情報処理学会論文誌　データベース　Vol. 2　No. 3　29–40 (Sep. 2009)

大量頻度分布データのための高速探索アルゴリズム

松　原　靖　子[†1]　　櫻　井　保　志[†2]　　吉　川　正　俊[†1]

分布データは気象，産業，金融をはじめとする様々な分野で発生している．本論文では，与えられた問合せデータに対して，複数の分布データの中から類似データを探索する問題を対象とする．本論文では，分布データの探索手法として D-Search を提案する．D-Search は（1）SVD を利用した KL 情報量の近似計算を高速に行い，(2) 分布データを様々な近似の粒度で探索し，類似度の低い分布を高速に枝刈りする．本研究ではさらに D-Search の拡張版として，(3）時系列分布データの類似探索手法を提案し，時系列分布データの中から任意の長さの類似部分シーケンスを探索する問題を扱う．様々な実データを用いた実験を行い，D-Search が時系列分布データから正確に部分シーケンスを検出し，そしてナイーブな手法と比較して大幅な性能向上を達成していることを明らかにした．

An Efficient Search Algorithm for Large Distribution Sets

Yasuko Matsubara,[†1] Yasushi Sakurai[†2]
and Masatoshi Yoshikawa[†1]

Distribution data naturally arise in countless domains, such as meteorology, biology, geology, industry and economics. However, relatively little attention has been paid to data mining for large distribution sets. Given n distributions of multiple categories and a query distribution Q, we want to find similar clouds (i.e., distributions), to discover patterns, rules and outlier clouds. We propose D-Search, an efficient algorithm for similarity search in large distribution datasets. Our main contributions are (1) approximate KL divergence, which can speed up cloud-similarity computations by using SVD coefficients, (2) multi-step sequential scan, which efficiently prunes a significant number of search candidates and leads to a direct reduction in the search cost. We also introduce an extended version of D-Search : (3) timeseries distribution mining, which uses the D-Search algorithm and finds similar subsequences in time-series distribution datasets. Extensive experiments on real multi-dimensional datasets show D-Search is significantly faster than the naive method.

1. はじめに

分布データは気象，産業，金融をはじめとする様々な分野において大量に発生している．分布データを扱うアプリケーションにおいてしばしば求められるのは，大量の分布データの中からパターンの発見，異常値検出等を行う技術である．しかしその一方で，大量の分布データを対象としたマイニング技術についてはあまり取り組まれていない．本論文では，与えられた問合せデータに対して，複数の分布データの中から類似データを探索する問題を対象とし，これらを解決する手法として D-Search を提案する．

図 1 を用いて本論文で扱う問題の説明を行う．図 1 はそれぞれ，モーションキャプチャにおける左右の足の運動エネルギーの散布図である．図中上の問合せデータと左下の分布データ 1 は running のモーション，そして図中右下の分布データ 2 は squatting のモーションを表している．本研究で解決したい問題は，問合せの分布が与えられたときに類似した分布を発見することである．図 1 の例では，問合せデータである running のモーションに対し，分布データ 1 が類似分布データとして発見されるが，一方分布データ 2 は類似分布として判断されないことになる．

本論文では主に，数値分布データの類似探索に焦点をあて手法の提案を行う．しかし，我々の提案する D-Search は，数値分布データと同様に，カテゴリ分布データに対しても適応することが可能である．

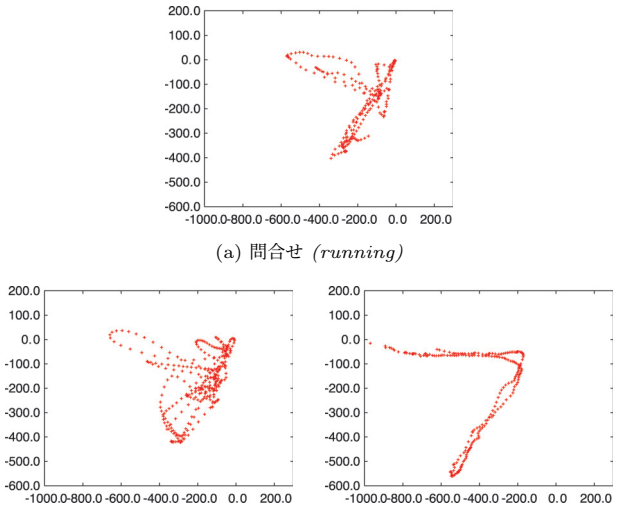
1.1 応　用　例

分布データの探索は，様々なアプリケーションに適応できる．本節では，D-Search の有用性を示すために，いくつかの領域とそのアプリケーションの例について紹介する．

- マルチメディア：マルチメディア分野におけるマルチモーダルデータのマイニングは，様々な形で研究がさかんに行われている[3),7),9),12)]．マルチメディアデータは，デジタル画像，オーディオ，ビデオ，テキストデータ等，様々なものがあり，これらのデータのマイニングは非常に有益である．マルチメディア分野における分布データ探索の例として，モーションキャプチャにおけるヒトの動きの検出が考えられる．体の各所の時間ごとの加速度の値は，d 次元の数値分布データに変換することができ，これらの分布データを用いて類似探索を行うと，アノテーションやそのほかのメタデータを用意することなく類似モーションを発見することができる．他の例として，デジタル画像につい

†1 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University
†2 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, NTT Corporation

© 2009 Information Processing Society of Japan



(a) 問合せ *(running)*

(b) 分布データ 1 *(running)*　(c) 分布データ 2 *(squatting)*

図 1　モーションキャプチャデータを用いた 3 つの分布データ
Fig. 1　Three distributions from the *MoCap*data.

ても同様に，2 次元のピクセル情報を利用した類似ヒストグラムの探索を行うと，容易に類似画像を抽出することが可能となる．

- 医療データ解析：大量の医療データの中から有益な情報を抽出することは，多くの医療領域において研究の中心的な課題となっている[16]．たとえば，脳波（EEG）のデータを用いたメンタルタスクの探索問題は，ヒトの脳機能の理解に非常に有益であり，様々な取り組みがなされている[4),20]．
- *Web* サービス：*Web* サービス のマイニングには，様々なアプリケーションが考えられる．1 つ目の例として，ブログ等の SNS サービスのマイニング問題が考えられる．それぞれのブログユーザは，1 日ごとの参照数，投稿数，投稿記事に含まれるリンクの URL，トラックバックの数等のいくつかの数値データ，カテゴリカルデータを保持しており，これらのデータを分布データとして調べることで，ユーザの傾向調査が可能となる．2 つ目の例として，オンデマンド TV サービスのマイニングが考えられる．日々のオンデマンドの視聴状況のレコードは，番組のジャンル，番組の視聴時間，視聴したユーザの ID 等のデータを保持し，これらのカテゴリカルな分布データを解析することで，どの視聴者がどのコミュニティに属しているか等の視聴者の傾向把握を実現できる．これらのマイニングによって，サービス提供者はサービスデザインやターゲットの絞り込みを行うことができる．
- 電子商取引：電子商取引におけるマイニングとして，顧客の購入傾向の解析が考えられる．すべての購入履歴には，顧客がその買い物にかけた時間，購入した商品の種類や数，価格等の情報が保持され，これらを 4 次元の分布データとして類似分布探索を行うと，どの顧客がどのような商品を買うかといった内容や，多くの買い物をする顧客は買い物にかける時間が長いかどうか等の検証が可能となり，顧客の購入傾向の調査やマーケット調査，異常検出が実現できる．

