1C1-4

半正定性を満たす類似性尺度の高速推定手法

Fast Estimation of Psitive Semi-Definite Similarity Measures

桑島　洋[*1]　　中西　耕太郎[*1*2]　　鷲尾　隆[*1]
Hiroshi KUWAJIMA　Koutarou NAKANISHI　Takashi WASHIO

[*1]大阪大学産業科学研究所
The Institute of Scientific and Industrial Research (ISIR), Osaka University

Under the development of ubiquitous sensing, electric documents and multi-media technologies, data sets consisting of high dimensional and massive instances have become available in various practical fields. Efficient evaluation of the similarity measures, *e.g.*, correlations and distances, among such instances is one of the most important tasks for the instance queries and clustering which are required to analyze the data. However, the computational complexity of the evaluation for n instances is $O(n^2)$ which is practically intractable under the high dimensional and massive data. The objective of this paper is to provide an efficient remedy to this problem. We propose a fast approach to estimate the similarity measures among n instances from their small portion by using a mathematical constraint called "positive semi-definiteness" governing the similarity measures. The superior performance of our approach in both efficiency and accuracy of the estimation is demonstrated though the evaluation based on artificial and real world data sets.

1. はじめに

近年，計算機や実験装置の進歩によって大規模次元の対象変数群を一度に測定可能になり，大規模事例かつ大規模次元のデータベースを扱うことの重要性が増している．例えば，情報検索やバイオインフォマティクス分野のデータが代表的である．データ分析の中でも特に，データベース内の各事例間の相関係数や距離などの類似性尺度の計算は，因果関係の分析，クラスタリング，分類，検索，また，これらのためのカーネル関数構成などに必須である．しかし，対象事例数を n とすると，類似性尺度の組合せ数は $n(n-1)/2$ になり膨大な計算を要する．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，ベクトル化された文書や画像などは超高次元になり，内積や距離などの類似性尺度値の計算に時間がかかる．あるいは，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．このような類似性尺度を用いたとき，現実的に全ての組み合わせで値を調べるのは難しい．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既存技術を用いることが考えられる．

範囲問い合わせ（range query）とは，ユークリッド距離空間のような計量空間において，ある事例との類似性尺度がある指定した値以内の事例を全て検索する問題である．範囲問い合わせには大きく分けてピボットとボロノイ図を利用する 2 つの方法がある [CNBM 99]．ピボットを利用した代表的な手法に Vantage Point Tree [Yianilos 93]，ボロノイ図を利用した代表的な手法には Voronoi Tree [DN 87] があり，それぞれ計算複雑性はインデックス構築が $O(n \log n)$，一回の問い合わせが $O(\log n)$，メモリ使用量は $O(n)$ である．主に類似した事例間の類似性尺度値を知りたい場合は，範囲問い合わせを行ったうえで類似性尺度を計算すれば大幅に計算時間を削減することができる．しかし，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．

一方，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値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．行列補完は全ての類似性尺度値を推定することができるが，計算複雑性から大規模データでは実用的ではない．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をするものであり，いずれの計算複雑性も $O(n^2)$ を超えてしまう [Graepel 02] [TAA 03]．

本研究では，任意の類似性尺度値を計算できる状況において，必要最小限の類似性尺度値を選択的に測定ないし実計算することで，大量事例間の全ての類似性尺度を効率的かつ必要な精度を確保して推定することを目途として，上記各手法の長所を兼ね備えた手法を提案する．本手法は，任意の類似性尺度値を測定ないしは計算することができるが，計算複雑性，測定や計算コスト，あるいはその他の理由から，最小限の類似性尺度値から全ての類似性尺度値を推定する必要がある場合に威力を発揮する．これに対して，上記の範囲問い合わせは高速だが完全性を持たず，行列補完やカーネル関数推定は任意の既知類似性尺度値からその他の全ての類似性尺度値を推定できるものの，直接計算より低速である．任意の類似性尺度値を測定ないし計算可能な状況で，高い類似性を持つ関係とその尺度値のみを知りたい場合には範囲問い合わせが有効である．既知の類似性尺度値以外を測定ないし計算することが困難な場合には，行列補完やカーネル関数推定を行う必要がある．本論文で提案する手法は類似性尺度行列が半正定である類似性尺度を対象とするが，ピアソンの相関係数やスピアマンの順位相関係数，コサイン距離をはじめ多くの相関係数や内積距離が半正定類似性尺度であることが知られている．また，広く用いられているユークリッド距離行列は半負定だが，半正定に容易に変換可能であることが知られており，本手法の適用範囲は非常に広い

