1C1-5

半正定性に基づく高速ユークリッド距離推定手法

Fast Estimation of Euclidean Distance based on Positive Semi-Definiteness

城戸 健太郎[*1] 桑島 洋[*1] 中西 耕太郎[*1*2] 鷲尾 隆[*1]
Kentarou Kido Hiroshi Kuwajima Kotarou Nakanishi Takashi Washio

[*1]大阪大学産業科学研究所知能システム科学研究部門高次推論方式研究分野
Department of Advanced Reasoning , Division of Intelligent Systems Science , Osaka University.
The Institute of Scientific and Industrial Research

Under the development of ubiquitous sensing and its associated technologies, data sets consisting of high dimensional and massive instances have become available in various practical fields. Efficient evaluation of the similarity measures such as Euclidean distance among the instances is one of the most important tasks for the instance queries and clustering. Under this circumstance, we proposed a fast approach to estimate the similarity measures among massive instances from their small portion by using a mathematical constraint called "positive semi-definiteness." However, this approach can not be directly applied to the problem to estimate Euclidean distance which is not positive semi-definite. The objective of this study is to investigate some approaches to transform the Euclidean distance to a positive semi-definite similarity measure and to apply the above estimation method. The performance of each transformation has been characterized through the evaluations based on artificial data sets.

1. はじめに

近年, ユビキタスセンシングを始め, 様々な技術が発達する中で, 大規模次元かつ大量事例データを取り扱う必要性が増大してきた. これらベクトルデータ間の類似性を何らかの類似性尺度で評価する手法は, 検索, クラスタリング, 分類などの種々のデータ解析手法の基礎となる. しかし,n 個の事例間の類似性尺度評価には n(n-1)/2 の計算時間複雑性を要し, 事例数の増加に伴い計算時間が膨大な量となる. 特に, 連続値を引数とするカーネル関数を離散近似で構成する際には, 非常に膨大な離散値の組み合わせについてカーネル関数の値を計算する必要がある. また, 測量や化学反応の速度など, 物理的に何らかの測定や実験を行わなければ得られない類似性尺度の場合は, その収集コストが膨大となる. 以上の問題を軽減する方法として, 範囲問い合わせ, 行列補完, カーネル関数推定などの技術が存在する.

しかし, 範囲問い合わせ (range query) を用いれば, 計算時間を大幅に削減することは可能であるが, 検索範囲にない事例との類似性尺度を知ることができないため, 完全性は失われる [CNBM 99][Yianilos 93][DN 87]. 一方で, 行列補完とは何らかの性質を満たす行列の未知要素値を他の既知要素から推定補完する問題である [Laurent 01]. 例えば, 半正定性を満たす相関係数行列や, 半負定性を満たすユークリッド距離行列の未知要素値を実際に直接計算することなく推定することができる. 多項式時間で半正定行列を任意の精度で推定補完する方法が知られているが, 計算複雑性は少なくとも行列の要素数 $O(n^2)$ を超える. そのため, 大規模データにおいては, 現実的な手法ではない. カーネル関数の推定に関しても研究が行われているが, 半正定行列補完を用いたものや, カーネル関数に関する補助データに EM アルゴリズムを適用して反復計算を行うことで推定するものであり, いずれの計算複雑性も $O(n^2)$ を超えてしまう [Graepel 02][TAA 03].

以上の背景から, 我々の研究室では一部事例間の類似性尺度値から他の事例間類似性尺度を高速かつ十分な精度で推定する手法の研究に取り組んでいる [Kuwajima 07]. しかしながら, この手法は類似性尺度行列が半正定であるものを対象としている. 代表的類似性尺度の代表的なものとして, ユークリッド距離があげられるが, ユークリッド距離は半負定であるため, 上記の手法を直接適用することは不可能である. そこで, 本稿では, 上記の手法を高速かつ十分な精度でユークリッド距離尺度の推定にも適用できるように, ユークリッド距離尺度の半正定類似性尺度への変換を導入する. そして, 種々のデータ解析を通じて, 導入する変換に対するユークリッド距離尺度の推定性能の依存性を明確化する.