1.2　提案手法の貢献

本論文では大量頻度分布データの探索手法として D-Search を提案する．D-Search は，SVD を利用した KL 情報量の近似計算を高速に行い，分布データを様々な近似の粒度で探索する．類似度の低い分布を効率的に枝刈りすることによって，大幅な探索コストの低減化を可能とする．

本論文の貢献は次のとおりである．(1) 理論的な分析を行い，ナイーブな手法と提案手法の探索コストを比較する．分布データの数を n，ヒストグラムのバケット数を m としたとき，ナイーブな手法が KL 情報量の計算に $O(mn)$ のコストを必要とするのに対し，提案手法では $O(n)$ の計算コストしか必要とせず，このことによって探索処理において効率的に枝刈りすることができる．(2) 実験において探索漏れを発生させることなく高い探索性能を達成した．本研究ではさらに D-Search の拡張版として，(3) 時系列分布データの類似探索手法を提案し，時系列分布データの中から任意の長さの類似部分シーケンスを探索する問題を扱う．

本論文の構成は次のとおりである．2 章では関連研究をあげる．3 章で本論文における問題設定を述べた後に，分布を高速に探索するための提案手法について説明する．4 章では実験結果について検証し，D-Search の有効性を示す．5 章は，本論文のまとめである．

2. 関 連 研 究

関連研究は大きく 2 種類に分類することができる．第 1 のグループは分布間の類似関数に関するものであり，第 2 のグループは確率的問合せに関するものである．

2.1　分布間の比較

2 つの分布が同じものであるかどうかの判断には，χ^2 検定や Kolmogorov-Smirnoff 等いくつかの統計的手法が用いられる[15]．しかしこれらの方法はスコア値を算出することがで

© 2009 Information Processing Society of Japan

きず，採択・棄却の 2 択の結果を与えるのみである．さらに，これらの手法は高い次元数においては容易に扱うことができない．

画像処理分野においては，*Earth Mover's Distance* と呼ばれる画像比較法が存在する[17]．これは，ある画像を別画像へ変換する際の最小コストを計算し，それらのコストの比較を行うことで，画像間の類似性を比較するというものである．

2 つの点集合 P と Q 間の距離・類似度の測定にはいくつかの方法があり，最も近いペアの距離（min-min distance）を用いるというものや，集合 P と Q のハウスドルフ距離を用いる等がある．最後に，tri-plots[22] は，2 つの大きな次元数の点集合について，距離値を算出することなく比較することができる．

本論文の問題設定に最も適した解決案として，カルバックライブラー（KL）情報量（式 (1) を参照）を利用して，2 つの分布間の類似性を評価することが考えられる．KL 情報量は主に情報理論[23] やパターン認識[10),19] 等で用いられ，2 つの確率密度関数（PDF）間の類似性を評価するための尺度である．

2.2　確率的問合せ

本論文に関連する問題として，確率的問合せがあげられる．Cheng ら[5] は，未知のモデルに基づく不確実データの元の分類問題を扱った．確率的問合せのための索引法については文献 6) の中で提案されている．さらに，Tao ら[21] は，不確実性データの確率密度関数に対する索引アルゴリズムについて示した．本論文における分布データは，ヒストグラムと同様に確率密度分布として表現することができるが，本研究と文献 5) の違いは，我々の研究が各分布間の距離の比較に焦点を当てているのに対し，Cheng らの研究は，与えられた閾値をもとに各分布がどのエリアに所属しているかに焦点を当てている点である．

大規模な分布データの探索とマイニングは，我々の知る限り，まだ取り組まれていない問題である．さらに，本節において述べた分布間の距離関数はどれも，確率密度関数をはじめとする連続分布を対象にしたものであり，これらは計算コストが非常に高い点が課題である．本研究のファーストステップとして，分布データの分類問題について取り組んでいる[18]．本論文では KL 情報量に基づく分布データの探索問題に取り組み，KL 情報量の近似手法と探索アルゴリズムを提案している．

3. 提 案 手 法

本章では，分布データの探索について，背景と問題設定を示し，そしてその問題を解決するためのアルゴリズムである D-Search について述べる．さらに，D-Search の拡張版である時系列分布データのマイニング手法について述べる．

表 1　主な記号の定義
Table 1　Symbols and definitions.

Symbol	Definition
n	分布データの個数
m	バケットの個数
P, Q	分布データのヒストグラム
$\hat{P}$	P の値の対数をとったヒストグラム
p_i, $\hat{p}_i$	P と $\hat{P}$ の i 番目のバケットの値
S_p	P の SVD 係数
$\hat{S}_p$	$\hat{P}$ の SVD 係数
sp_i	S_p の i 番目の係数の値
$d(P, Q)$	P と Q の対称的な KL 情報量
$d'_c(P, Q)$	P と Q の c 個の係数を用いた対称的な KL 情報量の近似値

3.1　背　　景

3.1.1　KL 情報量

関連研究の章で述べたように，2 つの分布データ P と Q が与えられたとき，その分布データのペアの類似度を測る指標にはいくつかのものが存在する．本論文では，情報理論を用い，なかでもカルバックライブラー（KL）情報量を用いることによって，分布データのペアの類似度の計算を行う．2 つの点集合 P，Q と，それぞれの確率密度関数 $\mathcal{P}$，$\mathcal{Q}$ を考える．もし確率密度関数 $\mathcal{P}$，$\mathcal{Q}$ が事前に分かっている場合，連続確率変数を用いて KL 情報量を計算すればよいが，$\mathcal{P}$，$\mathcal{Q}$ は事前に得ることができない．そこで本論文では分布 P, Q を m 個のバケットに分割してヒストグラムを作成し，それらの離散確率変数を用いて以下のように KL 情報量を計算する．

$$d_{KL}(P, Q) = \sum_{i=1}^{m} p_i \cdot \log\left(\frac{p_i}{q_i}\right) \tag{1}$$

ここで p_i，q_i はそれぞれ，分布データ P，Q の i 番目のバケットを示し，さらに，$\sum_{i=1}^{m} p_i = \sum_{i=1}^{m} q_i = 1$ となる．

しかし，上記の KL 情報量の定義式には対称性がないため，本論文における類似分布探索での使用には適していない．そこで本論文では対称 KL 情報量 d_{SKL} を使用する．

$$d_{SKL}(P, Q) = d_{KL}(P, Q) + d_{KL}(Q, P) = \sum_{i=1}^{m} (p_i - q_i) \cdot \log\left(\frac{p_i}{q_i}\right). \tag{2}$$

以下本論文では，$d_{SKL}(P, Q)$ を $d(P, Q)$ と簡略化して記述する．

© 2009 Information Processing Society of Japan

また，上記の KL 情報量の計算において，それぞれのバケットに 0 の値を入れることができないが，本論文で扱う分布データは各バケットに 0 の値が入る可能性がある．そこで本論文ではこれらの空値を避けるために Laplace 推定量[11),13)] を利用し空値のバケットに対するかさ増しを行う．

$$x'_i = \frac{x_i + 1}{|X| + m} \cdot |X|. \tag{3}$$

ここで，x_i（$i = 1, \ldots, m$）はオリジナルのヒストグラムの値を示し，x'_i は Laplace 推定量による x_i の変換後の値を示し，$|X|$ は分布データに含まれる点の総数を示す（つまり，$|X| = \sum_{i=1}^{m} x_i$ となる）．空値を避けるためのその他の解決法として，空値を持つバケットに対し，$n << 1/|X|$ となるような非常に小さな値 n を入れ込む方法等も考えられる．本論文では，探索の精度を高めるために Laplace 推定量を採用するが，他の方法も同様に適応することが可能であり，これらの選択は我々の提案するアルゴリズムとは完全に独立したものである．