連絡先: 氏名：桑島　洋，所属：大阪大学産業科学研究所，住所：大阪府茨木市美穂が丘 8-1，電話番号：+81-6-6879-8544，電子メールアドレス：kuwajima@ar.sanken.osaka-u.ac.jp
∗ 2 株式会社日本総合研究所所属

[Laurent 98].

本手法による全事例類似性尺度の十分な精度での推定は時間計算量 $O(n^2)$ である．類似性尺度行列やカーネル関数の全要素を実計算する時間計算量も $O(n^2)$ だが，高次元事例データであるなどの理由で個々の尺度値の測定や計算にコストがかかる場合には，本手法によって定数倍の高速化が期待できる．本手法における一回の推定のための時間コストを c_e，類似性尺度値を直接計算する時間コストの平均値を c_d とすると，$c_e < c_d$ の場合には本手法が有効になる．ここで，k を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，即ち本質次元とすると，本手法では $c_e \propto k$ である．これに対して，全事例数，即ち類似性尺度行列の大きさ n を表現次元という．一般にデータの広がる空間次元 k は表現次元 n やデータの属する属性空間次元 m より大きくなり得ないので $k \leq \min(n, m)$ である．これにより，相関係数やユークリッド距離のような類似性尺度値を直接計算する時間コスト c_d が $c_d \propto m$ であるような類似性尺度の場合は，本手法が直接計算よりも高速になることが多くなる．また，データが少数の部分空間領域に密集している場合は $k \ll n$ になり，高速化が期待できる．

2. 背景理論

本章では，後の章で我々が提案する類似性尺度値の推定手法で必要とされる，類似性尺度行列の半正定性と修正コレスキー分解について簡単に説明する．同時に，本手法が正規化された半正定行列のみを対象とすることの一般性，妥当性を述べる．

2.1 半正定類似性尺度行列

本稿において類似性尺度行列とは，類似性尺度を要素に持つ行列である．例えば，相関係数 r_{ij} を要素に持つ相関係数行列 $R = \left(r_{ij}\right)$，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 $D = \left(d_{ij}\right)$ などである．本手法は類似性尺度行列が半正定行列になる類似性尺度を対象とする．相関係数行列は半正定であり，距離行列は半負定だが半正定に変換可能である [Laurent 98].

半正定行列とは，任意の非零ベクトルに対して二次形式が非負であるような行列である．つまり，

$$x^T A x \geq 0, \forall x \neq 0 \tag{1}$$

であるような行列 A を半正定行列と呼ぶ．主小行列式が全て非負であることは半正定行列の必要十分条件である．ただし，主小行列とは元の行列の任意の行と列を削除した行列である．この判定法はシルベスターの判定法と呼ばれる．また，行列の全ての固有値が非負であることも半正定性の必要十分条件である．

2.2 半正定類似性尺度行列の対角成分正規化

ある行と列以外を全て削除した主小行列は，対角成分になる．従って，対角成分が非負であることは半正定性の必要条件である．行列 A の全ての対角成分が正ならば，その第 i 対角成分を $a_{i,i}$ としたとき，各対角成分が $1/\sqrt{a_{i,i}}$ であるような対角行列 Q によって A を各対角成分を 1 に正規化した $Q^T AQ$ に変換することができる．この変換は常に逆変換 ${Q^{-1}}^T Q^T AQQ^{-1} = A$ を持ち，また $n \times n$ の対角行列やその逆行列の掛け算の計算複雑性は $O(n^2)$ であるが個々の演算は簡易である．従って，A の既知要素について対角成分を正規化した後に以降で述べる種々の処理を施し，それを逆変換することで元の A の各要素を容易に推定できる．相関係数やコサイン類似度など，ほとんどの半正定性を満たす類似性尺度は正の対角成分を持つので，以降述べる手法において A の対角成分が正規化されていることを仮定しても一般性は失われない．