連絡先: 大阪大学産業科学研究所
〒 567–0047 大阪府茨木市美穂ヶ丘 8-1
E-mail: k-kido@ar.sanken.osaka-u.ac.jp
*2 株式会社日本総合研究所所属

2. ユークリッド距離尺度の変換

本稿では, 類似性尺度行列を, 事例 i と事例 $j(i, j = 1, 2, ..., n)$ の間の類似性尺度値を第 ij 要素に持つ $n \times n$ 行列とする. ここでは類似性尺度行列として, ユークリッド距離 d_{ij} を要素に持つユークリッド距離行列 $D = (d_{ij})$ を取り上げる. 相関係数行列は半正定行列であるが, ユークリッド距離行列は半負定行列である. 半正定行列とは, 任意の非零ベクトル x に対して,$x^T Ax$ が常に 0 以上となる行列 A である. 一方, 半負定行列とは, 任意の非零ベクトル x に対して,$x^T Ax$ が常に 0 以下となる行列 A である. ただし, これら 2 つの行列の要素には以下に述べる関係性が存在する [Laurent 98].

2.1 半正定行列要素と半負定行列要素の対応関係（1）

$n \times n$ の半負定行列を $n \times n$ の半正定行列に変換するための各行列要素の 1 つの対応式を導入する. ユークリッド距離行列 D の i 行 j 列目の要素を d_{ij} とし, 半正定行列 P の i 行 j 列目の要素を p_{ij} とすると, これら 2 つの行列の要素には以下の関係を与える [Laurent 01].

$$p_{ii} = d_{i,n+1}$$
$$p_{ij} = \frac{1}{2}(d_{i,n+1} + d_{j,n+1} - d_{ij})$$

ただし,$d_{i,n+1}$ は i 番目の点と $n+1$ 番目の点のユークリッド距離を示す. ここで, 扱っているユークリッド距離行列は n 個の点間の距離しか表さないため, 任意の $n+1$ 番目の点と他の点との距離を付加する必要がある. 本稿では, この $n+1$ 番目

の点を既存の n 個の点の重心となるように設定する．この方法を「重心参照変換法」と呼ぶことにする．

2.2 半正定行列要素と半負定行列要素の対応関係（2）

$n \times n$ の半負定行列を $n \times n$ の半正定行列に変換するためのもう一つの各行列要素の対応式として，2つの行列の要素に以下の関係を導入する．

$$p_{ij} = \exp(-\alpha d_{ij})$$

ただし,α は $\alpha > 0$ となる任意の実数である．この方法を「指数変換法」と呼ぶことにする．

3. 半正定行列の推定手法の概要説明

半正定行列の推定手法 [Kuwajima 07] では, 本質次元の基底となる代表的な事例を選択し, 基底とその他の事例間の類似性尺度値を用いて, 全ての事例間の類似性尺度値を推定する．ここで, 本質次元とは, 類似性尺度空間上のデータ分布が有する実質的な広がり空間次元であり, 基底とは, その空間次元を表現する代表的な事例である．つまり, 選択した基底数を k とすると,$kn - k(k+1)/2$ 個の類似性尺度値を用いることで, 残りの事例間の類似性尺度を推定する．推定方法としては, 類似性尺度行列の半正定性と修正コレスキー分解を用いることで, 未知の類似性尺度値を任意の精度で推定することが可能となる．以下, その原理を簡単に説明する．

3.1 半正定類似性尺度行列の対角成分正規化

半正定行列の推定手法では, 推定対象である行列の対角要素を正規化し, 推定を行う．そのため, 本手法においてもユークリッド距離行列を半正定行列に変換後, 正規化を行う必要がある．ただし,2.2 節における変換では, 変換後の対角要素は自動的に 1 になるため, 正規化の必要はない．また,2.1 節の変換を用いる場合は, 変換後の各要素 d_{ij} を $\sqrt{d_{i,n+1}}\sqrt{d_{j,n+1}}$ で割ることで対角要素の正規化を行う．

