様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月　 ７ 日現在

機関番号：１２６１２
研究種目：基盤研究(C)（一般）
研究期間：2015～2018
課題番号：１５Ｋ００１４８

研究課題名（和文）辞書の類似性に着目した圧縮率ベース特徴空間の最適な構築方法の探求

研究課題名（英文）Optimal Construction of Compression-based Feature Space

研究代表者
古賀　久志（Koga, Hisashi）

電気通信大学・大学院情報理工学研究科・准教授

研究者番号：４０３６１８３６

交付決定額（研究期間全体）：（直接経費）　　2,100,000円

研究成果の概要（和文）：圧縮ベースパターン認識は分析対象に対する事前知識なしに、データ分析を実現する非教示データ分析技術である。その基本は、データ間の類似性を圧縮率から測定することである。とくに本研究では、SVMやk-meansなど既存パターン認識技術を活用するため、データを圧縮率を並べたベクトルとして表現する圧縮率特徴空間を考え、その適切な空間の構築法をテーマとした。そして、特徴空間の軸を定義する圧縮辞書の単語を入れ替えて、軸間の独立性を高めることにより既存手法よりもパターン認識精度を7～8%向上することに成功した。

研究成果の学術的意義や社会的意義
近年、人工知能が大流行しているが、学習データを用意する手間が大きいのが課題である。とくにIoTの時代となり、これまで分析対象とされなかった新種のデータを分析する必要に迫られているが、そのようなデータはそもそも性質が不明なので学習データを用意するのが難しい。圧縮ベースパターン認識は非教示でデータ分析をする技術であり、上記のような性質が不明なデータを分析するのが得意であり、学習データを構築するための要素技術として重要である。本研究はそのパターン認識やクラスタリング（データ分類）の精度改善に貢献した。

研究成果の概要（英文）：Compression based pattern recognition is an unsupervised data analysis technique which realizes data analysis without prior knowledge about the data to be analyzed. Its primary point is to measure the similarity between two data based on the compression rate. In particular, in order to exploit the standard pattern recognition algorithms such as SVM and k-means, this research deals with compression-based feature spaces in which an object is represented as a compression vector consisting of multiple compression ratios and studies their effective construction. As the main result, by exchanging the words among the compression dictionaries each of which is responsible for one dimension so that they may be more independent one another, we succeeded in improving the pattern recognition accuracy by 7 to 8% as compared with the previous method in literatures.

研究分野：アルゴリズムとデータ構造

キーワード：圧縮ベースパターン認識　圧縮辞書　トライ　特徴空間

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

IoT の時代になり、センサデータなどこれまで解析対象でなかったデータを分析する必要が出てきている。こうした新しいデータには過去の資産がないため、非教示でデータを分析できることが重要になる。この背景の下、データ圧縮技術を活用した圧縮ベースパターン認識は、データの種類を問わずに適用可能で汎用性が高いことから注目されている。その基本原理は、2つのオブジェクト間の類似度を圧縮率から測ることである。圧縮ベースパターン認識において最も有名な手法は、NCD(Normalized Compression Distance)である。しかし、NCD をはじめとした多くの従来手法では、オブジェクト間の距離を求めるだけなので、類似度行列を前提とするパターン認識技術しか適用できない。つまり、特徴ベクトルを入力とする手法は適用不可能である。例えば、k-means クラスタリングや SVM(サポートベクタマシン)などは使用できない。

この問題に対し、当研究室ではオブジェクトを圧縮率特徴ベクトルとして表現する手法 PRDC を考案した。具体的には複数種類のオブジェクト T1,T2,…,Tn を LZW 圧縮して得られる辞書 D1,D2,…,Dn をあらかじめ用意する。そして、任意のオブジェクトをｎ個の辞書で圧縮し、n 個の圧縮率が並んだ n 次元ベクトルとして表現する。PRDC はこれまで航空画像解析やネットワーク管理といったアプリケーションに適用された実績を持つ。

その一方で、圧縮ベースパターン認識はその汎用性の高さゆえに、特定のアプリケーションに特化した専用手法と較べて認識性能が劣るという欠点があり、圧縮ベースパターン認識の性能を向上させることは重要性の高い研究課題になっている。