3.1.2　特異値分解

m 個のバケットに分割された n 個のヒストグラムの集合 $\mathbf{P}$ は，特異値分解（SVD）を用いて次のように変換することができる．

あらゆる行列 $\mathbf{P} \in \mathbb{R}^{n \times m}$ は，以下のような分解が存在する．

$$\mathbf{P} = \mathbf{U\Sigma V}^T$$

ここで，$\mathbf{U} \in \mathbb{R}^{n \times r}$，$\mathbf{V} \in \mathbb{R}^{m \times r}$，$\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ であり，r は $\mathbf{P}$ のランクで，$r \leq min(n, m)$ となる．$\mathbf{V} \equiv [\mathbf{v_1} \cdots \mathbf{v_r}]$ の列 $\mathbf{v_i}$ は，$\mathbf{P}$ の右特異ベクトルであり，$\mathbf{P}$ の正規直交基底となる．同様にして，$\mathbf{U} \equiv [\mathbf{u_1} \cdots \mathbf{u_r}]$ の列 $\mathbf{u_i}$ は，$\mathbf{P}$ の左特異ベクトルであり，$\mathbf{P}$ の正規直行基底となる．最後に，$\mathbf{\Sigma} \equiv diag[\sigma_1 \cdots \sigma_r]$ は，$\mathbf{P}$ の特異値と呼ばれ，非負の数 σ_i をとる対角行列である．上記の SVD により変換された値 $\mathbf{S_P} \equiv [\mathbf{sp_1} \cdots \mathbf{sp_r}]$（つまり，$\mathbf{P}$ の写像）は，$\mathbf{S_P} = \mathbf{U\Sigma}$ として与えられる．

3.2　D-Search

本節では，以下の問題の解決に焦点を当てる．

問題 1（分布データの探索）　n 個の分布データ集合と問合せ分布データ Q が与えられたとき，問合せ Q に対して対称 KL 情報量が最小となる k 個の分布を探索する．

この問題は，(a) 分布データの圧縮と距離計算の高速化，(b) 探索のコスト削減のための効率的な枝刈り処理，(c) 提案手法のメモリ量と計算コストの検証の 3 つに分割される．以下では，これらの問題の解決法について順に述べる．

3.2.1　SVD を用いた KL 情報量の近似計算

2 つの分布データ P と Q の類似性を効率的に評価するため，本論文では 2 つの課題に焦点を当てる．1 つ目の課題は計算処理をいかにして高速化するか，2 つ目の課題はどのようにして分布データの情報を少ないメモリ量で蓄積するかである．

ナイーブな手法：ナイーブな手法は，m 個にバケット化されたヒストグラムの分布データそれぞれに対して厳密に KL 情報量を計算し，それらの値を用いて近傍探索等のアルゴリズムに適用する方法である．本論文ではこの手法を特に，*NaiveDense* と呼ぶ．

しかし，*NaiveDense* は非常に多くのメモリ量を必要とするため，特に次元数の多い分布データに対しては不適切な手法といえる．そこで，1 つの解決案として，最も重要な上位 c 個のバケットのみを用いることで，効率的に KL 情報量の近似を求める方法が考えられる．この手法を本論文では *NaiveSparse* と呼ぶ．

さらなる解決案として，本論文では SVD を用いた手法を提案する．この手法は，各ヒストグラムに対して SVD を用いて圧縮し，その後必要な係数値 c 個のみを蓄積して利用するというものである．この手法は後に示すとおり，探索処理を高速化するだけでなく，メモリ量も削減することができる．ただし，ヒストグラム $P = (p_1, \ldots p_m)$ に対して SVD の係数を算出するだけでは，KL 情報量の近似精度を向上させることはできない．そこで本論文では，ヒストグラムの対数（$\log p_1, \ldots \log p_m$）の SVD 係数も同時に蓄積することで，近似精度における問題を解決する．詳細は次のとおりである．

$\hat{P} = (\hat{p}_1, \ldots, \hat{p}_m)$ をヒストグラム $P = (p_1, \ldots, p_m)$ の各バケットの値の対数をとったもの，つまり，$\hat{p}_i = \log p_i$ とする．さらに，S_p と $\hat{S}_p$ をそれぞれ，P と $\hat{P}$ の SVD 係数とする．提案手法は，この S_p と $\hat{S}_p$ を用いて処理を行う．

提案手法（D-Search）：提案手法は，それぞれの分布データ P とその対数 $\hat{P}$ に対してまず SVD の係数値 S_p と $\hat{S}_p$ を計算し，その後これらの中から必要な係数値上位 c 個のみを用いて KL 情報量の近似計算を行う．最後に，近傍探索等のアルゴリズムへ適用する．

以下に示す定理 1 は提案手法の基盤要素である．本手法では，SVD の係数を用いることで対称 KL 情報量の効率的な計算を実現する．

定理 1　$S_p = (sp_1, \ldots, sp_m)$，$\hat{S}_p = (\hat{sp}_1, \ldots, \hat{sp}_m)$ をそれぞれ，P，$\hat{P}$ の SVD の係数とすると以下の式が得られる．

$$d(P, Q) = \frac{1}{2} \sum_{i=1}^{m} f_{pq}(i) \tag{4}$$

$$f_{pq}(i) = (sp_i - \hat{sq}_i)^2 + (sq_i - \hat{sp}_i)^2 - (sp_i - \hat{sp}_i)^2 - (sq_i - \hat{sq}_i)^2.$$

証明：定義より，

© 2009 Information Processing Society of Japan

$$d(P,Q) = \sum_{i=1}^{m}(p_i - q_i)\cdot\log\left(\frac{p_i}{q_i}\right).$$

さらに，上記の式より次式が得られる．

$$d(P,Q) = \sum_{i=1}^{m}(p_i - q_i)\cdot(\log p_i - \log q_i)$$

パーセバルの定理により，

$$\sum_{i=1}^{m}(sp_i - \hat{sq}_i)^2 = \sum_{i=1}^{m}(p_i - \log q_i)^2,\quad \sum_{i=1}^{m}(sq_i - \hat{sp}_i)^2 = \sum_{i=1}^{m}(q_i - \log p_i)^2$$

$$\sum_{i=1}^{m}(sp_i - \hat{sp}_i)^2 = \sum_{i=1}^{m}(p_i - \log p_i)^2,\quad \sum_{i=1}^{m}(sq_i - \hat{sq}_i)^2 = \sum_{i=1}^{m}(q_i - \log q_i)^2$$