3.2 修正コレスキー分解

修正コレスキー分解は, 任意の正則な行列を上三角行列と下三角行列に分解する手法である LU 分解を対称行列に特化した手法である. 修正コレスキー分解によって任意の対称行列 A は

$$A = LDL^T$$

のように下三角行列 L と固有値を対角要素に持つ対角行列 D に分解可能である. 修正コレスキー分解では, 元の行列 $A = A^{(1)}$ から最終的な解である上三角行列 L^T と固有値を要素に持つ対角行列 D をまとめた行列 $A^{(n)}$ を求める.$A^{(n)}$ の第 i 対角成分 λ_i は A の固有値, 各上三角成分は L^T の上三角要素である．途中段階 $A^{(k)}(1 < k < n)$ を

$$A = (a_{i,j}^{(1)}) = A^{(1)} \to A^{(2)} \to \cdots \to A^{(k)} \to \cdots \to A^{(n-1)} \to A^{(n)}$$

と漸化的に計算することで $A^{(n)}$ を求めることができる．

$$A^{(k)} = \begin{pmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & \cdots & a_{1,k}^{(1)} & a_{1,k+1}^{(1)} & \cdots & a_{1,n}^{(1)} \\ & a_{2,2}^{(2)} & \cdots & a_{2,k}^{(2)} & a_{2,k+1}^{(2)} & \cdots & a_{2,n}^{(2)} \\ & & \ddots & \vdots & \vdots & & \vdots \\ & & & a_{k,k}^{(k)} & a_{k,k+1}^{(k)} & \cdots & a_{k,n}^{(k)} \\ & & & & a_{k+1,k+1}^{(k)} & \cdots & a_{k+1,n}^{(n)} \\ & & & & \vdots & & \vdots \\ & & & & a_{n,k+1}^{(k)} & \cdots & a_{n,n}^{(k)} \end{pmatrix}$$

ただし,$a_{i,j}^{(k+1)}$ は以下で定義される．

$$a_{i,j}^{(k+1)} = a_{i,j}^{(k)} - \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}} a_{k,j}^{(k)}$$

3.3 推定原理

本節では, 実際に行列の要素を推定するための原理を示す. 以下に示す行列 $B^{(k)}$ を用いて類似性尺度行列 A の要素を推定する．ここで, 行列 $B^{(k)}$ は行列 $A^{(k)}$ の第 1 行から第 k 行までを取り出した行列である．

$$B^{(k)} = \begin{pmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & \cdots & a_{1,k}^{(1)} & a_{1,k+1}^{(1)} & \cdots & a_{1,n}^{(1)} \\ & a_{2,2}^{(2)} & \cdots & a_{2,k}^{(2)} & a_{2,k+1}^{(2)} & \cdots & a_{2,n}^{(2)} \\ & & \ddots & \vdots & \vdots & & \vdots \\ & & & a_{k,k}^{(k)} & a_{k,k+1}^{(k)} & \cdots & a_{k,n}^{(k)} \end{pmatrix}$$

さらに, 行列 $B^{(k)}$ の第 1 列から第 k 列のみから成る行列の右側に, $B^{(k)}$ 第 $k+1$ 列から第 n 列から任意の 2 列を付加したブロック行列 $\left(B^{(k)}(*, 1..., k)B^{(k)}(*, i)B^{(k)}(*, j)\right)(k+1 \leq i, j \leq n)$ を基に, 事例 i, 事例 j の未知類似尺度値を $s_{i,j}$ とおいて, 新たな $k+2$ の $A'^{(k)}$ を構成する．

$$A'^{(k)} = \begin{pmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & \cdots & a_{1,k}^{(1)} & a_{1,i}^{(1)} & a_{1,j}^{(1)} \\ & a_{2,2}^{(2)} & \cdots & a_{2,k}^{(2)} & a_{2,i}^{(2)} & a_{2,j}^{(2)} \\ & & \ddots & & \vdots & \vdots \\ & & & a_{k,k}^{(k)} & a_{k,i}^{(k)} & a_{k,j}^{(k)} \\ & & & & 1 & s_{i,j} \\ & & & & s_{i,j} & 1 \end{pmatrix}$$