２．研究の目的

本研究では、PRDC の性能向上を目指す。PRDC でオブジェクトを圧縮率特徴ベクトルとして表現する場合、特徴空間の軸(を定める辞書)をどう選ぶかが認識性能に大きく影響する。そこで本研究では、圧縮特徴空間を生成するための最適な軸選択方式を探求することを研究目的とする。とくに、圧縮ベースパターン認識がパラメータフリーで使えるという長所を維持すべく、学習不要な非教示型の軸決定アルゴリズムを考案する。具体的には独立性が高い辞書を生成することで次元の独立性を高め、低次元でもパターン判別能力に優れた圧縮特徴空間を構築する。

上記と平行して、PRDC 以外の既存圧縮ベースパターン認識の性能向上にも取り組む：（１）まず、NCD のような距離ベースの手法に関して、パターン認識に有効な距離が得られるよう圧縮距離の計算方法を修正する。（２）次に特徴ベクトルベースの PRDC において、ベクトル要素となる特徴量を改善する。

３．研究の方法

本来、圧縮ベースパターン認識は IoT におけるセンサデータなど性質が未知なデータに適用してその有効性を示すべきであるが、そのようなデータは入手が困難である。また、仮に入手ができたとしても、データにアノテーションをしてパターン認識性能を論じるための正解データを用意するのが難しい。そこで、本研究では正解データの入手が容易な画像データセットを用いて実験評価を実施する。

４．研究成果

(1)圧縮特徴空間の構築アルゴリズム

PRDC は、特徴空間の軸を定める圧縮辞書をどう決定するかに、特徴空間の識別性能が強く影響される。一般的に特徴空間は各次元が互いに独立していることが望ましい。従来の PRCD では、種類が異なる複数のオブジェクトをデータセットから選択し、それらのオブジェクトから抽出した圧縮辞書を使用していた。つまり、種類が違うオブジェクトから抽出した辞書は互いに似ていないだろうという訳である。

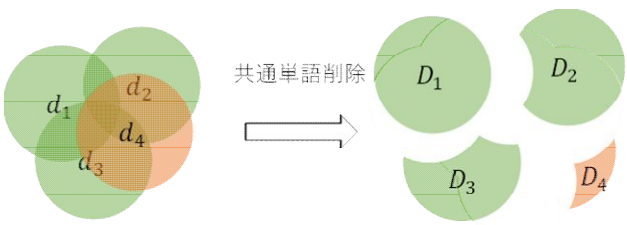

図１：共通単語の削除

これに対して、本研究ではさらに圧縮辞書間で重複した単語を削除することで、全ての辞書が単語を共有しないことを保証し、辞書間の独立性を高める手法を構築した。共通単語を排除する手順は次のようになる。辞書 Di (1 ≦ i ≦ n) に対しては、Di より若い辞書 Dj (1 ≦ j ≦ i) との共通単語を調べ、共通単語が存在した場合は Di から削除する。しかし、これだけだと、任意の圧縮辞書ペアが共通単語を持たない一方で、

添字の大きい後半の圧縮辞書ほど単語削除の結果、保有単語数が少なくなる（図１）。そして、保有単語数が少ない圧縮辞書は、圧縮効率が悪いため、対応する特徴次元の識別能力も低下することが懸念される。

そこで本研究では、共通単語を削除した後の圧縮辞書間で単語を再分配し、各圧縮辞書が持つ単語数を均等化した。単語再分配アルゴリズムは、

1. 全圧縮辞書に含まれる全単語をアルファベット順で分割する手法（辞書順ソート法）
2. 保有単語の最もすくない辞書から順番にランダムに消された単語を復活させる手法（単語復活法）

の 2 種類を考案した。

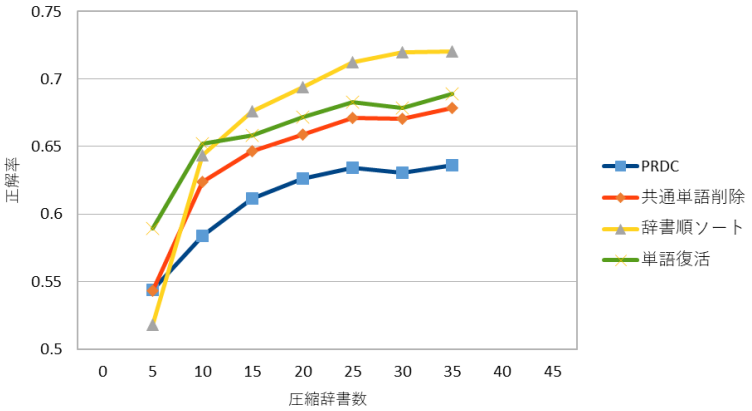

図 2:画像オブジェクト分類における正解率

提案手法を画像オブジェクト分類に適用した結果、共通単語の削除により認識率が 4%程度向上し、さらに単語数均等化により認識性能が 3%程度向上し、合計で従来の PRDC より認識率を 7%以上改善できることを示した（図 2）。