であるため，

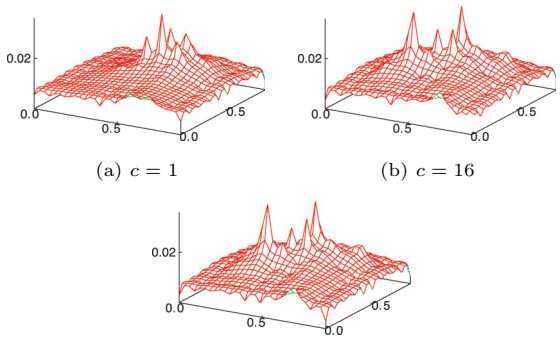
$$d(P,Q) = \frac{1}{2}\sum_{i=1}^{m} f_{pq}(i).$$

したがって，定理 1 が成り立つ．□

ヒストグラムの SVD 係数を用いた KL 情報量の近似計算は，式 (4) から得られる．ヒストグラム中のバケットの数 m は，次元数が増加するに従い非常に大きくなるが，その一方で，一般的に実データには偏りがある．そこで本手法では，SVD 係数のうちの限定した個数のみを用いることで，効率的な KL 情報量の近似計算を行う．オリジナルのヒストグラムを用いて得られた SVD 係数から，固有値が大きい順に上位 c 個の係数値を選ぶ．固有値の大きい基底を選ぶことで，少ない係数の数で高い精度を保つことができる．

式 (4) から，上位 c 個の SVD 係数を用いることで KL 情報量の計算を行うことができる．具体的には，上位 c 個に該当する係数 sp_i，sq_i，$\hat{sp}_i$，$\hat{sq}_i$（$i = 1, ..., c$）についてのみ $f_{pq}(i)$ を計算し，それ以外の係数については係数はほとんど 0 に近い値をとるため計算を省く．このようにして，最終的に P と Q の KL 情報量の近似式は次のようにして得られる．

$$d'_c(P,Q) = \frac{1}{2}\sum_{i=1}^{c} f_{pq}(i). \tag{5}$$



(a) $c = 1$　(b) $c = 16$

(c) *Original data*

図 2　SVD 係数を用いた確率密度分布の近似
Fig. 2　Approximation of probability distribution.

このとき特に $m = c$ の場合，次式が得られる．

$$d'_m(P,Q) = d(P,Q).$$

3.2.2　Multi-Step Sequential Scan

D-Search では分布データ探索において，類似性の低い分布データを枝刈りすることによって効率的な探索処理を可能とする．本論文では，この手法を multi-step sequential scan と呼ぶ．

c 個の SVD 係数を用いた KL 情報量の近似計算処理により高速な類似探索が可能になるが，近似計算には近似精度と計算時間のトレードオフが存在する．すなわち使用する SVD 係数を過度に縮退させると近似計算のコストは小さくなるが，誤差が大きくなってしまう．そのため D-Search では探索処理において，使用する SVD 係数の数（すなわち c）を徐々に大きくしていき，KL 情報量の近似の精度を上げていく．

図 2 は *MoCap* の running のモーションに対する確率密度分布である．図 2 の例では，2 次元の分布データを 10×10 個のバケットに分割してヒストグラムを作成し，それらに対して SVD を用いて基底変換を施す．図 2 の場合，SVD の係数値の最大個数は $c_{max} = m = 100$ となるが，ランク r は一般に c_{max} より少ない値をとる．そこで本手法では，限られた係数値のみを使用して探索を行うことで，効率的な計算処理を実現する．本手法は，各ステップごとに近似精度を徐々に向上させて，類似度の低い分布を枝刈りしながら効率良く探索処理を行っていく．まず最初のステップでは，図 2 (a) のような少ない係数値を用いて KL 情報

© 2009 Information Processing Society of Japan

量の近似を荒い精度で計算する．ここで得られた KL 情報量の近似値が，現時点での k-近傍探索の解候補の類似距離 d_{cb} よりも遠い場合は，枝刈り処理を行いその後の探索処理を効率化する．そして次のステップに進み，枝刈りされなかったデータに対してさらに多くの係数値を用いて（図 2 (b)），より精度の高い KL 情報量の近似を計算していく．最終的に，各ステップの中で枝刈りされなかったデータについては，オリジナルのヒストグラム（図 2 (c)）を用いて厳密な KL 情報量を計算する．

より具体的には，以下の処理を各ステップごとに繰り返し，類似度の低いデータを徐々に枝刈りしながら最終的な解を算出する．

- まず初めに，上位 c 個の SVD 係数を用いて KL 情報量の近似を計算し，それらの値を用いて k-近傍探索の近似の解候補（NNL'）を集める．
- 次に，近似の解候補（NNL'）と問合せ分布間の厳密な KL 情報量を計算し，もし現時点での厳密な解候補の距離 d_{cb} よりも類似度の近い分布があった場合，厳密な解候補（NNL）を更新する．
- すべての分布データについて，もし現時点での厳密な解候補の距離 d_{cb} よりも KL 情報量の近似値が大きい値をとる場合，それらの分布は k-近傍探索の結果として適していないため破棄し，それ以降の探索処理では KL 情報量の計算を行わない．

本手法では上記の処理を h 回繰り返すことで効率的な探索を行う．ここで，各ステップにおける係数の数は等差数列に基づいて，$c_i := i \cdot c_1$（$i = 1, 2, ..., h$）とする．

図 3 は，3.2.1 項で述べた KL 情報量の近似計算を用いた探索アルゴリズムである．最初に与えられた問合せ分布 Q の SVD の係数値を計算し，その後各ステップごとに KL 情報量の近似計算を行い解候補となる類似分布を抽出する．このとき，解候補とならない分布データについては枝刈りを行い，探索を効率化する．探索アルゴリズムは，ステップが 1 つ上がるごとに少しずつ KL 情報量の近似精度を上げていき，類似度の低い分布データの枝刈りを行うことで，KL 情報量の計算コストを削減する．すべてのステップにおいて枝刈り処理を終えたら，最後に残った分布データに関して厳密な KL 情報量の計算を行い，最終的な解を算出する．

本項では k-近傍探索の処理についてのみ述べたが，D-Search は範囲探索も同様に処理することができる．範囲探索の場合は探索の閾値を ϵ に置き換えて KL 情報量が ϵ 以下の分布データすべてを検出する．

3.3　時系列分布データのマイニング

時系列分布データは，様々な分野で発生するデータである．たとえば，金融分野における株価のデータや，気象分野における温度の数値データ，商品の売り上げに関するカテゴリカルなデータ等，様々なものが存在している．時系列分布データは，時間ごとのトレンドやパ

```
Algorithm D-Search(Q, k)
    compute the SVD coefficients S_Q and Ŝ_Q from Q
    // NNL is the sorted nearest neighbor list
    initialize NNL;
    // Multi-step sequential scan
    for i := 1 to h do
        Scan(Q, k, c_i);
    endfor
    // Compute exact KL divergence
    for each P ∈ database do
        compute d(P, Q);
        if d(P, Q) ≤ d_cb then
            add P to NNL, and update d_cb;
        endif
    endfor
    return NNL;

Algorithm Scan(Q, k, c)
    // NNL' is the sorted nearest neighbor list
    initialize NNL';
    // Compute approximate KL divergence
    for each P ∈ database do
        compute d'_c(P, Q);
        if d'_c(P, Q) ≤ d'_cb then
            add P to NNL', and update d'_cb;
        endif
    endfor
    //Compute exact KL divergence
    for each P ∈ NNL' do
        compute d(P, Q);
        if d(P, Q) ≤ d_cb then
            add P to NNL, and update d_cb;
        endif
    endfor
    //Prune the search candidates
    for each P ∈ database do
        if d'_c(P, Q) ≥ d_cb then
            remove P from database;
        endif
    endfor
    return NNL;
```

図 3　分布データ集合のための探索アルゴリズム
Fig. 3　Search algorithm for distribution data sets.

