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　(指導教員：渡辺知恵美)

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おり，その管理により一層の効率化が求められている．効率化のために，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．

このような手法を実現するために，これまで多くの研究者たちがさまざまな表現手法を提案してきた．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，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．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．そこで本研究では、SAX の適用時に起こりうる問題点の指摘と実験での検証結果について述べる．

2　SAX

2.1　SAX の適用手順

SAX とは，時系列データを文字列に変換する手法である．SAX の時系列データへの適用手順を以下に示す．

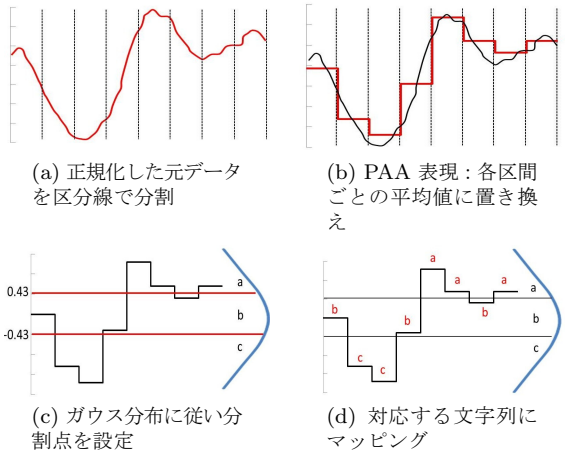
(1) データを平均 0，分散 1 の正規分布に従うように正規化する

(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))



(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

2 個のパラメータ値である，変換後文字列長 w，文字種類数 a はあらかじめ設定しておく．例えば，長さ $n = 128$ の時系列 $C = c_1, c_2, \cdots, c_{128}$ に $w = 8$，$a = 3$ として SAX を適用すると，$\hat{C} = \hat{c_1}, \hat{c_2}, \cdots, \hat{c_8}$ = bccbaaba と表せる．

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．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．

2.2　データ間の距離

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 間の距離は以下の図 2 で表されるユークリッド距離である．また，SAX 適用後データでの Q, C 間の距離は図 3 で表される文字列間の距離である．

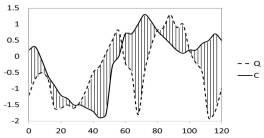


図 2: 元データ間の距離

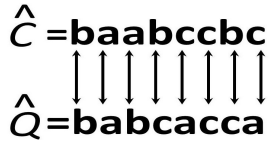


図 3: SAX 適用後のデータ間の距離

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 4 で，文字 a と文字 d の距離は $dist(a, d)$ によって定義される．

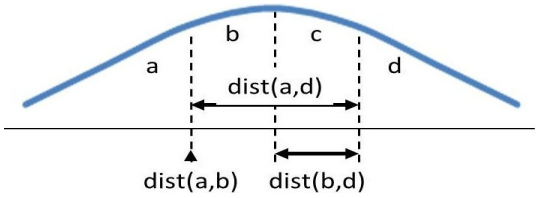


図 4: 文字列間の距離 $dist(\hat{q_i}, \hat{c_i})$

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．

3　SAX の注意点の検証

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をそれぞれ正規化し，ガウス分布に従うことを示している．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．

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．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する．

3.2 距離の逆転現象

例えば図 5 のようなデータ上の 3 点 X, Y, Z を考える．実際の距離関係は明らかに $|XY| < |YZ|$ である．一方 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，距離が $dist(b,d)$, $dist(d,e)$ によって求められる．すなわち XY 間の距離は $dist(b,d) = 0.5$，YZ 間の距離は $dist(d,e) = 0$ となる．従って距離関係は $|XY| > |YZ|$ となり，実データと SAX 適用後データでの距離が逆転してしまう．

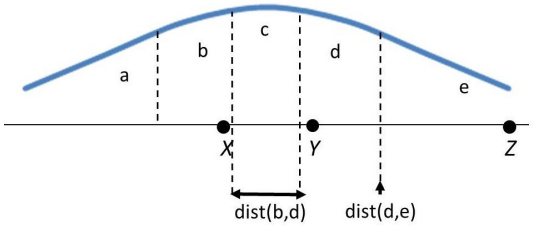


図 5: 距離の逆転現象 ： 距離関係が，実データでは $|XY| < |YZ|$ であるのに，SAX 適用後では $|XY| > |YZ|$ のように表されてしまう

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．

4 実験結果と考察

実験で使用したデータは Control Chart データセット [2] である．このデータは，100 例ずつ 6 つの傾向に分類されたデータから成るデータセットである．今回はこの中から 102 例を取り出して使用した．

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，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．

ここで，ジャック-ベラ検定とは，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．

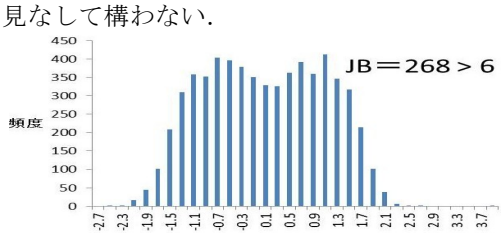


図 6: ヒストグラムとジャック-ベラ検定結果

結果は図 6 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 7 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．

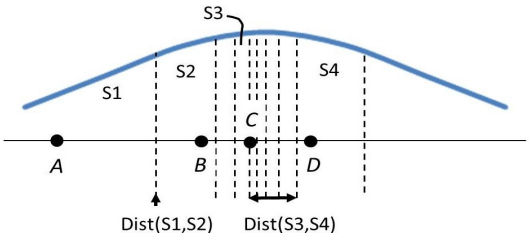


図 7: 実距離関係は $|AB| < |CD|$ だが，SAX 文字列間距離は逆になっている

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 0.10 %であった．

表 1 より，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られる．

	逆転現象が起こる確率
理論値	0〜0.87 %
実際の計算値	0.10 %

表 1: 距離の逆転現象が起こる確率

5 まとめと今後の課題

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，SAX 適用時に考慮する必要があるといえる．また，距離の逆転現象が起こる確率はかなり低いが，データに依存する可能性を否定できない．従って今後の課題として，距離の逆転現象が起こる確率の定量的分析で，筆圧データなどの偏った傾向を持つデータにおいても同じ結果が得られるかを検証したい．

参考文献

[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] $Indexof/ml/databases/synthetic_control$: http://archive. ics. uci. edu/ml/databases/synthetic_control/