DEIM Forum 2013 E1-3

SAX適用において利用者が注意すべき特性についての検討

松田　成美†　　大西　史花†　　渡辺知恵美†

† お茶の水女子大学理学部情報科学科　〒 112-8610　東京都文京区大塚 2-1-1
E-mail: †{g0920537,onishi.ayaka,chiemi}@is.ocha.ac.jp

あらまし　時系列データの表現手法はこれまで多くの研究者たちによって提案されてきた．その中の一つに， Lin らが提案した， Symbolic Aggregate Approximation （ SAX ） がある．SAX は有効な手法であるとして広く研究の場で使われているが，多様な時系列データ全てにおいて万能に適用できると限らない．SAX の利点を最大限に生かす為，利用者が注意すべき SAX の特性を明らかしたい．よって，本稿では SAX 適用時に起こりうる 2 つの問題点について指摘，検証を行う．1 つめに,「正規化された時系列データはガウス分布に従う」という前提条件は正しいのか，という点，2 つめに，実データと SAX 適用後のデータで，距離関係が逆転している可能性がある，という点である．検証の結果，前者は成り立たない可能性が高く，その結果後者のような影響が生じるということが分かった．

キーワード　時系列データ

Narumi MATSUDA†, Ayaka ONISHI†, and Chiemi WATANABE†

† Humanities and Sciences, Ochanomizu University　2-1-1 Ohtsuka, Bunkyo-ku, Tokyo, 112-8610 Japan
E-mail: †{g0920537,onishi.ayaka,chiemi}@is.ocha.ac.jp

1. はじめに

私たちの日常には，株価や医療，気象データのような，さまざまな時系列データがあふれている．また，スマートフォンなどの急速な普及により，一人一人が扱うデータ量も増え，その管理により一層の効率化が求められている．また，できるだけその情報を失うことなく正しく保持したまま，データ量を削減して処理を高速化したいという要求が高まっている．そのため，データ量の削減や処理時間短縮の面から索引の付与が不可欠となっている．

データに索引を付与するさまざまな表現手法は，これまで多くの研究者たちによって提案されてきた．その中の一つに， Lin らが提案した， Symbolic Aggregate Approximation （ SAX) [1] がある．SAX は有効な手法であるとして，広く研究の場で使われている．しかしながら，多様なデータ全てにおいて汎用性の高い手法を提案することは難しく，SAX がすべての時系列データにおいて万能に適用できると限らない．扱うデータによっては，常に正しいマイニング結果を得られない可能性もあるといえる．利用者が注意すべき SAX の特性がわかれば，SAX の利点を最大限に生かすことができるだろう．そこで本稿では， SAX 適用時に起こりうる問題点を指摘するため，以下の 2 点について検証する．

まずひとつめに,「正規化された時系列データはガウス分布に従う」という前提条件は正しいのか，という点である．SAX では，まずデータを正規化した後，文字列に変換するための操作を行う．文字列変換の際，変換後の各文字が等確率で出現するようにデータの分割区間を設定する．この分割点は,「正規化された時系列データはガウス分布に従う」という前提に基づき，ガウス曲線に従って決定している．しかしここで，正規化された時系列データは本当にガウス分布に従うのか，という疑問が出てくる．SAX 論文 [1] 中では，この前提条件が成り立つことを，実験結果を以て示している．しかしながら，これはわずか 8 種類の時系列データについて検証した結果に基づいた見解であり，他のデータでは成り立たない可能性がある．

ふたつめに，実データと SAX 適用後のデータで，距離関係が逆転している可能性がある，という点である．データ上の点 X, Y, Z があるとき，その距離関係が，実データでは $|XY| < |YZ|$ であるのに，SAX 適用後に $|XY| > |YZ|$ のように逆転してしまうことがある．このような距離の逆転現象は，階層的クラスタリングや k 近傍法のように順序関係を重視する手法では，逆転によって生じた誤差を必要以上に拡大してしまう可能性がある．

以上の可能性を検証するため，我々はいくつかの時系列データを用いて， SAX を適用した場合と元データでの結果を比較する．そして，先に述べた可能性がどういうときに起こりうるのかを調査する．この検証により，利用者が注意すべき SAX の特性がわかり，その利点を最大限に生かすことができる．本

