# 時系列データ次元圧縮方式の評価および業務適用性の考察

佐藤　重雄† 高山　茂伸† 東　辰輔† 藤森　敬悟† 早川　孝之† 白井　健治†
三菱電機株式会社 情報技術総合研究所†

## 1. はじめに

プラントなどの現場においては、センサにより一定間隔の時間で測定された大量のデータが存在する。このような時間の経過とともに値の変化するデータ（時系列データ）に対して、一定期間での値の推移の類似性に着目することにより、時間による規則性の検出、将来の予測を行うことが可能となる。時系列データの類似検索処理では、類似度判定のために膨大な計算が必要となるため、一般には次元圧縮手法により計算量を削減することが研究されている[1]。

本稿では、従来の次元圧縮手法及び新たに提案する手法を、実測データに適用した場合について評価し、特徴、適用可能性について述べる。

## 2. 次元圧縮手法

時系列データ $X = x_1,...,x_n$ と $Y = y_1,...,y_n$ の類似度を表す指標として、n 次元空間上でのユークリッド距離が用いられる[1]。ここで、$X$ と $Y$ のユークリッド距離は以下で定義される。

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

ある時系列データに対して、蓄積されたデータから類似度の高いデータを検索するためには、すべてのデータに対するユークリッド距離計算が必要となる。この計算量を削減するため、従来より以下の次元圧縮手法が提案されている[1]。

- SVD(Singular Value Decomposition)
- PAA(Piecewise Aggregate Approximation)

これらの手法は次のような特徴を持つ。

- PAA は SVD と比較して次元圧縮アルゴリズムでの計算負荷が少ないため、次元圧縮処理に要する時間が短い。
- 次元圧縮空間で得られる候補解の中で、実際の検索結果となる割合は SVD の方が良い。これは、次元圧縮後に検索解を検出する処理性能は SVD の方が良いことを意味する。

本稿では、PAA と SVD を組み合わせた新たな手法（以降 PAA+SVD）を提案し、その特性の評価も行う。PAA+SVD は、次の手順で次元圧縮を行う。

(1) 時系列データを等間隔に分割し、PAA により次元圧縮を行う（第 1 段階の次元圧縮）
(2) PAA の結果得られたデータに対して SVD を実行することにより更に次元圧縮を行う（第 2 段階の次元圧縮）

## 3. 次元圧縮手法の適用性評価

### 3.1 評価内容

蓄積された時系列データベース内から類似した時系列データを検索する処理において、次元圧縮の各手法の評価を実施した。

評価を行った次元圧縮アルゴリズムは、SVD、PAA、PAA+SVD の 3 種類である。上記 3 種類のアルゴリズムについて、以下の二つの側面から適用性の評価を行った。

- ◆ 次元圧縮性能
  時系列データから各アルゴリズムに従って次元圧縮データを作成する時間の評価。
- ◆ 検索性能
  次元圧縮空間で得られる候補解に対する検索解の割合（ヒット率と定義）の評価。類似検索の種類は range queries を用い、時系列データ間の距離が指定された距離以内にあるデータの検索処理を実行した。

評価内容を表 1 に示す。評価対象データはセンサ等で測定された実際の時系列データである。なお、PAA+SVD では、PAA を適用する間隔を 4、8 の二通りにした測定を実施した（検索時系列長が 128 の場合は PAA 適用間隔 4 のみ実施）。

### 3.2 次元圧縮性能

各データの次元圧縮処理時間を図 1 に示す。

### 3.3 検索性能

各データでのヒット率を図 2、図 3 に示す。

表 1　評価内容

| 項目 | 内容 |
|---|---|
| 対象データ規模 | データ 1：1 時間毎 20000 時間<br>データ 2：1 時間毎 8760 時間 |
| 検索時系列長 | 3 種類（128/256/512） |
| 検索対象時系列 | 2 種類（ID1、ID2） |
| 圧縮後の次元 | 8 に固定 |

Evaluation of Dimensionality Reduction in Time Series Database, and consideration of applicability for Business Data
† Shigeo Sato
† Shigenobu Takayama
† Shinsuke Azuma
† Keigo Fujimori
† Takayuki Hayakawa
† Kenji Shirai
†Mitsubishi Electric Corporation Information Technology R&D Center