(2)圧縮ベースの時系列データ解析

通常、圧縮ベースパターン認識では LZW など 1 次元列を対象とする可逆圧縮アルゴリズムが使用される。このため、画像を取り扱う際には，事前に画像を 1 次元データへ変換する必要があり、その過程で情報が失われるという欠点があった。そこで、非可逆圧縮である MPEG-1 を用いて 2 枚の画像間の類似度を測る CK-1 距離が提案された。CK-1 距離では 2 枚の画像を連結した動画像を生成し，その圧縮後のファイルサイズから距離計算を行う。さらに近年，この CK-1 距離を使用した圧縮ベースの時系列分類手法 Recurrence Plots Compression Distance (RPCD) が提案された。RPCD 手法は時系列データから Recurrence Plots と呼ばれる特徴画像を生成し、Recurrence Plots 間の非類似度を CK-1 距離で計測して、時系列データを分類する。RPCD 手法は単純であることが利点であり、パターン認識の非専門家でも市販の MPEG-1 エンコーダさえあれば簡単に使用できる。しかし、MPEG-1 エンコーダをどう設定するべきか等の詳細は既存研究では明らかになっていなかった。

そこで、本研究ではまず RPCD 手法を実装し，その性質を調査した。調査の結果、MPEG-1 において動画の品質をコントロールする q 値と呼ばれる品質パラメータが、認識性能に大きく影響することを発見した。とくに q 値を不適切に設定してしまうと、認識性能がランダムな分類器より低くなってしまう場合すらあった。上記の観察結果を踏まえ、学習データを leave-one-out 法でクラス分類し、分析対象のデータセットに対し、適応的に q 値を定める qRPCD という手法を提案した。そして、RPCD を発案した論文と全く同じデータセットを使って、時系列データの分類実験を行い、qRPCD が従来の RPCD より認識性能を約 4%改善できることを示した

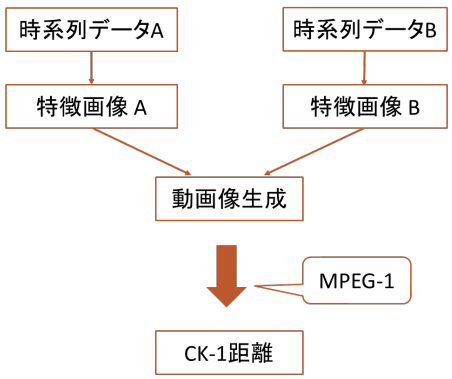

図 3：RPCD 手法の処理

(3)単語長を考慮した辞書間距離 WNMD

本報告書の冒頭で述べた NCD(Normalized Compression Distance)は、2 個のデータ x,y を結合したファイルを圧縮しそのファイルサイズから x と y の距離を計測する。しかし、NCD は類似度計算の度に大きなファイルを圧縮するためオーバーヘッドが大きい。そこで、データ x の圧縮辞書 Dx を x の特徴表現と見なし、 x と y の距離を圧縮辞書の非類似度から求める辞書間距離が近年出現した。辞書間距離の肝は、圧縮辞書 Dx は x 内に出現する単語集合なので、x の特徴量として有効であるということである。その一方で、圧縮辞書が元データから情報を捨てている事が辞書間距離の欠点である。例えば、圧縮辞書 Dx は各単語が x 内に何個存在したかという情報は持たない。そこで、圧縮辞書 Dx に各単語が x 内に出現した回数を記録して辞書間距離を求める Normalized Multiset Distance (NMD) が提案された。NMD は最先端の辞書間距離に位置付けられる。