© 2009 Information Processing Society of Japan

ターンを表現しており，これらのデータの解析を行うことで，ユーザは様々な情報を取得することができる．本節では，D-Search の拡張版として，時系列分布データの中から類似部分シーケンスを抽出する手法を提案する．

時系列分布データとは離散的なヒストグラムのシーケンス $P_1, P_2, \ldots, P_n$ であり，ここで P_i は時刻 i の分布ヒストグラム，P_n は最も新しい分布ヒストグラムである．時刻 i から j までの分布ヒストグラムを $P_{(i,j)}$ とするとき，本節で扱う問題は以下のとおりである．

問題 2（時系列分布データ探索）　分布シーケンス $P_1, P_2, \ldots, P_n$，問合せ分布データ Q が与えられたとき，Q に対して KL 情報量が最小となる k 個の部分シーケンスを検出する．

時系列分布データは，時間ごとに頻度分布が変化し，異なる時間幅の中にそれぞれ特徴の異なる分布が並んでいると考えられる．このような分布シーケンスの中からパターンを発見するためには，様々なウィンドウサイズの部分シーケンスの中から，いかに効率良く類似分布の探索を行うかを考慮する必要がある．まず第 1 に考えられるのは，等差数列を用いて，$w_0, w_0+1, w_0+2, w_0+3, \ldots$ というウィンドウサイズそれぞれに対して部分シーケンスを生成し，類似分布を探索する手法がある．しかし，この手法は，大量の部分シーケンスに対して処理が必要となるため，効率的な探索とはいえない．そこで本論文では等比数列を用いて $w_0, 2w_0, 4w_0, \ldots$，より一般的には $w_l = w_0 \cdot W^l$（$l = 0, 1, 2, \ldots$）という条件のもと，効率的に部分シーケンスの探索を行う[14]．本手法を用いると，ウィンドウのセット $\mathcal{W}$ の数を大きくすることなく，柔軟なウィンドウサイズでの探索処理を実現することができる．

各レベル l において，分布シーケンスをウィンドウサイズ w_l の部分シーケンスに分割し，部分シーケンス各々について KL 情報量を計算する．KL 情報量の計算には SVD に基づく近似手法を用いる．探索処理では multi-step sequential scan によって類似部分シーケンスを効率的に検出する．提案手法のアイデアを**図 4** に例示する．図では，問合せ分布と類似する部分シーケンスとして，レベル 0（$l = 0$）の階層の $P_{(8,12)}$，$P_{(12,16)}$，レベル 1（$l = 1$）の $P_{(8,16)}$ が検出される（図中のグレーで示された部分シーケンス）．

提案手法はさらに，オプションとしてスライディングウィンドウを用いることもできる．スライディングウィンドウは，時系列処理において一般的なモデルであり[8),24)]，この手法を用いると，ウィンドウサイズの大きさ以下の小さな時間の遅延や変動をとらえながら類似部分シーケンスを検出することが可能になる．

3.4　理論的な分析

n を分布データの数，m をヒストグラムのバケットの数，c を使用する SVD 係数の数とする．本節では探索処理における空間計算量および時間計算量を考察する．

本論文では固定した分布データ集合を探索対象としているため，n 個の分布データの SVD，および射影のための計算は問合せ分布データを受け取る前に実行することができる．したがって，n 個の分布データの SVD 係数を得るための処理は探索処理に含めないものとする．

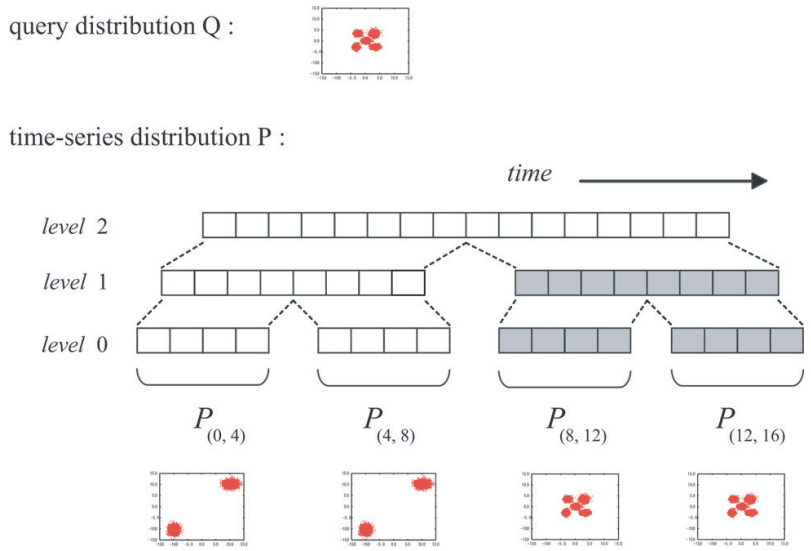


図 4　時系列分布の探索
Fig. 4　Time-series distribution search.

3.4.1　空間計算量

ナイーブな手法（Naive Dense）

補助定理 1　ナイーブな手法は探索処理に $O(mn)$ のメモリ量を要する．

証明：ナイーブな手法は，m 個のバケットに分割された n 個のヒストグラムの配列を保持するため，$O(mn)$ のメモリ使用量が必要となる．□

提案手法（D-Search）

補助定理 2　提案手法は探索処理に $O(m+n)$ のメモリ量を要する．

証明：提案手法は，最初に問合せ分布データの m 個のバケットに分割されたヒストグラムを射影し，$O(c)$ 個の SVD 係数を得る．探索処理では，問合せデータと n 個すべての分布データとの KL 情報量を計算するため，全部で $O(cn)$ のメモリ量となる．ただしここで，c は一般に非常に小さな定数であるため，無視することができる．最終的に，$O(m+n)$ のメモリ量が必要となる．□

© 2009 Information Processing Society of Japan

3.4.2　時間計算量

ナイーブな手法（Naive Dense）

補助定理 3　ナイーブな手法は，k 近傍探索における KL 情報量の計算に $O(nm)$ の時間を要する．

証明：分布データのペア 1 つあたりの KL 情報量の計算には $O(m)$ の時間を要する．そのため n 個の分布データの探索には，$O(mn)$ の計算時間が必要となる．□

提案手法（D-Search）

補助定理 4　提案手法は，k 近傍探索における近似計算に $O(n)$ の時間を要する．

証明：c 個の SVD 係数を用いて n 個の分布データの k 近傍探索をする場合，$O(cn)$ の時間を要する．ここで c は非常に小さな値をとるため，計算量は単純に $O(n)$ と表せる．□

補助定理 3 と補助定理 4 は，提案手法が KL 情報量の計算コストを大幅に削減していることを示している．実用面では，提案手法の探索コストは枝刈りの効果に依存する．次章では，様々な数値およびカテゴリデータを用いた実験によって，提案手法が精度を犠牲にすることなく大幅な探索コストの低減化を達成していることを示す．

4. 評 価 実 験

D-Search の有効性を検証するため，実データを用いた実験を行った．実験は 4 GB のメモリ，Intel Core 2 Duo 1.86 GHz の CPU を搭載した Linux マシン上で実施した．

