Universal SAX：空間充填曲線を使用した SAX の多次元時系列データへの適用

理学専攻・情報科学コース　大西 史花

1　はじめに

センサーデバイスやシミュレーション技術の発展に伴い，GPS ログなどの軌跡データ，車の動きなどのムービングデータなど大量の多次元時系列データの入手が容易になっているが，これら膨大なデータの活用には効果的な問い合わせや索引が不可欠である．その中の 1 つに Lin らが提案した Symbolic Aggregate Approximation(SAX)[2] がある．SAX は時系列データを文字列に変換する．この手法はアルゴリズムが単純である一方で元データに対してマイニングを行った結果と同等の結果を生成することが出来るなど有用な特徴を多々持つ．しかしながら，SAX は 1 次元の時系列データしか扱うことができない．既存の手法では多次元時系列データの次元を重心距離法や主成分分析を用いて 1 次元に次元削減した後 SAX を適用しているが [3][4]，これらの手法はデータ間の位置関係を損ないマイニングに必要なデータを欠落させてしまう場合がある．

よって本稿では，空間充填曲線を用いて多次元空間内のデータの位置関係を損なうことなく文字列に変換する Universal SAX（USAX）を提案する．空間充填曲線とは，多次元空間を曲線で埋め尽くすように辿ることで空間の順序付けを行う手法である．本手法において文字列に変換後の各文字は元のデータ空間に適用した空間充填曲線による順序値の範囲によって割り当てられる．空間充填曲線による順序値の範囲は元空間において多面体を成すため，2 つの USAX 文字間の距離は 2 つの多面体間の最小距離として計算することが出来る．このように，USAX 文字列化後の距離測度として元の多次元空間上の距離を用いるので，本提案手法による多次元時系列データの 1 次元化は元データのマイニング等に有用な情報を欠落させることなく行うことが出来る．

2　SAX

SAX とは，時系列データの表現手法の 1 つで，時系列データを文字列に量子化する手法である．SAX による時系列データの文字列化を図 1 によって示す．図 1 では破線が SAX を適用する時系列データである．SAX を使用して時系列データを文字列に変換する手順は以下の通りである．まず時間軸を等間隔に区分し（縦実線），区間毎の平均値を算出する（横太線）．その後，正規分布の各面積が等しくなるような分割線（横実線）を定め，その区間に，a，b，c …とアルファベットを割り振る．最後に，求めた平均値を該当する区間に割り振られている文字に変換する．なお，時系列データは位置やスケールのズレを修正するためにあらかじめ正規化されている．

以上のような手順で変換される SAX 文字列は，本来実数値である時系列データに対して文字列用の検索・分析手法が適用できるという特徴を持つ．また，アルゴリズムが単純である一方で元データに対してマイニングを行った結果と同等の結果を生成することが出来る．しかしながら，SAX は 1 次元の時系列データしか扱うことができない．

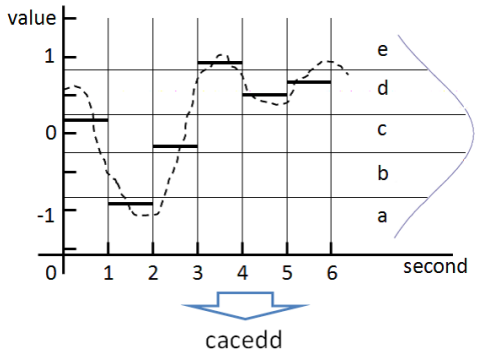

図 1: SAX による 1 次元時系列データの文字列化

3　Universal SAX

本節では多次元時系列データを文字列に変換する手法である Universal SAX を提案する．本提案手法では空間充填曲線という手法のひとつであるヒルベルト曲線を用いている [1]．空間充填曲線は多次元空間から 1 次元空間へのマッピング方法として広く使われている手法である．USAX 文字列の作成手順の概略を 2 次元データを例にして図 2 に示す．まず対象データをそれぞれの次元ごとで標準正規分布に正規化し，対象データ（黒実線）を $t_0, t_1, \cdots, t_n$ まで等間隔に区分，区間ごとの中央値 $x_0, x_1, \cdots, x_(n-1)$（黒点）を決定する（図 2 (1)）．次に多次元空間を $2^k \times 2^k$ の格子状になるよう量子化する（図 2 (2)）．その後ヒルベルト曲線を用いて正規化された対象データを区間に対応する数列 (この例では {51,46,40,38,27,28,10,17,15,1}) に変換，時系列状のデータとして表す（図 2 (3)）．最後に，各次元で行った正規化の領域が等しくなるよう分割を行い，それぞれの領域にアルファベットを割り振り（図 2 の (4) 左図）文字列に変換する（図 2 の (4) 左図下）．図 2 の (4) の右図は正規分布を分割した後の多次元空間のイメージである．

Universal SAX 文字列間の距離計算は，その文字列を構成する文字間の距離の合計として行われる．文字間の距離は，それぞれの文字が割り振られている空間上の領域間の最小距離として計算される．例えば図 2 の (4) の右図のように分割された空間上での文字 a と d の距離 $dist(\text{"}a\text{"},\text{"}d\text{"})$ は，マス目 1 つを距離 1 とすると 1 となり，文字 a と g の距離 $dist(\text{"}a\text{"},\text{"}g\text{"})$ は空間上で領域が隣接しているため 0 となる．このように，USAX は元の空間上の距離関係を用いて文字列データに変換しているため文字列変換時にデータ間位置情報の欠落が起きない．