これより，2 段の修正コレスキー分解を進めて最終的な解を求める時,$A'^{(1)}$ が半正定性を満たすためには, $\lambda_j^{k+2} = a_{k+2,j}^{(k+2)} \geq 0$ を満たす必要がある．この制約から次式が導かれる．

$$\left\langle v_i^{(k)}, v_j^{(k)} \right\rangle - \sqrt{\lambda_i^{(k)}\lambda_j^{(k)}} \leq s_{i,j} \leq \left\langle v_i^{(k)}, v_j^{(k)} \right\rangle + \sqrt{\lambda_i^{(k)}\lambda_j^{(k)}}$$

ここで，事例ベクトル $\mathbf{v}_i^{(k)}$ と固有値 $\lambda_i^{(k)}$ は

$$\mathbf{v}_i^{(k)} := a_{1,i}^{(1)}/\sqrt{a_{1,1}^{(1)}}a_{2,i}^{(2)}/\sqrt{a_{2,2}^{(2)}}\ldots a_{k,i}^{(k)}/\sqrt{a_{k,k}^{(k)}}$$

$$\lambda_i^{(k)} := 1 - ||\mathbf{v}_i^{(k)}||^2$$

で与えられる. このように事例 i と事例 j 間の類似性尺度 $s_{i,j}$ の範囲が定まり,λ によって推定精度が決定されることが分かる．

3.4 アルゴリズム

類似性尺度値推定アルゴリズムを図 1 に示す．類似性尺度値の推定値と真値の取りうる範囲と, 基底の追加に伴い固有値 $\lambda_i^{(k)}$, 及び推定誤差が単調に減少することより, 基底数 $k = 1$ からはじめて徐々に増やしながら修正コレスキー分解を拡張することで，十分な推定精度で全類似性尺度を推定できる．その際の基底を選択する基準やアルゴリズム終了条件は以下の通りとする．

基底選択基準：本手法では，次の基底として固有値が最大である事例を選択する方法を採用した．この理由は，固有値が最も大きい事例は現在の基底が張る空間から最も外れている事例であり，その事例を基底として追加することで基底が張る空間がデータの本質次元空間のより良い近似になる，と考えられ

るからである．最初の基底選択には何の事前情報もないため，ランダムに 1 つの事例を選択する．

終了条件：十分な精度を得るために全ての固有値（固有値の最大値）がある固有値閾値 λ_{th} を下回るまで基底を追加し続け，コレスキー分解を反復する．これにより，定理 1 から，最悪でも $2\lambda_{th}$ の誤差を持つ推定を保証できる．この条件に従ってアルゴリズムが停止したとき，全ての類似性尺度値を十分な精度で推定，つまりデータ全体を表現できる最小数の基底を選択したことになり，基底数 k はデータの本質次元に概ね一致する．