稿では，SAX 適用時に起こりうる問題点についての具体例を用いた説明と，実験による検証結果について述べる．

本稿の構成は以下の通りである．第 2 節では本研究の研究対象手法である SAX について述べ，第 3 節では SAX を使用する際の問題点について述べる．第 4 節ではその問題が起こる場合の評価実験について述べ，そして第 5 節でまとめと今後の課題を述べる．

2. SAX

SAX とは， 時系列データを文字列に変換する手法である．同じく時系列データの表現手法として，離散フーリエ変換や離散ウェーブレット変換などがある．それら 2 つとは異なり，最終的に文字列に変換されるということが SAX の大きな特徴である．これにより，パターン検索など，自然言語処理分野における文字列処理アルゴリズムが適用可能となる．その他の特徴としては， 中間段階で Piecewise Aggregate Approximation (PAA) [4] によって次元削減を行うことが挙げられる．PAA は時系列データをある区間で平均化する手法であり，離散フーリエ変換や離散ウェーブレット変換などのような複雑な技術と比較すると直観的で単純である．それによって SAX 自体も単純なアルゴリズムとなっており，他の既存手法と比較してマイニングのしやすい有効な手法として，多くの研究分野で使われている．

2.1 SAX の適用手順

SAX の時系列データへの適用手順を以下に示す.

(1) データを平均 0，分散 1 の正規分布に従うように正規化する

(2) 正規化したデータを PAA 表現に変換する

i データの時間軸を等間隔に区分する (図 1(a))

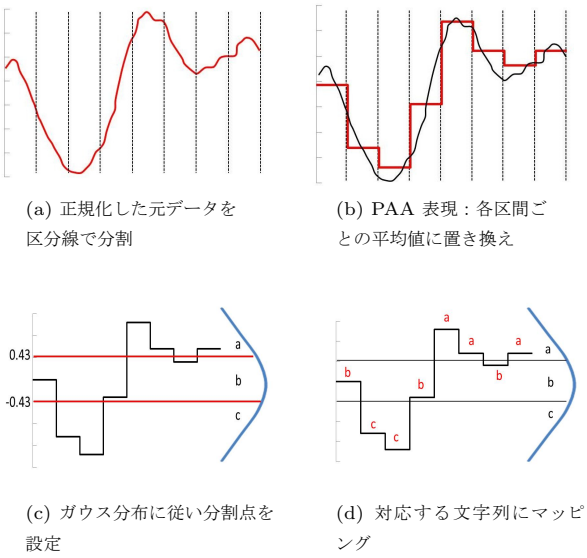
ii 各区間ごとに平均値を計算し，データの値をその平均値に置き換える (図 1(b))

(3) 文字列化

i 分割点を設定 (図 1(c))

ii 対応する文字列にマッピング (図 1(d))

たとえば，長さ $n = 128$ の時系列 $C = c_1, c_2, \cdots, c_{128}$ に SAX を適用したいとする．準備として，2 個のパラメータ値をあらかじめ設定しておく必要がある．変換後のデータの長さあるいは何個の文字に変換したいかを表す文字列長 w と，何種類の文字に変換するのかを表すアルファベットサイズ a である．ここでは，文字列長 $w = 8$，アルファベットサイズ $a = 3$ とする．つまり，変換後のデータは 3 種類のアルファベット 8 個からなる文字列となる．まず，この時系列データ C を平均 0，分散 1 の正規分布に従うように正規化する．次に，正規化したデータの時間軸を等間隔に 8 分割（ $w = 8$ ）する．図 1(a) で，赤線が正規化した元データ C，点線が区分線になっている．次に，各区間ごとに値を平均値に置き換える．これによって得られたデータは， $\bar{C} = \bar{c_1}, \bar{c_2}, \cdots, \bar{c_8}$ と表せる．ここでは，アルファベットサイズ $a = 3$ に従って分割点を設定し，それぞれの領域を a , b , c とする．(図 1(c)) そして，データを対応する文字列に置き換える．(図 1(d)) ここで得られた文字列は，$\hat{C} = \hat{c_1}, \hat{c_2}, \cdots, \hat{c_8} =$ bccbaaba と表せる．