2.3 修正コレスキー分解

修正コレスキー分解は，任意の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 分解を対角対称行列に特化した手法である．半正定行列は対角対象行列なのでこの分解が適用可能である．修正コレスキー分解によって任意の行列 A は

$$A = LDL^T \tag{2}$$

のように下三角行列 L と固有値を対角成分に持つ対角行列 D に分解することができる．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

$$A^{(n)} = \begin{pmatrix} \lambda_1 & l_{2,1} & \dots & l_{n,1} \\ & \lambda_2 & \ddots & \vdots \\ & & \ddots & l_{n,n-1} \\ & & & \lambda_n \end{pmatrix} = \begin{pmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & \dots & a_{1,n}^{(1)} \\ & a_{2,2}^{(2)} & \ddots & \vdots \\ & & \ddots & a_{n-1,n}^{(n-1)} \\ & & & a_{n,n}^{(n)} \end{pmatrix} \tag{3}$$

を求める．ここで各 λ_i は A の固有値，$l_{i,j}(i > j)$ は L の要素である．次のように途中段階 $A^{(k)}(1 < k < n)$ を $A = \left(a_{i,j}^{(1)}\right) = A^{(1)} \rightarrow A^{(2)} \rightarrow \ldots \rightarrow A^{(k)} \rightarrow \ldots \rightarrow A^{(n-1)} \rightarrow A^{(n)}$ と漸化的に計算することで $A^{(n)}$ を求めることができる．

$$A^{(k)} = \begin{pmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & \dots & a_{1,k}^{(1)} & a_{1,k+1}^{(1)} & \dots & a_{1,n}^{(1)} \\ & a_{2,2}^{(2)} & & a_{2,k}^{(2)} & a_{2,k+1}^{(2)} & & a_{2,n}^{(2)} \\ & & \ddots & \vdots & \vdots & & \vdots \\ & & & a_{k,k}^{(k)} & a_{k,k+1}^{(k)} & \dots & a_{k,n}^{(k)} \\ & & & & a_{k+1,k+1}^{(k)} & \dots & a_{k+1,n}^{(k)} \\ & & & & \vdots & \ddots & \vdots \\ & & & & a_{n,k+1}^{(k)} & \dots & a_{n,n}^{(k)} \end{pmatrix} \tag{4}$$

ただし，$a_{i,j}^{(k+1)}$ は以下で定義される．

$$a_{i,j}^{(k+1)} = a_{i,j}^{(k)} - \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}} a_{k,j}^{(k)} \tag{5}$$

3. 類似性尺度値の推定

本手法は，本質次元の基底事例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．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て残りの事例間の類似性尺度値を推定する．後述するように，k はデータの本質次元に概ね一致するように決めることができ，類似性尺度行列の半正定性と修正コレスキー分解を利用することで未知の類似性尺度値の推定とその範囲を決定できる．また，基底数を増やせばその存在範囲を単調に狭めることができ，類似性尺度値を任意の精度で推定可能である．

3.1 推定原理

本手法は，式（6）に示す行列 $A^{(k)}$ の第 1 行から第 k 行までを取り出した行列 $B^{(k)}$ を用いて類似性尺度行列 A の残り部

分を推定する．

$$B^{(k)} = \begin{pmatrix} a^{(1)}_{1,1} a^{(1)}_{1,2} \dots a^{(1)}_{1,k} a^{(1)}_{1,k+1} \dots a^{(1)}_{1,n} \\ a^{(2)}_{2,2} \quad a^{(2)}_{2,k} a^{(2)}_{2,k+1} \quad a^{(2)}_{2,n} \\ \ddots \quad \vdots \quad \vdots \quad \vdots \\ a^{(k)}_{k,k} a^{(k)}_{k,k+1} \dots a^{(k)}_{k,n} \end{pmatrix} \tag{6}$$