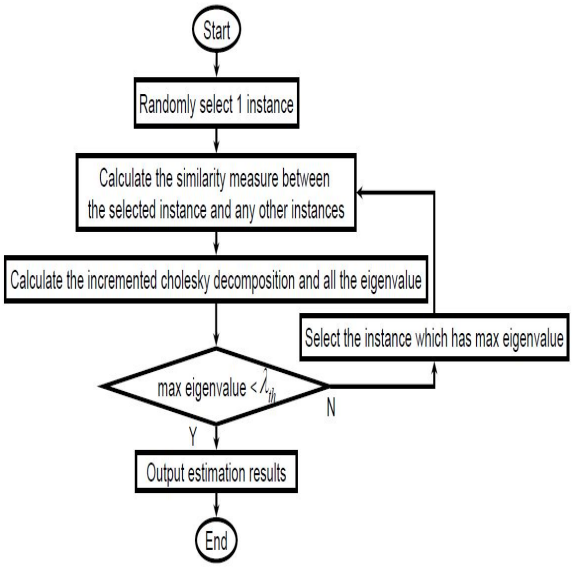

図 1: フローチャート

4. 評価結果

4.1 データの概要

評価実験に用いた人工データは，事例数 n，属性空間次元 m，クラスタ数 k を指定し，最初に k 本の m 次元ベクトルをランダムに発生させ，各クラスタの芯とした．ただし，各次元における値は 0 から 1 の範囲においてランダムな値を取る．各クラスタを構成する事例の数と何れのクラスタにも属さないバックグラウンドデータ数は，事例数 n 中，それぞれランダムな割合とした．さらに，クラスタを構成する事例は，クラスタの芯に対して，各次元において標準偏差の最大値が 0.01 の正規乱数を誤差として加えた．本実験では，$n = 300, 1000, 3000, 10000$，$m = 300, 1000, 3000, 10000$，$k = 3, 10, 30, 100$ の人工データで実験を行った．その際，ユークリッド距離行列を半正定行列に変換する指数変換法において，α も様々な値に変えて実験を行った．

4.2 考察

人工データによる評価実験では，データのデフォルト値として，指数変換法の場合，（事例数，次元数，クラスタ数，アルファの値，精度）$= (1000, 1000, 10, 0.1, 0.1)$，重心参照変換法の場合，（事例数，次元数，クラスタ数，精度）$= (1000, 1000, 10, 0.1)$ とし，各パラメータを変化させて実験を行った．

評価実験の結果より，本推定手法によって，条件によっては類似性尺度推定計算時間が大幅に短縮できることが分かった．図 2，3 から，事例数や属性次元数が変化した場合は，指数変換法よりも重心参照変換法の方が高速であることが分かる．これは，重心参照変換法では，より少ない基底で，適切に他の事例間の距離関係を表すことができるためである．ただし，図 4 に示すようにクラスタ数を変化させた場合は，両変換手法に差異は見られない．これより，推定対象データに存在するクラスタの種類数，即ち対象の本質次元の影響は，両変換とも同じであると考えられる．一方，図 5 より，指数変換法の計算時間は，α の値に大きく影響を受けることが分かる．また，α はデータによって適正な値が異なり，特にデータの値域に大きく依存する．そのため，本研究では，推定の際，ユークリッド距離を各次元の最大値と最小値の差の平均で割り，規格化した．この変換方法では，α を適正値に設定すれば，高速な推定が可能となるが，データごとに異なる α の適正値を一意に決定できない．一方，重心参照変換法では，変換パラメータがないため，安定して高速な計算時間を実現している．図 6 より，固有値閾値 λ_{th} を変化させた場合は，λ_{th} が大きい，即ち低精度では両変換手法は同様に高速であるが，λ_{th} が小さい，即ち高精度推定を行う場合は，重心参照変換法が極めて有利であることが分かる．以上のように，人工データを通じた評価では重心参照変換法が有利であるが，これは参照点を重心に取ることで多くの事例間の半正定類似性尺度の差異が明確になるような変換がなされるためと考えられる．

5. おわりに

本稿では，半正定行列の推定手法をユークリッド距離行列に拡張し，直接に距離計算するよりも高速に距離推定することができた．今後の課題としては，何れのクラスタにも属さないバックグラウンドデータを基底として選択せず，推定することができれば，さらに計算時間の短縮につながると考えられる．