(a) 正規化した元データを区分線で分割

(b) PAA 表現：各区間ごとの平均値に置き換え

(c) ガウス分布に従い分割点を設定

(d) 対応する文字列にマッピング

図 1 時系列データへの SAX 適用

分割点は，ガウス曲線のもとで各記号が等確率で出現するような領域に分割されるよう決定する．つまり，正規化された時系列はガウス分布に従うという前提条件に基づき，ガウス曲線の下の領域が同じ面積になるように分割する．したがって，分割点はアルファベットサイズ a によって決まっており，入力データに依存しない．図 2 のように，分割点の値は，アルファベットサイズに応じてあらかじめ計算しておくことができる．

β_i \ α	3	4	5	6	7	8	9	10
β 1	−0.43	−0.67	−0.84	−0.97	−1.07	−1.15	−1.22	−1.28
β 2	0.43	0	−0.25	−0.43	−0.57	−0.67	−0.76	−0.84
β 3		0.67	0.25	0	−0.18	−0.32	−0.43	−0.52
β 4			0.84	0.43	0.18	0	−0.14	−0.25
β 5				0.97	0.57	0.32	0.14	0
β 6					1.07	0.67	0.43	0.25
β 7						1.15	0.76	0.52
β 8							1.22	0.84
β 9								1.28

図 2 $a = 3 \sim 10$ のときの分割点の値 β_i

例えば，アルファベットサイズを 5 に設定した場合，4 つの分割点 $\beta_1 \sim \beta_4$ は，$(\beta_1, \beta_2, \beta_3, \beta_4) = (-0.84, -0.25, 0.25, 0.84)$ のように求めておくことができる．

2.2 データ間の距離

ともに長さ n の 2 つの時系列データ Q, C 間の距離の定義を，元データと SAX 適用後データそれぞれについて説明する．

まず，元データでの Q, C 間の距離 $D(Q, C)$ は以下の図 3 で表されるユークリッド距離であり，式 (1) で求められる．

$$D(Q, C) \equiv \sqrt{\sum_{i=1}^{n} (q_i - c_i)^2} \qquad (1)$$

また，SAX 適用後データでの Q, C 間の距離 $MINDIST(\hat{Q}, \hat{C})$ は図 4 で表される文字列間の距離であり，

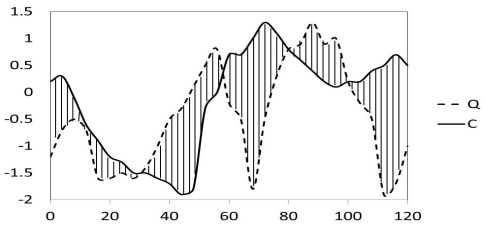


図 3　元データ間の距離（ユークリッド距離）

式 (2) で求められる．

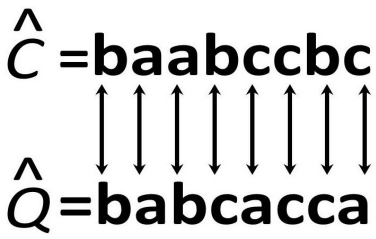


図 4　SAX 適用後データ間の距離（文字列間の距離）

$$MINDIST(\hat{Q},\hat{C}) \equiv \sqrt{\frac{n}{w}}\sqrt{\sum_{i=1}^{w}(dist(\hat{q}_i,\hat{c}_i))^2} \qquad (2)$$

変換後の 2 つの文字 $\hat{q}_i, \hat{c}_i$ 間の距離 $dist(\hat{q}_i, \hat{c}_i)$ は文字 $\hat{q}_i$ がとりうる値と文字 $\hat{c}_i$ がとりうる値の差が最少となるときの最少値で表わされる．たとえば図 5 で，文字 a と文字 d の距離は $dist(a,d)$ ，文字 a と文字 b の距離は $dist(a,b)$ ，文字 b と文字 d の距離は $dist(b,d)$ によって定義される．