本研究では、NMD が各単語の重要度を考慮していない点に着目し、各単語に重要度に応じて重み付けする Weighted NMD (WNMD) という手法を開発した。具体的には、n 文字の単語には $\sqrt{n}$ の重みを与え、長い単語ほど重要度を高くした。例えば画像処理の場合、1 文字の単語は 1 画素分の情報しか持たないが、5 文字の単語は 5 画素分の情報を持つのでより重要ということである。WNMD を類似画像検索に適用して評価した結果、適合率が表 1 に示すように NMD を上回り、単語の重要性を考慮する WNMD の妥当性を示せた。

表1: 全画像に対する平均適合率

NMD	WNMD
0.5086	0.5249

(4)再圧縮率を利用した PRDC

PRDC ではデータ x を複数の圧縮辞書による圧縮率を並べた圧縮率特徴ベクトルで表現する。しかし、圧縮率は x 内に圧縮辞書に含まれる単語が何回出現したかだけで決まり、x 内でどの単語が隣接しているかという情報を捨てている。そこで、本研究では x 内での単語の隣接関係から定まる新しい特徴量を圧縮後のファイルから抽出した。具体的には、x を圧縮辞書 D で圧縮した後のファイル(符号列) を再圧縮した時の圧縮率を特徴量として利用する。再圧縮率がデータ内の単語の隣接関係から定まる特徴であることを図 4 の例から述べる。図 4 は異なるテキスト x と y を再圧縮まで行った時の様子を示す。最下段が元のテキストであり，真ん中の段が辞書 D で圧縮後のファイル、最上段が再圧縮後の出力ファイルである。なお、再圧縮時には自己圧縮を行うため、x の再圧縮と y の再圧縮では異なる辞書が用いられる。辞書 D で x、y を圧縮後の符号化列はそれぞれ"123123", "132231" と異なっている。しかし，圧縮率はどちらも 6/16 であり、x と y は区別できない。これは符号語'1','2','3' の出現回数が同じであることが理由であり、単語の出現回数のみで決定される圧縮率の限界である。一方，x の再圧縮率は 2/6、y の再圧縮率は 4/6 と異なり、x と y を区別できる．この違いは x と y の単語順序の違いに起因する．例えば，x の再圧縮率は符号語列'123'が 2 回出現した事実を反映する．このように再圧縮率は単語の隣接関係から定まる。

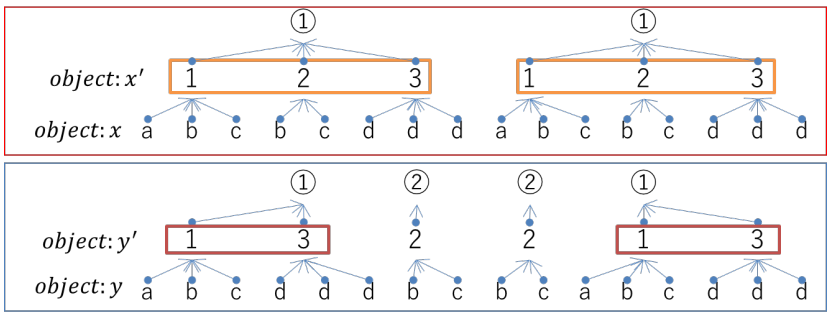

図 4: 圧縮と再圧縮

提案手法では、圧縮率と再圧縮率を n 個ずつ並べた 2n 次元のベクトルでオブジェクトを表す。この圧縮率ベクトルは、(1)通常の圧縮率により単語の頻度を考慮し、(2)再圧縮率により単語の隣接関係を考慮することができる。この性質から、提案手法を HOPRDC (High-Order PRDC，高階 PRDC)と名付ける。提案手法を画像認識に適用した結果、従来の PRDC より、正解率を 4%向上できた（表 2)。

表2: クラス分類の正解率

PRDC	HOPRDC
0.639	0.680

図５は Flower クラスのクエリ画像（図５における一番左上の画像）に対して、最も類似した上位 9 位の画像が HOPRDC と PRDC でどう変わったかを示している。類似検索結果はどちらの手法でも赤色の画像が多いが、HOPRDC の方がクエリと形状が似ている画像を多く見つけている。これは、HOPRDC では単語の出現順序まで考慮するため、構造情報を特徴ベクトルに畳み込めたた