3.4 節でも述べたとおり，本論文では固定した分布データ集合を探索対象としているため，n 個の分布データの SVD，および射影のための計算は問合せ分布データを受け取る前に実行することができる．したがって，n 個の分布データの SVD 係数を得るための処理時間は探索処理時間に含めないものとする．

評価実験では，以下の諸問題に取り組む．

- 時系列分布データにおけるパターン発見
- 大量分布データの探索における計算時間の検証
- SVD に基づく KL 情報量の近似精度と枝刈りの効果の検証

4.1　時系列分布データにおける類似分布の探索

本節では，時系列分布データにおける D-Search の有効性を立証するため，実データを用いたケーススタディを示す．**図 5** は D-Search が適合する部分シーケンスを検出している様子を示している．D-Search による類似部分シーケンスの検出結果は，ナイーブな方法による出力結果と完全に一致しており，類似分布の探索に対し高い精度を保つことが示された．

MoCap

MoCap は，1 秒 120 フレームでヒトの動きを計測したモーションキャプチャのデータセットである[1]．実験で利用したデータは，ヒトの典型的な動き（walking，running，twisting，jogging，jumping 等）がランダムに繰り返されるというものであり，それぞれ，各フレームごとの両足各部の運動エネルギーの値を 4 次元の分布データに変換して使用した．

図 5 (a) に示すとおり，D-Search は類似分布を検出していることが分かる．たとえば，D-Search は問合せ分布データ（data No.14-20，2-4sec）に対し，分布データ 1（data No.14-14，2-4sec.）を類似分布として検出したが，分布データ 1 と同じデータシーケンス中の違うフレーム区間である分布データ 2（data No.14-14，4-6sec.）は，類似分布として検出しなかった．実際，検出された部分シーケンスを確認してみると，分布データ 1 は，問合せ分布データと同じ jumping のモーションだったのに対し，分布データ 2 は，問合せ分布データと異なり jogging のモーションであった．

EEG

このデータセットは，アルコール依存症の遺伝的素因と EEG の関連性の大規模調査の結果である[2]．各データは，1 秒間に 256 Hz（3.9-msec epoch）で 64 カ所の電極の測定値を持ち，そのシーケンス長は 256 である．すべてのデータはアルコール依存傾向（alcohol），非アルコール依存傾向（control）の 2 つのグループに分けられる．

D-Search は分類問題にも非常に有効である．図 5 (b) に示すように，提案手法は問合せ分布データ（co2a0000364 の部分シーケンス 36-37sec.）と分布データ 1（co3a0000451 の部分シーケンス 56-57sec.）を同じグループと判定し，分布データ 2（co2c0000364 の部分シーケンス 75-76sec.）は別グループであると判定した．実際前者のグループについて検証してみると，問合せと分布データ 1 はともに alcohol の分布だったのに対し，分布データ 2 は別のカテゴリである control の分布データとなった．

Ondemand TV

このデータセットは，オンラインの TV 配信サービスの視聴に関するデータである．このデータは 6 カ月（2007/5/14-2007/11/15）の期間の番組視聴履歴であり，ランダムに選出された 10,000 名の匿名ユーザそれぞれに対し，どの番組をいつ視聴したかの情報が蓄積されている．本実験では，各番組ごとにどのユーザが何回その番組を視聴したかという情報をカテゴリカルな分布データに変換して用いた．

図 5 (c) に示すとおり，提案手法は類似する TV 番組を検出することができた．たとえば，問合せ分布「全豪オープンテニス 2007 女子決勝（放送期間：2007/2/1-2008/4/1）」に対し，最もユーザの分布が類似しているものは，分布データ 1 「全豪オープンテニス 2007 女子準決勝（放送期間：2007/2/1-2008/4/1）」であるという結果になり，一方で，分布データ 2 「デスパレートな妻たちシーズン 1（放送期間：2006/8/21-2008/12/5）」のような他の番組は，類似分布として検出されなかった．

© 2009 Information Processing Society of Japan

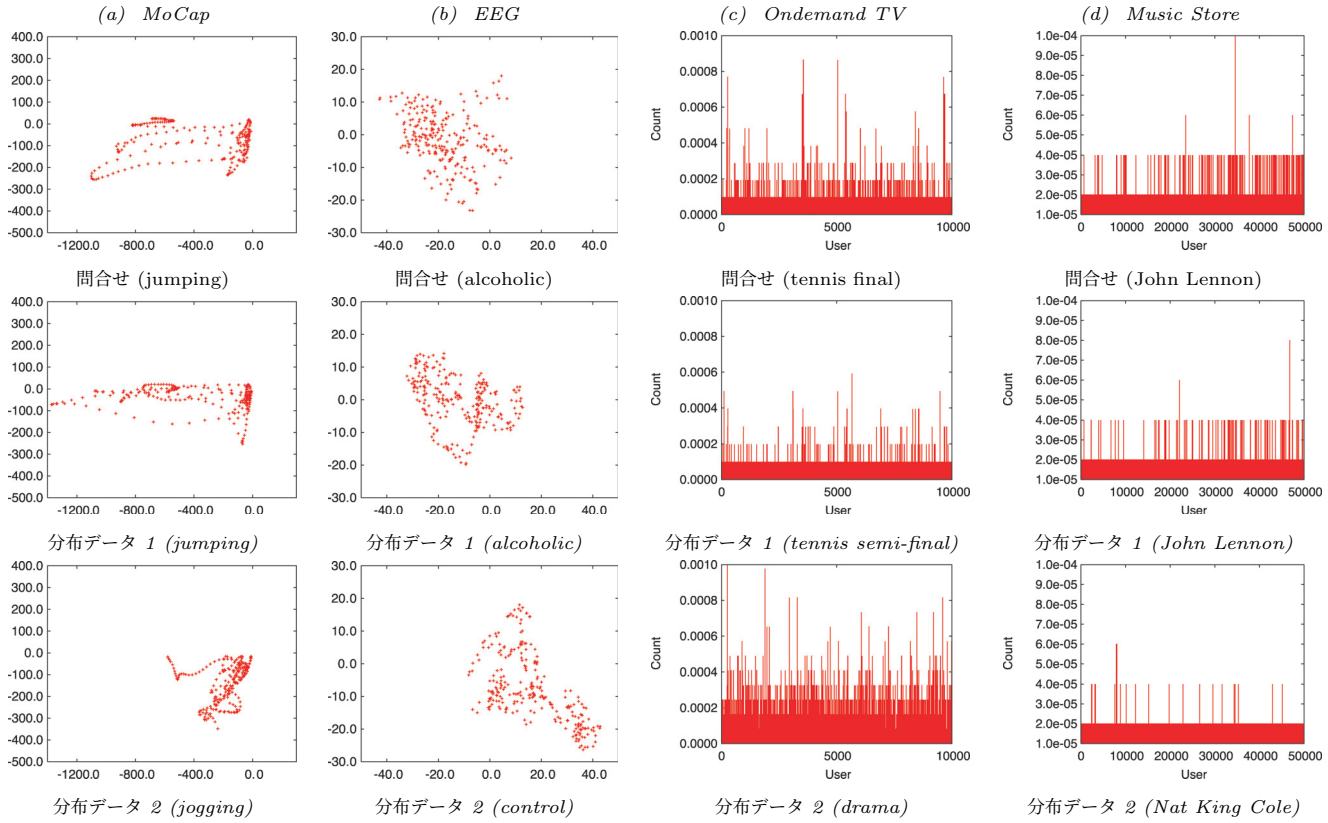


図 5　*MoCap*，*EEG*，*Ondemand TV*，*Music Store* を用いた類似部分シーケンスの発見．上の図が問合せ分布データであり，中央と下の図がそれぞれ類似した分布，類似していない分布を示している

Fig. 5　Discovery of subsequence in *MoCap*, *EEG*, *Ondemand TV* and *Music Store*. The *Query*, *Distribution 1*, *Distribution 2* show the query distributions and similar distributions and dissimilar distributions, respectively. In experiments of *MoCap*, *EEG*, *Ondemand TV* and *Music Store*.