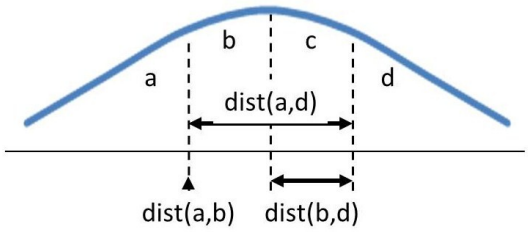
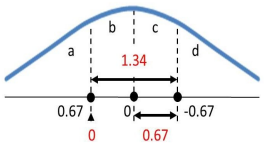


図 5　文字列間の距離 $dist(\hat{q}_i, \hat{c}_i)$

ここで第 2. 節で述べたように,「正規化した時系列はガウス分布に従う」という前提条件に基づいて分割点を決定していることに改めて注目する．分割点の値はデータに依存せず，分割数，すなわちアルファベットサイズに応じて決まっている（図**??**）ため，文字列間の距離もアルファベットサイズに応じて決まる．以下の図 6 でアルファベットサイズ $a = 4$ の時の例を示す．

3 つの分割点は (0.67 , 0 , -0.67) である．文字列間の距離は距離表のように決まっており，$dist(a,d) = 1.34$ ， $dist(a,b) = 0$ ， $dist(b,d) = 0.67$ となる．



(a) 文字列間距離

	a	b	c	d
a	0	0	0.67	1.34
b	0	0	0	0.67
c	0.67	0	0	0
d	1.34	0.67	0	0

(b) 距離表

図 6　アルファベットサイズ $a = 4$ のとき

3.　SAX の注意点の検証

SAX を時系列データに適用する際，その特徴から考えられる 2 つの可能性について指摘する．

3.1 「正規化した時系列はガウス分布に従う」という前提条件

SAX 論文中では，8 個の異なる時系列データから抽出した長さ 128 のサブシーケンスデータを用い，それぞれ正規化してガウス分布に従うか調べる実験を行っている．正規化したデータ値の累積分布の正規確率をプロットし，その視覚的特徴から用いたデータがガウス分布に従っていることを示している．その結果，全ての時系列データでも同じようにいえるとして議論を進めている．

正規化した時系列データがガウス分布に従っているという前提条件が成り立つことの利点は，文字列化の際，変換する文字の種類数に応じた分割点をあらかじめ決定できる，ということであるまた，分割点に応じてあらかじめ距離表を作成しておくことが可能であり，計算時間の短縮につながる．

しかしながら， SAX 論文中で用いたデータが偏った結果を導いている可能性を否定できない．つまり SAX 論文中で用いたデータ以外のデータでも同じような結果が得られるとは限らないということである．そこで他のデータで同じように正規確率がガウス分布に従うか検証し，従わない場合もあることを示す．

たとえば，同じような傾向を持つデータを考える．お茶の水女子大学情報科学科（お茶大情報科）1 年生 5 人分の，1 日のツイッターでのツイート数を時間ごとに表した時系列データ（図 7）があるとする．

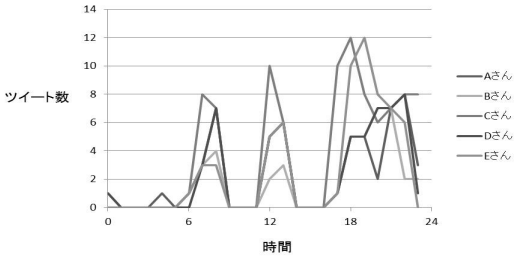


図 7　お茶大情報科 1 年生の 1 日のツイート数

この 5 人は通学，授業，睡眠など生活リズムがほぼ同じである．そのため，授業中や睡眠時間のツイート数は少なく，休み時間や放課後，夕方から夜にかけてのツイート数が多くなって

いるなど，グラフの傾向が似ていることがわかる．この 5 つの時系列データを正規化し，その頻度をヒストグラムで示した．（図 7）見た目から，ガウス分布に従っていないことがわかる．