$A^{(k)}$ の第 1 行から第 k 行は一度計算されると値が変わることはないので，$B^{(k)}$ は $A^{(n)}$ の第 1 行から第 k 行でもある．前述の通り半正定行列の固有値は全て非負であり，$B^{(k)}$ の $a^{(i)}_{i,i}(1 \leq i \leq k)$ 成分，すなわち $A^{(n)}$ の対角成分は固有値なので，これらも非負である．

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

$$A'^{(k)} = \begin{pmatrix} a^{(1)}_{1,1} a^{(1)}_{1,2} \dots a^{(1)}_{1,k} a^{(1)}_{1,i} a^{(1)}_{1,j} \\ a^{(2)}_{2,2} \dots a^{(2)}_{2,k} a^{(2)}_{2,i} a^{(2)}_{2,j} \\ \ddots \quad \vdots \quad \vdots \\ a^{(k)}_{k,k} a^{(k)}_{k,i} a^{(k)}_{k,j} \\ 1 \quad s_{i,j} \\ s_{i,j} \quad 1 \end{pmatrix} \tag{7}$$

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

$$A'^{(k+2)} = \begin{pmatrix} a^{(1)}_{1,1} a^{(1)}_{1,2} \dots a^{(1)}_{1,k} \quad a^{(1)}_{1,i} \quad a^{(1)}_{1,j} \\ a^{(2)}_{2,2} \dots a^{(2)}_{2,k} \quad a^{(2)}_{2,i} \quad a^{(2)}_{2,j} \\ \ddots \quad \vdots \quad \vdots \\ a^{(k)}_{k,k} \quad a^{(k)}_{k,i} \quad a^{(k)}_{k,j} \\ a^{(k+1)}_{k+1,i} \, a^{(k+1)}_{k+1,j} \\ a^{(k+2)}_{k+2,j} \end{pmatrix} \tag{8}$$

を求めるとき，$A'^{(1)}$ が半正定性を満たすためには $s_{i,j}$ が $\lambda^{k+2}_j = a^{(k+2)}_{k+2,j} \geq 0$ を満たす必要がある．$\lambda^{k+1}_i = a^{(k+1)}_{k+1,i}$ は $s_{i,j}$ に無関係に決まるのでここでは考慮する必要はない．式 (5) と行列の対称性 $a^{(l)}_{i,l} = a^{(l)}_{l,i}$ より

$$a^{(k+1)}_{k+1,j} = s_{i,j} - \sum_{l=1}^{k} \frac{a^{(l)}_{l,i}}{a^{(l)}_{l,l}} a^{(l)}_{l,j} \tag{9}$$

であることを用いると，同じく式 (5) と行列の対称性 $a^{(l)}_{j,l} = a^{(l)}_{l,j}$ より，

$$a^{(k+2)}_{k+2,j} = 1 - \sum_{l=1}^{k} \frac{{a^{(l)}_{l,j}}^2}{a^{(l)}_{l,l}} - \frac{{a^{(k+1)}_{k+1,j}}^2}{a^{(k+1)}_{k+1,k+1}} \tag{10}$$

$$= -\frac{1}{a^{(k+1)}_{k+1,k+1}} (s_{i,j} - \sum_{l=1}^{k} \frac{a^{(l)}_{l,i}}{a^{(l)}_{l,l}} a^{(l)}_{l,j})^2 + 1 - \sum_{l=1}^{k} \frac{{a^{(l)}_{l,j}}^2}{a^{(l)}_{l,l}} \geq 0 \tag{11}$$