Music Store

このデータセットは，*Music Store* における，16 カ月間（2005/4/4-2006/7/1）の楽曲購入履歴である．購入履歴データは，ユーザ ID（50,000 名の匿名ユーザ），楽曲 ID（全 43,896 曲），購入日時の 3 つの属性を持つ．本実験では，期間ごとの各楽曲における購入ユーザの頻度分布を作成して利用した．

図 5 (d) は，D-Search が正しく類似楽曲グループを発見できたことを示している．たとえば，問合せとして「ジョン・レノンの Woman」の楽曲を設定すると，図 5 (d) 中の分布データ 1 「ジョン・レノンの Love」が類似楽曲として検出された．一方，分布データ 2 「ナッ

© 2009 Information Processing Society of Japan

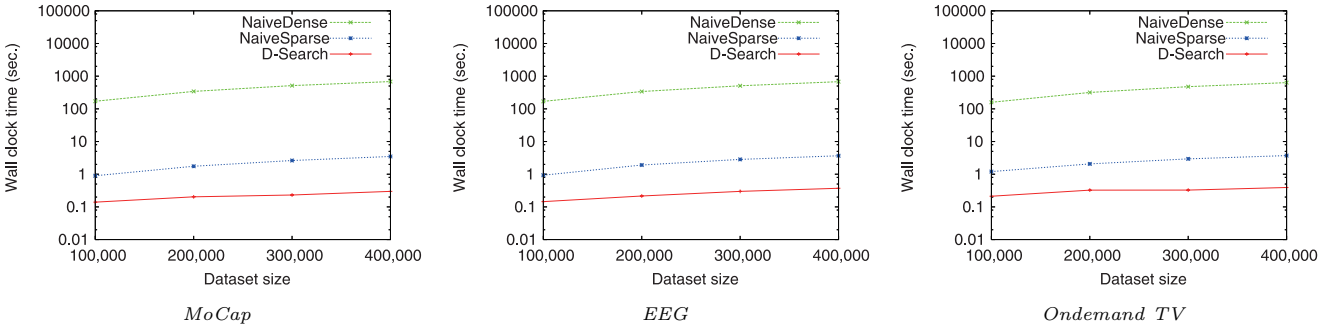


図 6　データ数を変化させた場合の探索時間
Fig. 6　Scalability: wall clock time vs. dataset size n (= number of distributions).

ト・キング・コールの L-O-V-E」のようなジャンルの違う楽曲は類似分布として検出されなかった．

4.2　性　　能

本節では，以下の問題に対し検証を行う．

- **速度：**提案手法と 2 種のナイーブな手法での探索速度の比較（*NaiveDense*：m 個のバケットすべてを用いて KL 情報量を計算する手法，*NaiveSparse*：m 個のバケットのうち，値の大きい順に上位 c 個のみを用いて KL 情報量を計算する手法）
- **精度と効果：**SVD を用いた KL 情報量の近似精度と枝刈り効果について，提案手法と *NaiveSparse* との比較

ただし，*NaiveSparse* の探索処理については，提案手法の一部である Multi-Step Sequential Scan の手法を適用する．

4.2.1　探索処理時間

探索処理の性能を検証するため，D-Search，ナイーブな手法（*NaiveDense*，*NaiveSparse*）の 3 手法についての比較を行う．本論文では，分布データセットのサイズを変化させた場合についての実験を行った．**図 6** は，D-Search とナイーブな手法での計算コストを示している．データセットのサイズ n は，100,000 から 400,000 まで変動させた．本実験では，ヒストグラムのバケット数を $m = 10,000$ とし，探索の初期係数を $c_1 = 50$，ステップ数を $h = 4$ とした．速度の計測には，上位 $k = 1$ 個の類似分布の探索 100 回の試行の平均を用いた．

図 6 は，データセットのサイズ n に対する探索時間である．結果を見ると D-Search は *NaiveDense* と比較して約 2,300 倍，*NaiveSparse* と比較して約 10 倍の高速化を達成している．特に，*NaiveSparse* の探索処理については，提案手法の一部である Multi-Step Sequential Scan の手法を適用している．すなわち，単純に上位 c 個のバケットのみを用いた *NaiveSparse* の近似よりも SVD を使う提案手法のフィルタリング効果がより大きいことが明確となっている．

探索の高速化に加え，提案手法は高い探索精度を示している．D-Search によって検出された類似分布は，2 種のナイーブな手法（*NaiveDense*，*NaiveSparse*）によって検出された類似分布と完全に同じものであることが確認された．

4.2.2　提案手法の性能

ここでは，(1) SVD を用いた KL 情報量の近似精度，(2) 枝刈りの効果の 2 つの課題の検証を行う．

まず (1) の検証として，**図 7** に近似精度と計算コストに関する実験結果を示す．x 軸は KL 情報量の計算コストを示し y 軸は厳密な KL 情報量と近似 KL 情報量との誤差率を示す．図 7 は D-Search と NaiveSparse の比較結果であり，係数値の数を増減させることにより，KL 情報量の計算時間と近似精度を変化させている．図は，精度と計算時間がトレードオフの関係にあることを示しているが，それと同時に，D-Search と *NaiveSparse* を比較すると同じ誤差率でも D-Search のほうがより高速に処理ができることが分かる．

次に (2) の検証として，**図 8** にデータサイズ n =100,000 における各ステップでの枝刈り効果についての実験結果を示す．図 8 に示すとおり，各ステップ $c = \{50, 100, 150, 200\}$ において KL 情報量計算の総数は徐々に減少していき，最終的にはほとんどのデータについて，近似 KL 情報量のみで枝刈りすることに成功している．近似精度が荒いと計算コストは低く抑えられるが，精度を上げると計算コストは非常に大きくなる．提案手法は，徐々

© 2009 Information Processing Society of Japan

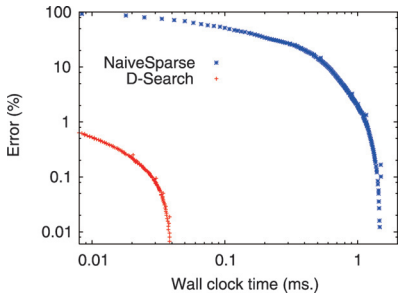


図 7　KL 情報量近似の精度と計算時間
Fig. 7　Approximation quality.