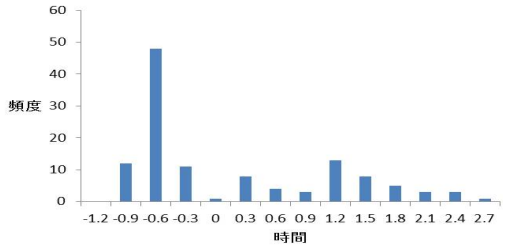


図 8　ツイート数の頻度分布

また，ガウス分布に従うことを数値的に確認するために，ジャック-ベラ検定を用いる．以下で，ジャック-ベラ検定について簡単に説明する．

ジャック-ベラ検定

ジャック-ベラ検定とは，統計学において標本データがガウス分布に従う尖度と歪度を有しているかどうかを調べる適合度検定である．検定統計量 JB は，以下のように定義される．

$$JB = \frac{n}{6}\left[S^2 + \frac{1}{4}(K-3)^2\right]$$

N：標本数, S：歪度, K：尖度

もし標本分布がガウス分布であれば，この検定統計量は自由度 2 のカイ二乗分布に従う．有意水準が 5 %として，この数値が大体 6 よりも小さければ，ガウス分布と見なして構わない．

お茶大情報科 1 年生の 1 日のツイート数データを正規化したデータにジャック-ベラ検定を適用してみると，$JB = 16.9 > 6$ という結果が得られた．従って，このデータがガウス分布に従っていないことが数値的にも示された．

我々は，この例のように，データの傾向が同じであるデータセットを使用する場合,「正規化した時系列はガウス分布に従う」という前提条件が成り立たないのではないかと予測している．

3. 2　距離の逆転現象

第 2. 節で述べたとおり， SAX による符号変換はガウス分布の等面積分割によって行われる．この時，それぞれの領域の幅は中心が密，外側が疎となる．例えば図 9 のようなデータ上の 3 点 X, Y, Z を考えると，実際の距離関係は明らかに $|XY| < |YZ|$ である．しかし， SAX 適用後では X は文字 b ， Y は文字 d ， Z は文字 e に変換されるため，$dist(b,d)$, $dist(d,e)$ によって求められる．2. で述べた距離表に従ってこれを計算すると，XY 間の距離は $dist(b,d) = 0.5$ ， YZ 間の距離は $dist(d,e) = 0$ となる．従って $|XY| > |YZ|$ となり，距離関係が逆転してしまう．

このような逆転現象はクラスタリングの精度に悪影響を与えると考える．以下に簡単な例を示す．

下の図 10 のような 4 つのデータからなるデータセットがあるとする．元データでの距離関係が， data1 , data2 間の距離

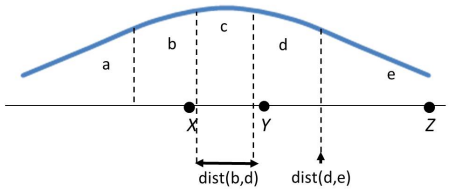


図 9　距離の逆転現象：距離関係が，実データでは $|XY| < |YZ|$ であるのに，SAX 適用後では $|XY| > |YZ|$ のように表されてしまう

のほうが data3 , data4 間の距離より大きいとする．しかし，SAX 適用後に距離を測定するとき，このデータ上のある地点で，先ほど述べたような距離の逆転現象が起こる，ということが頻繁に起こったとする．このとき，データ全体同士での距離関係が，SAX 適用後， data1 , data2 間の距離より data3 , data4 間の距離のほうが大きくなり，データ間の距離の逆転が起こってしまう．

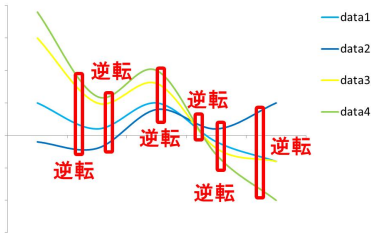


図 10　元データでは (data1,data2 の距離)>(data3,data4 の距離) だが， SAX 適用後， (data1,data2 の距離)<(data3,data4 の距離) となり，データ間の距離の逆転現象が起こっていることがわかる

このようなデータセットに，階層的クラスタリングを行う．階層的クラスタリングは，もっとも距離の近いデータをグループ化していく，ということを繰り返すようなクラスタリング法である．元データでは図 11 のような結果になるのは明らかである．しかし， SAX 適用後のデータでは，元データと異なる結果となってしまう．