めであると考えられる。

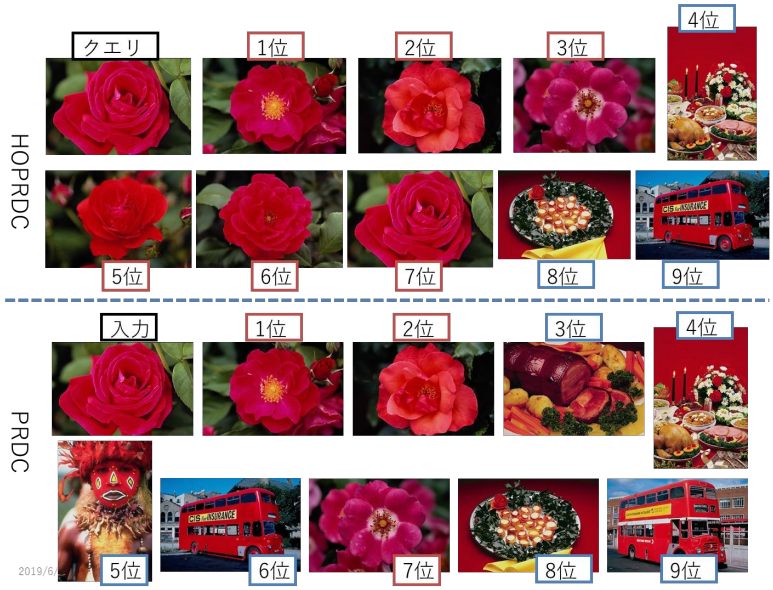

図 5: クエリと類似した画像

５．主な発表論文等

〔雑誌論文〕（計 3 件）

1. T. Yamazaki, H. Koga and T. Toda, "Fast Exact Algorithm to Solve Continuous Similarity Search for Evolving Queries", in Proc. Asia Information Retrieval Symposium (AIRS2017), springer LNCS 10648, pp.84-96, 2017. 査読有
   DOI: 10.1007/978-3-319-70145-5_7
2. T. Uchino, H. Koga, T. Toda, "Improved Compression-Based Pattern Recognition Exploiting New Useful Features ”, in Proc. Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2017), springer LNCS 10255, pp.363-371, 2017. 査読有
   DOI: 10.1007/978-3-319-58838-4_40
3. H. Koga, Y. Nakajima and T. Toda, "Effective Construction of Compression-based Feature Space", in Proc. International Symposium on Information Theory and Its Applications (ISITA2016), pp.116-120, 2016. 査読有
   URL: https://ieeexplore.ieee.org/document/7840397

〔学会発表〕（計 8 件）

1. 大内晶太, 古賀久志, "品質パラメータの学習による動画像圧縮技術に基づく時系列分類手法の改良", 第 39 回情報理論とその応用シンポジウム(SITA2018), 2018.
2. 村井建応, 古賀久志, 戸田貴久, "品質パラメータの学習による動画像圧縮技術に基づく時系列分類手法の改良", 第 39 回情報理論とその応用シンポジウム(SITA2018), 2018.
3. 藤原勇二, 古賀久志, 戸田貴久, "ユークリッド距離に基づく多観点非類似度とその分割最適化クラスタリングへの応用", 人工知能学会研究資料 SIG-FPAI-B509, pp.51-56, 2018.
4. 鈴木 聡, 古賀 久志, Gibran FUENTES PINEDA Gibran, 戸田 貴久, "共通要素を類似度とするハッシュベース集合間類似検索手法の改善",第 10 回データ工学と情報マネジメントに関するフォーラム(DEIM2018) 2018.
5. 藤原勇二, 古賀久志, 戸田貴久, "多観点類似度を用いた凝集型階層クラスタリング",第 16 回情報科学技術フォーラム(FIT2017), 第 2 分冊, pp.121-124, 2017.
6. 山崎智博, 古賀久志, 戸田貴久 "集合間類似度を用いたストリームデータの top-k 類似検索に対する高速な厳密解アルゴリズム", 信学技報 COMP2017-1, pp.1-9, 2017.
7. 内野太智, 古賀久志, 戸田貴久, "圧縮ベースパターン認識に有用な新しい特徴量の抽出", 第 39 回情報理論とその応用シンポジウム(SITA2016), 2016.
8. 板橋 大樹, 古賀 久志, Gibran Fuentes Pineda GIBRAN, 戸田 貴久,"共通要素数を重視したハッシュベース集合間類似検索",第 8 回データ工学と情報マネジメントに関するフォーラム(DEIM2016) 2016.

６．研究組織

(1)研究分担者　なし