型空間への埋め込みによるテキストデータの文書分類, 九州工業大学, 卒業論文 (2024)

[7] 中俣尚己・太田陽子・加藤恵梨・澤田浩子・清水由貴子・森篤嗣 (2021)「『日本語話題別会話コーパス：J-TOCC』」『計量国語学』33 巻 1 号, pp.11-21, 計量国語学会.