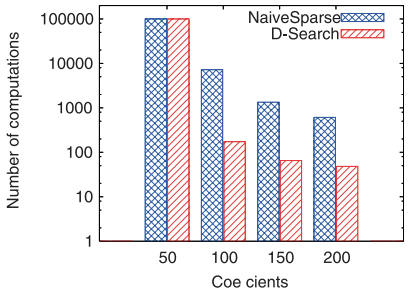


図 8　枝刈りの効果：各ステップごとの計算個数
Fig. 8　Pruning quality.

に近似精度を向上させながら枝刈りをしていくことで，精度を犠牲にすることなく大幅なコスト削減に成功している．さらに，SVD を用いた提案手法は，*NaiveSparse* と比較してより効果的に枝刈りが行えることが分かる．

5. おわりに

本論文では，分布データのための探索問題について扱い，その問題を解決するための手法である D-Search を提案した．ナイーブな手法と異なり，D-Search は，SVD による KL 情報量の計算の高速化と，類似度の低い分布の枝刈り操作により，精度を犠牲にすることなく適切な類似分布データを高速に探索する．さらに本論文では D-Search の拡張版として，時系列分布データにおける類似シーケンス探索手法の提案も行った．本論文では理論的な検証だけでなく，実データを用いた実験でも提案手法の優位性を確認しており，ナイーブな手法と比べて約 2,300 倍まで高速に探索できることを確認した．

参　考　文　献

1) CMU Graphics Lab Motion Capture Database. http://mocap.cs.cmu.edu/
2) UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/
3) Barbic, J., Safonova, A., Pan, J.-Y., Faloutsos, C., Hodgins, J.K. and Pollard, N.S.: Segmenting Motion Capture Data into Distinct Behaviors, *Graphics Interface*, pp.185–194 (2004).
4) Bay, S.D., Kibler, D.F., Pazzani, M.J. and Smyth, P.: The UCI KDD Archive of Large Data Sets for Data Mining Research and Experimentation, *SIGKDD Explorations*, pp.81–85 (2000).
5) Cheng, R., Kalashnikov, D.V. and Prabhakar, S.: Evaluating Probabilistic Queries over Imprecise Data, *Proc. ACM SIGMOD*, San Diego, California, pp. 551–562 (2003).
6) Cheng, R., Xia, Y., Prabhakar, S., Shah, R. and Vitter, J.S.: Efficient Indexing Methods for Probabilistic Threshold Queries over Uncertain Data, *Proc. VLDB*, Toronto, Canada, pp.876–887 (2004).
7) Fischer, S., Lienhart, R. and Effelsberg, W.: Automatic Recognition of Film Genres, *ACM Multimedia*, pp.295–304 (1995).
8) Gao, L. and Wang, X.S.: Continuous Similarity-Based Queries on Streaming Time Series, *IEEE Trans. Knowl. Data Eng.* (*TKDE*), pp.1320–1332 (2005).
9) Guo, Z., Zhang, Z., Xing, E.P. and Faloutsos, C.: Enhanced max margin learning on multimodal data mining in a multimedia database, *KDD*, pp.340–349 (2007).
10) Huang, X., Li, S.Z. and Wang, Y.: Jensen-Shannon boosting learning for object recognition, *Proc. IEEE Computer Society International Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vol.2, pp.144–149 (2005).
11) Ishikawa, Y., Machida, Y. and Kitagawa, H.: A Dynamic Mobility Histogram Construction Method Based on Markov Chains, *Proc. Int. Conf. on Statistical and Scientific Database Management* (*SSDBM*), pp.359–368 (2006).
12) Li, C., Zhai, P., Zheng, S.-Q. and Prabhakaran, B.: Segmentation and recognition of multi-attribute motion sequences, *ACM Multimedia*, pp.836–843 (2004).
13) Manning, C.D. and Schütze, H.: *Foundations of Statistical Natural Language Processing*, The MIT Press (1999).
14) Papadimitriou, S. and Yu, P.S.: Optimal Multi-Scale Patterns in Time Series Streams, *SIGMOD*, pp.647–658 (2006).
15) Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P.: *Numerical Recipes in C*, 2nd edition, Cambridge University Press (1992).
16) Raymer, M.L., Doom, T.E., Kuhn, L.A. and Punch, W.F.: Knowledge discovery

© 2009 Information Processing Society of Japan

in medical and biological datasets using a hybrid Bayes classifier/evolutionary algorithm, *IEEE Trans. Systems*, pp.802–813 (2003).
17) Rubner, Y., Tomasi, C. and Guibas, L.J.: The Earth Mover's Distance as a Metric for Image Retrieval, *Int. J. Comput. Vision*, Vol.40, No.2, pp.99–121 (2000).
18) Sakurai, Y., Chong, R., Li, L. and Faloutsos, C.: Efficient Distribution Mining and Classification, *SIAM International Conference on Data Mining* (*SDM*), pp.632–643 (2008).
19) Sun, Z.: Adaptation for multiple cue integration, *Proc. IEEE Computer Society International Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vol.1, pp.440–445 (2003).
20) Sykacek, P. and Roberts, S.J.: Adaptive Classification by Variational Kalman Filtering, *NIPS*, pp.737–744 (2002).
21) Tao, Y., Cheng, R., Xiao, X., Ngai, W.K., Kao, B. and Prabhakar, S.: Indexing Multi-Dimensional Uncertain Data with Arbitrary Probability Density Functions, *Proc. VLDB*, Trondheim, Norway, pp.922–933 (2005).
22) Traina, A., Traina, C., Papadimitriou, S. and Faloutsos, C.: Tri-Plots: Scalable Tools for Multidimensional Data Mining, *KDD* (2001).
23) Vert, J.-P.: Adaptive context trees and text clustering, *IEEE Trans. Information Theory*, Vol.47, No.5, pp.1884–1901 (2001).
24) Zhu, Y. and Shasha, D.: StatStream: Statistical Monitoring of Thousands of Data Streams in Real Time, *VLDB*, pp.358–369 (2002).

(平成 21 年 3 月 20 日受付)
(平成 21 年 7 月 10 日採録)

(担当編集委員　立石 健二)

松原　靖子
2006 年お茶の水女子大学理学部情報科学科卒業．2009 年同大学院人間文化創成科学研究科理学専攻博士前期課程修了．2009 年より京都大学大学院情報学研究科社会学情報専攻博士後期課程に在籍．データストリーム処理に関する研究に従事．日本データベース学会学生会員．

櫻井　保志（正会員）
1991 年同志社大学工学部電気工学科卒業．1991 年日本電信電話（株）入社．1999 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（工学）．2004〜2005 年カーネギーメロン大学客員研究員．本会平成 18 年度長尾真記念特別賞，本会平成 16 年度および平成 19 年度論文賞，電子情報通信学会平成 19 年度論文賞，日本データベース学会上林奨励賞，KDD 2008 best paper awards 等受賞．索引技術，データストリーム処理，センサーデータ処理技術の研究に従事．ACM，電子情報通信学会，日本データベース学会各会員．

吉川　正俊（正会員）
京都大学大学院工学研究科博士後期課程修了．工学博士．京都産業大学，奈良先端科学技術大学院大学，名古屋大学を経て 2006 年より京都大学大学院情報学研究科教授．この間，南カリフォルニア大学客員研究員，ウォータルー大学客員准教授．The VLDB Journal および Information Systems（Elsevier/Pergamon）の編集委員．XML データベース，異種情報源の統合等の研究に従事．電子情報通信学会，ACM，IEEE Computer Society 各会員．

© 2009 Information Processing Society of Japan