図 11　クラスタリング結果

図 11 のようなクラスタリング結果の違いが，距離の逆転現象によるものなのか，また，距離の逆転現象がどういうときに起こるのかを，今後検証する．そのための実験方法について第 4. 節で述べる．

4. 実　　験

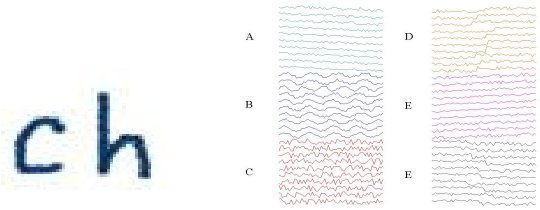
実験のために，以下の 2 つのデータを用いる．

データ 1　筆圧データ [2]

UCI の Machine Learning Repository 中の Character Trajectories Data Set である．(図 12(a)) このデータは同一人物から得た 2858 個の筆跡データセットで，ペンタブレットによる入力を軌跡と筆圧からなる 3 次元のデータとして取得したものである．本実験ではこの中から筆圧のデータだけを取り出して，1 次元のデータとして扱うことにする．

データ 2　Control Chart [3]

Robert Alcock（1999）により生成された 600 の合成データである．(図 12(b)) 100 ずつ，ノーマル，サイクリック（周期的），増加傾向，減少傾向，上方移動，下方移動の 6 種類に分類されている．本実験では簡単のため，6 種類から 17 個ずつ，計 102 個のデータを取り出して扱った．



(a) 筆圧データ　　(b) ControlChart

図 12　デ ー タ 例

4.1 「正規化した時系列はガウス分布に従う」という前提条件の検証

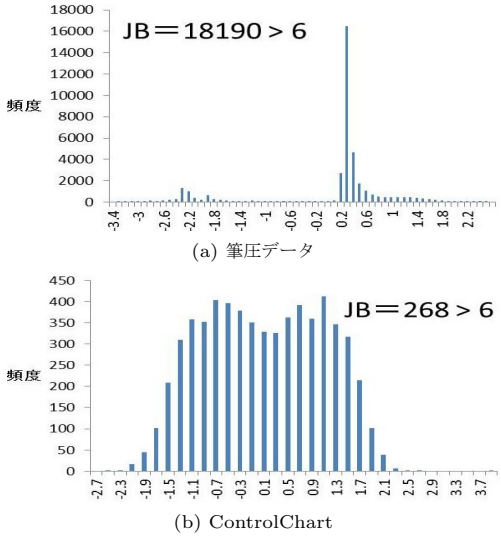
それぞれのデータにおける値の頻度分布を調査する．調査方法は，ヒストグラムを作成し，視覚的に確認する方法と，3.1 で述べたジャック-ベラ検定を用いて，数値的に確認する方法の 2 つである．結果は図 13 のようになり，ガウス分布に従わないことが分かる．

4.2　距離の逆転現象が起こる場合と確率の検証

まず定性的分析として，距離の逆転現象が起こる理論確率を求めた．この理論確率とは，4 つの時系列データにおけるある 1 時点を考えたとき，図 14 の例のような逆転現象がデータ全体で起こる確率のことである．SAX 適用後の文字種類数が 50 個であるとき，その確率は 0〜0.87 %であった．この程度であれば，逆転が起こっていてもデータ全体には大きな影響は与えないといえる．

次に定量的分析として，データを用いて距離の逆転現象が起こる実際の確率を求めた．実データ間の距離と SAX 適用後データ間の距離それぞれの測定は，タイムワーピングに基づいて行った．その結果，逆転が起こる確率は表 1 のようになった．

この結果より， Control Chart では実際の計算値による確率と理論値で大きな差がないことが分かる．従って距離の逆転現象が起こる確率は非常に低く，データへの影響は少ないと考え