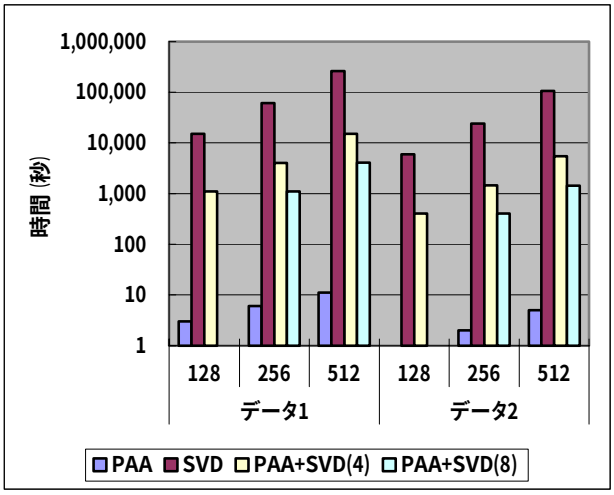


図 1：次元圧縮処理時間の比較

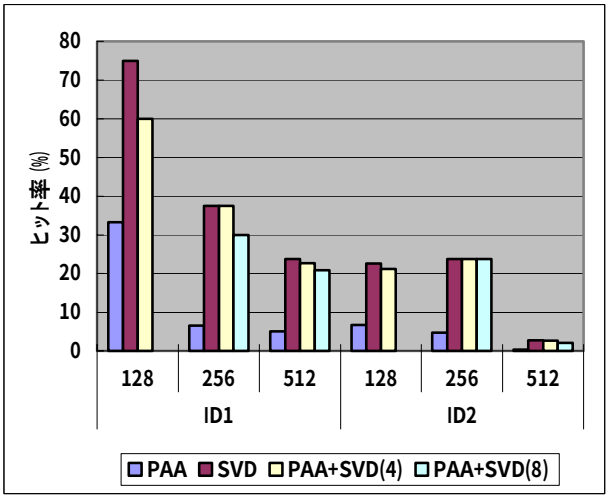


図 2：ヒット率の比較（データ 1）

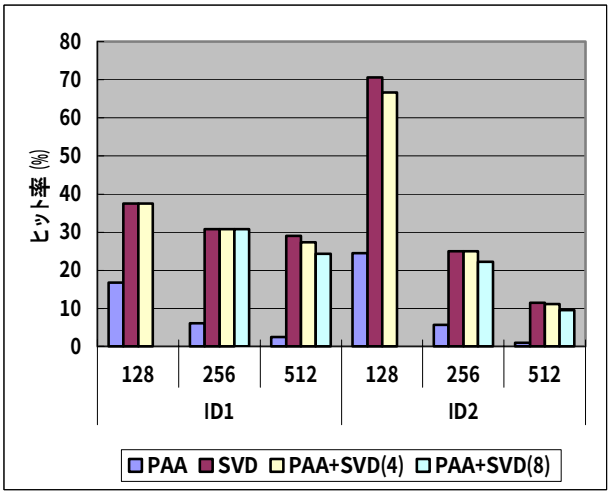


図 3：ヒット率の比較（データ 2）

（測定環境）

OS：Windows2000 Professional

PC：Pentium4 1.6GHz／Memory 640MBytes

## 4. 考察

次元圧縮処理時間は、アルゴリズムの特性により、PAA が最も短い。特に、検索時系列長が長い場合、SVD は時系列長の自乗に比例して次元圧縮時間が増加するため差が顕著になる。

一方、検索性能を決める要因となるヒット率は、PAA が最も悪く、SVD と PAA+SVD はほぼ同等の結果が得られた。PAA では、区間の平均値を求めて次元圧縮を行うため、時系列データの特性が失われる場合があり、類似しないデータを候補解に含める可能性が高いことが原因と推測される。特に、検索時系列長が長い場合は、平均値を計算する区間も長くなり、この傾向が顕著になると考えられる。

また、従来手法の組み合わせである PAA+SVD では、検索時系列長が長い場合の次元圧縮処理時間を SVD よりは抑えつつ、かつ、ヒット率を SVD とほぼ同等に保てる可能性があることが得られた。

本評価で得られた結果より、次元圧縮手法の適用に関する指針として次のことが考えられる。

◆ 類似検索期間（検索時系列長）の指針

週単位あるいは月単位のように長期間での測定データの類似性把握の目的では、次元圧縮データ作成時間を考慮して PAA 又は PAA+SVD の適用を考える。

◆ 検索時系列の特性に関する指針

類似パターンが比較的多いと予想される場合は、ヒット率を重視して、SVD 又は PAA+SVD の適用を考える。通常は安定して運転されているが、稀に発生する特異なデータの類似性を検索し異常検知を行う場合は、類似データの総数は少ないと考えられるため、ヒット率よりは次元圧縮時間を考慮した PAA の適用を検討する。

## 5. おわりに

本稿では、3 種類の次元圧縮処理方式について、次元圧縮性能、検索性能の測定を行い、実データへの適用性について評価した。

今後は、対象データを拡張した評価を実施するとともに、今回の評価で得られた結果から新たな次元圧縮の方式を提案し、評価を行っていく予定である。

## 参考文献

[1] E.Keogh, K.Chakrabarti, M.Pazzani and S.Mehrotra: Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases, Knowledge and Information Systems Journal(2001)