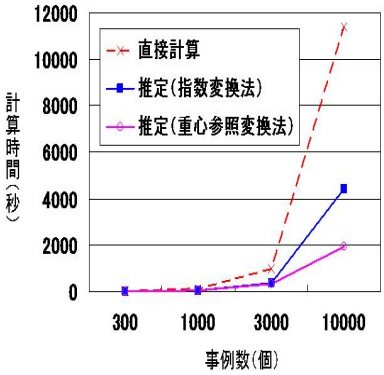

図 2: 事例数を変化させた計算時間

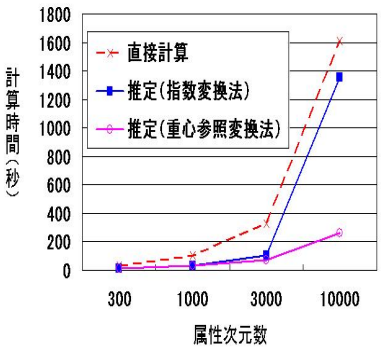

図 3: 属性次元数を変化させた計算時間

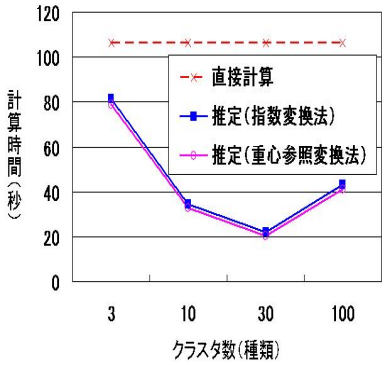

図 4: クラスタ数を変化させた計算時間

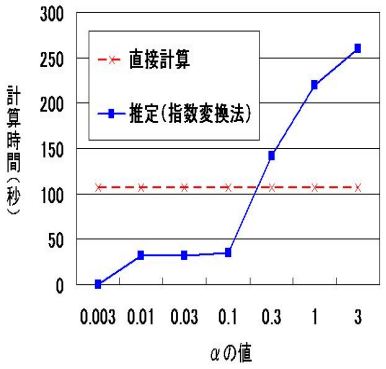

図 5: 変換定数 α を変化させた計算時間

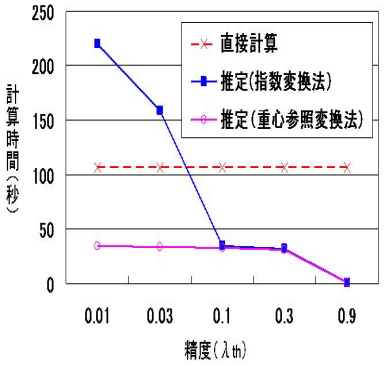

図 6: 精度を変化させた計算時間

参考文献

[CNBM 99] E. Chavez, G. Navarro, R. Baeza-Yates, J. Marroquin: Searching in metric spaces, Technical Report TR/DCC-99-3, Dept. of Computer Science, Univ. of Chile, 1999.

[Yianilos 93] P. N. Yianilos: Data structures and algorithms for nearest neighbor search in general metric spaces, Proc. 4th ACM-SIAM Symposium on Discrete Algorithms (SODA '93), pp.311-321, 1993.

[DN 87] F. Dehne, H. Nolteimer: Voronoi trees and clustering problems, Information Systems, 12(2), pp.171-175, 1987.

[Laurent 98] Monique Laurent: A Connection Between Positive Semidefinite and Euclidean Distance Matrix Completion Problems, Linear Algebra and its Applications, 273, pp.9-22, 1998.

[Laurent 01] Monique Laurent: Matrix Complition Problems, C. Floudas and P. Pardalos, editors, The Encyclopedia of Optimization, volume III, pp.221-229. Kluwer, 2001.

[Graepel 02] T. Graepel: Kernel matrix completion by semidefinite programming, J.R. Dorronsoro, editor, Artificial Neural Networks-ICANN 2002, pp.687-693, Springer Verlag, 2002.

[TAA 03] K. Tsuda, S. Akaho, K. Asai: The em Algorithm for Kernel Matrix Completion with Auxiliary Data, J. of Machine Learning Research, vol.4, 2003.

[Kuwajima 07] H. Kuwajima,K. Nakanishi,T. Washio:Fast Estimation on Positive Semi-Definite Similarity Measures.