(a) 筆圧データ

(b) ControlChart

図 13　ヒストグラムとジャック-ベラ検定結果

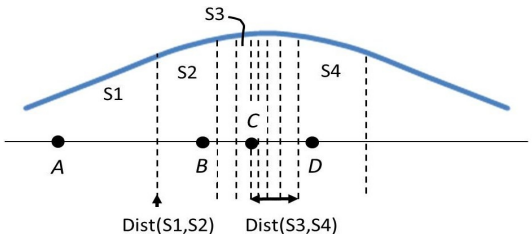


図 14　実距離関係は $|AB|<|CD|$ だが，SAX 文字列間距離は逆になっている

られる．しかしながら，筆圧データでは 44.8 %となり，半分近くが逆転しているという結果になっている．

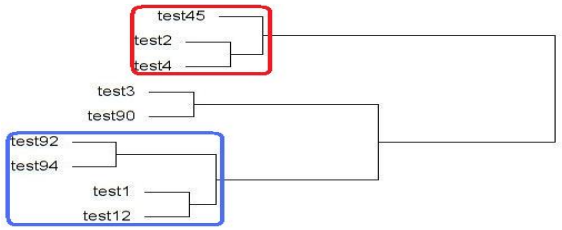
	逆転現象が起こる確率
理論値	0〜0.87 %
筆圧データ	44.80 %
Control Chart	0.10 %

表 1　距離の逆転現象が起こる確率

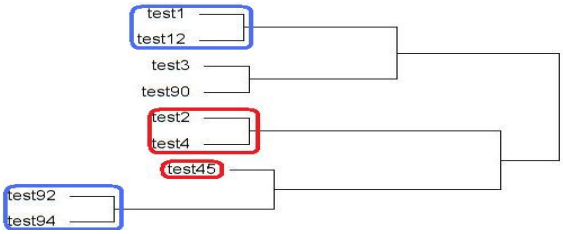
この原因として考えられる可能性として，実際はガウス分布に従っていないデータに対し，従っているものとみなしてＳＡＸを適用したため，ということが考えられる．SAX 適用時の分割点の設定は，4. でも指摘したように，ガウス分布に従うという前提条件に基づいて行う．従って，前提条件が成り立たないデータへの SAX 適用では，本来なら細かく分割されるべき部分の分割が大まかに分割されるなど，分割点の設定が不適当であった可能性がある．

また，筆圧データに階層的クラスタリングを適用した結果は図 15 のようになった．簡単のため，9 データ（ test1，2，3，4，12，45，90，92，94）のみをクラスタリングした．元データを基準とし， SAX 適用後に結果が異なっている部分を見ていく．test2，4 と test45 に着目すると，元データでは階層の下位でクラスタリングされているが，SAX 適用後では，かなり上位でクラスタリングされている．また， test92，94 と test1，12

も同様である．このように，SAX 適用によって距離の逆転現象が頻繁に起きた場合，クラスタリング結果にも影響を与えてしまうことが分かった．



(a) 元データ



(b) SAX 適用後データ

図 15　筆圧データにおけるクラスタリング結果：元データと SAX 適用後データでクラスタリング結果が異なっている

5. まとめと今後の課題

本稿では， SAX の概要について説明し，SAX 適用時に起こりうる問題点について具体例とともに説明した．また，扱うデータによっては，クラスタリング結果が実際と大きくかけ離れてしまう可能性を指摘した．実験の考察より，「正規化したデータがガウス分布に従う」という前提条件は成り立たない可能性が高く，それが距離の逆転現象を引き起こす原因となる可能性が高いといえる．従って SAX 適用時には，正規化したデータがガウス分布に従っているか否かを調べ，従っていない場合には分割点の設定方法を変えるなどの考慮をする必要があるといえる．今後の課題として，他の多様なデータにおいて距離の逆転現象が起こる確率を求め，確率が高くなるときのデータの傾向を分析したい．また，クラスタリング結果へ大きく影響を与える時の傾向についても分析したい．最終的には，正規化したデータがガウス分布に従わない場合の分割法について，有効な手法を提案していきたい．

文　献

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

[2] Ben H Williams, UCI Machine Learning Repository. : http://archive. ics. uci. edu/ml/index. html, February 2012

[3] Dr Robert Alcock, Synthetic Control Chart Time Series. : http://archive. ics. uci. edu/ml/databases/synthetic_control/

[4] Yi, B, K., and Faloutsos, C. : Fast Time Sequence Indexing for Arbitrary Lp Norms. In proceedings of the 26st VLDB Conference, Cairo, Egypt, 2000.