である．この二次不等式を解き，$a^{(k+1)}_{k+2,j} = 1 - \sum_{l=1}^{k} {a^{(l)}_{l,k+2}}^2 / a^{(l)}_{l,l}$ を用いると次の不等式が得られる．

$$\sum_{l=1}^{k} \frac{a^{(l)}_{l,i} a^{(l)}_{l,j}}{a^{(l)}_{l,l}} - \sqrt{a^{(k+1)}_{k+1,i} a^{(k+1)}_{k+1,j}} \leq s_{i,j} \leq \sum_{l=1}^{k} \frac{a^{(l)}_{l,i} a^{(l)}_{l,j}}{a^{(l)}_{l,l}} + \sqrt{a^{(k+1)}_{k+1,i} a^{(k+1)}_{k+1,j}} \tag{12}$$

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

$$\mathbf{v}^{(k)}_i := \left(a^{(1)}_{1,i} / \sqrt{a^{(1)}_{1,1}}\, a^{(2)}_{2,i} / \sqrt{a^{(2)}_{2,2}} \dots a^{(k)}_{k,i} / \sqrt{a^{(k)}_{k,k}}\right)^T \tag{13}$$

$$\lambda^{(k)}_i := 1 - ||\mathbf{v}^{(k)}_i||^2 \tag{14}$$

と定義して式 (12) に代入すると，次のように事例 i と事例 j 間の類似性尺度 $s_{i,j}$ の範囲が定まる．

定理 1：類似性尺度値の推定値と真値の取りうる範囲

事例 i と事例 j の類似性尺度値 $s_{i,j}$ の中央推定値は $\langle \mathbf{v}^{(k)}_i, \mathbf{v}^{(k)}_j \rangle$ であり，その真値の範囲は

$$\langle \mathbf{v}^{(k)}_i, \mathbf{v}^{(k)}_j \rangle - \sqrt{\lambda^{(k)}_i \lambda^{(k)}_j} \leq s_{i,j} \leq \langle \mathbf{v}^{(k)}_i, \mathbf{v}^{(k)}_j \rangle + \sqrt{\lambda^{(k)}_i \lambda^{(k)}_j} \tag{15}$$

である． ■

更に，式 (14) に式 (13) を代入して計算すると式 (16) が導かれ，それから式 (17) が導かれる。これより定理 2 が成立する。

$$\lambda^k_i = 1 - \sum_{l=1}^{k} \frac{{a^{(l)}_{l,i}}^2}{a^{(l)}_{l,l}} \tag{16}$$

$$\lambda^k_i - \lambda^{k+1}_i = \frac{{a^{(k+1)}_{k+1,i}}^2}{a^{(k+1)}_{k+1,k+1}} \geq 0 \tag{17}$$

定理 2：固有値及び真値範囲の単調性

定理 1 から k の増加，つまり基底の追加に従って，固有値 $\lambda^{(k)}_i$ は単調減少する．また，これにより真値の取りうる範囲幅も単調減少する． ■

3.2 アルゴリズム

提案する類似性尺度値推定アルゴリズムを図 1 に示す．定理 1 と定理 2 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：本論文で提案する手法では，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．

4. 実験評価結果

4.1 データと評価方法

実験データとして，実データと人工データを用いて実験を行った．人工データは事例数 n，属性空間次元 m，クラスタ数 k，を指定して，最初に k 本の m 次元ベクトルをランダムに発生させ，各クラスタの芯とする．次にこれらの芯の周辺

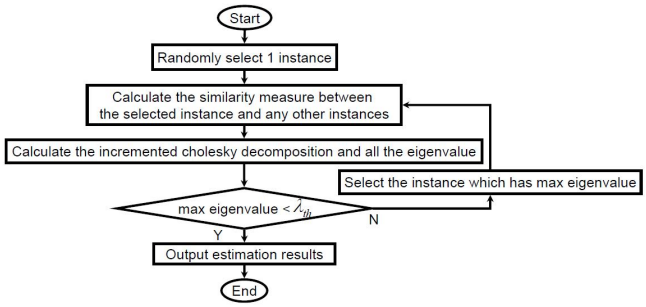

図 1: アルゴリズム

にランダムに事例ベクトルを n/k 本生成し，クラスタを構成する．このようにして事例が k 個の空間領域に密集しているデータを生成することができる．本実験では，$n = 1000, m = 1000, k = 10$ を標準データとして，$n = 300, 1000, 3000, 1000$, $m = 300, 1000, 3000, 1000$, $k = 3, 10, 30, 100$ の範囲で種々のパラメータ特性を持った実験データを作成した．