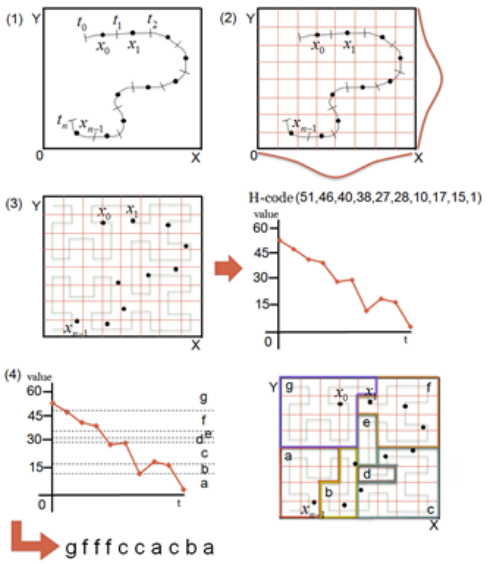

図 2: 2 次元データへの USAX の適用例

4 Universal SAX の有用性

本節では Universal SAX の有用性を示す実験結果について述べる．実験に使用したデータを図 3 に示す．このデータセットは疑似的に生成された 3 次元時系列データセットで，データの進む方向によって 4 種類に分類される．それぞれ {data0a, data0b}，{data1a, data1b}，{data2a, data2b}，{data3a, data3b} を同グループとして扱った．我々はこのデータセットに対して，元データと，既存手法を用い元データの次元を削減した後 SAX を適用した文字列データ，USAX を適用した文字列データの 4 種のデータで階層的クラスタリングの作成実験を行った．結果を図 4 に示す．赤線で囲ってあるデータは正解とみなした．この図から，唯一 USAX だけが元データと同等の結果を得ていることが分かる．これは実験に使用したデータが既存手法に不利なデータであるためと考えられる．この結果から，我々の提案手法である USAX が既存手法に比べてデータの特徴に左右されない汎用的で有用な手法であることが分かる．

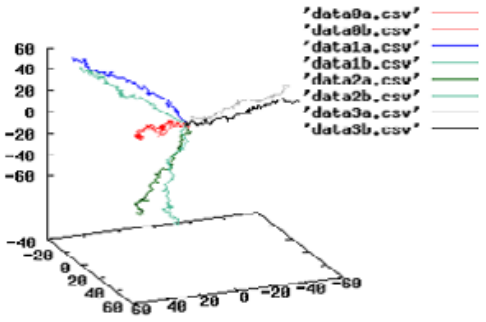

図 3: 疑似 3 次元時系列データの可視化

5 高次元データへの USAX の適用

多次元時系列データの文字列化を行う USAX では，データの高次元化に伴い必要な文字の種類が増大する．よって高次元のデータを扱う際に以下の 2 つの問題

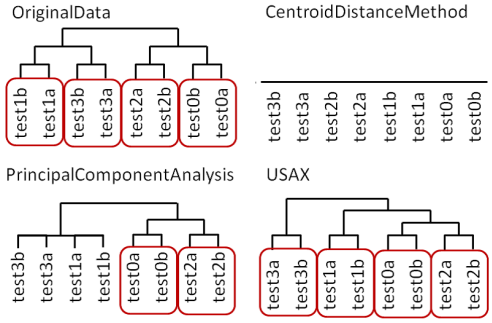

図 4: 階層的クラスタリングの実験結果

が生じる．1 つ目は，データの高次元化に伴い必要な文字の種類が増大する点である．この問題に対して，USAX では文字表現をバイナリ化することにより最大 $2^{12} = 4096$ の文字種類への対応を可能にしている．しかし，前述のように USAX 文字列化に必要な文字種類は次元が高くなるにつれ指数関数的に増大するため，高次元のデータには対応できない．2 つ目の問題は，高次元データの信頼性が次元の呪いにより低くなってしまう点である．次元の呪いとは．高次元のデータほどデータ間の距離が互いに等しくなってしまい汎化誤差が向上しなくなる現象のことである．つまりそもそも高次元のデータに対して直接 USAX を適用することは有効でないと考えられる．これらの問題を解決するために，我々は 5 次元以上の高次元データについて以下の手順を行うことを提案する．まず高次元元データに対して主成分分析を適用し， 4 次元以下に次元を削減する．その後 USAX を適用し 1 次元の文字列データに変換する．主成分分析と USAX の組み合わせにより，高次元データの位置関係情報を損なわずかつ 1 次元まで次元を削減できるようになる．

6 まとめ

本稿では，多次元空間内のデータの位置関係を損なうことなく文字列に変換する Universal SAX（USAX）を提案し，その有用性と展望を示した．今後は，提案手法の改良とそれに伴う評価実験を行っていく．多次元空間を区分するアプローチはヒルベルト曲線以外にも多くの種類があるため，それらの手法を検討し，より有効な変換手法を模索する．

参考文献

[1] D. Hilbert. : “Uber die stetige Abbildung einer Linie auf ein Flachenstuck,” In Math. Ann. 38, 1891, pp 459-460.

[2] Jessica Lin, Eamonn Keogh, Stefano Lonardi, Bill Chiu. : “A Symbolic Representation of Time Series, with Implications for Streaming Algorithms,” In SIGMOD workshop, 2003.

[3] Li Wei, Keogh, E. and Xiaopeng Xi. : “SAXually Explicit Images: Finding Unusual Shapes,” In the IEEE International Conference on Data Mining series (ICDM), 2006.

[4] Yoshiki Tanaka and Kuniaki Uehara. : “Discover Motifs in Multi Dimensional Time-Series Using The Principal Component Analysis And The MDL Principle,” In Workshop on Machine Learning and Data Mining in Pattern Recognition(MLDM), 2003.