実データは UCI Machine Learning Repository で提供される 4 種のデータ：musk，isolet，spambase，inonosphere を用いた．各実データの (事例数，属性空間次元) はそれぞれ musk(6598，167)，isolet(6238，618)，spambase(4601，58)，ionosphere(351，34) である．

以上の人工データと実データに対する相関係数行列，つまり全事例間の相関係数を推定するのに要する時間を計測し，直接計算で相関係数行列を求める時間と比較した．以下では推定の精度として 0.1，即ち $\lambda_{th} = 0.05$ の場合のみを示す．

4.2 結果

図 2 に人工データに関する事例数，属性空間次元，クラスタ数と実行時間関係の実験結果を示す．事例数や属性空間次元が大きい場合に，本手法は直接計算より遥かに高速な計算が可能であることが判る．クラスタ数が十分小さい場合，つまりデータが少数の空間領域に密集している場合にも本手法がは極めて高速である．

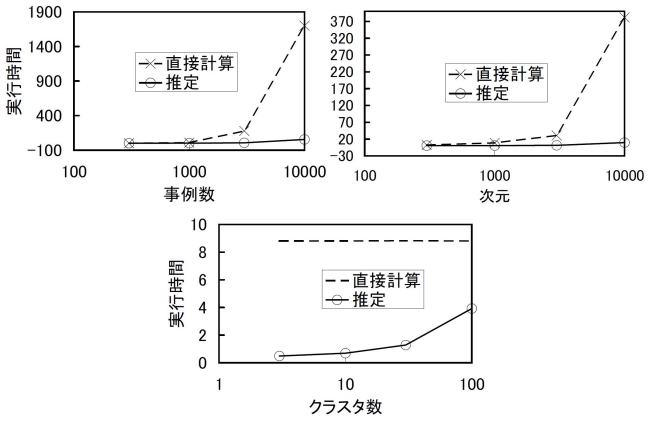

図 2: 人工データ実験結果

表 1 に計算時間に関する実データの実験結果を示す．基底数が次元よりも十分に小さいときに本手法が直接計算よりも高速になることが分かる．musk と isolet は表現次元や属性空間次元に対して基底数が十分小さいので直接計算より推定が高速だが，spambase と ionosphere は基底数が属性空間次元と大きく変わらないため，推定が直接計算より低速になっている．これは，各々の類似性尺度値の推定による高速化効果の合計が，アルゴリズム前半部の基底選択とコレスキー分解構築部分に要する計算時間コストを超えた場合に，全体として高速化できるからである．この問題は本論文で扱ったような比較的小規模の実データを扱う場合は無視できないが，データが大規模事例になるほど本手法が有利になる．なぜなら，基底選択とコレスキー分解構築部分の時間計算量は高々$O(kn)$ だが，各類似性尺度値の高速化は類似性尺度の個数 $O(n^2)$ 全てに亘るからである．musk と isolet で良い結果が得られたのは，事例数が比較的大きいことも原因と考えられる．

	推定 (s)	直接計算 (s)	基底数
musk	67.55	86.66	66
isolet	175.14	286.14	196
spambase	24.12	11.64	45
ionosphere	0.11	0.03	33

表 1: 実データ実験結果

5. 結論

本論文では相関係数などの半正定類似性尺度を高速推定する手法を提案した．実装と実験の結果，事例数と次元に関して優れた高速性を示すこと，特に大規模な事例数に対して威力を発揮することがわかった．

我々は現在，高い類似性尺度値のみを精度良く推定し，低い類似性尺度値の推定精度を落とすことでさらに高速化する手法を開発中である．

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.