様式 C-19

# 科学研究費補助金研究成果報告書

平成２４年　５月２０日現在

機関番号：１１３０１
研究種目：若手研究（B）
研究期間：2009　～2010
課題番号：２１７００００４
研究課題名（和文）　階層的セグメンテーションルールを用いた数理モデルの構築と知識抽出への応用
研究課題名（英文）　Construction for numeric data model using layard structure segmentation rule and Application to knowledge extraction
研究代表者
全　眞嬉（CHUN JNHEE）
東北大学・大学院情報科学研究科・准教授
研究者番号：８０４３１５５０

研究成果の概要（和文）：
本研究では階層的セグメンテーション理論研究として数値データに対する階層化ルール理論研究としてデジタルな星領域および山方地形図の最適近似アルゴリズムを与えた．数値データに対する階層化ルール理論に対する定式化に関する研究は非交差図形の認識アルゴリズムの設計を行った．長方形やタブローの隅の位置が与えられる場合に動作する効率的なアルゴリズムの設計を行った．研究成果を国際会議で積極的に発表を行った．

研究成果の概要（英文）：
In this research gives optimal approximation algorithms for pyramid and digital ray on numerical data as a theoretical hierarchical segmentation. Research on formulating the theory of hierarchical rule for the numerical data of the recognition algorithm is designed in a non-cross shape. We designed an efficient algorithm to work if given the position of the corner of the rectangle or tableau. We presented research at international conferences.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 2009 年度 | 1,800,000 | 540,000 | 2,340,000 |
| 2010 年度 | 1,600,000 | 480,000 | 2,080,000 |
| 年度 | | | |
| 年度 | | | |
| 年度 | | | |
| 総　計 | 3,400,000 | 1,020,000 | 4,420,000 |

研究分野：総合領域
科研費の分科・細目：　情報学・情報学基礎
キーワード：アルゴリズム，計算理論，知識検出、情報システム，情報基礎

## １．研究開始当初の背景

データマイニングは大規模なデータから情報をコンパクトな知識として抽出する技術であり，情報化社会における最重点技術の一つである．近年，大規模なデータベースに蓄積されたデータから傾向や頻出パターンを法則として高速に効率的に発見するデータマイニングに注目が集まり，盛んに研究されている．

■現在の知識抽出の問題点
・現行のデータベース技術では知識を引き出すために必要な規則や価値などの自動的な抽出能力に乏しい．
・人工知能での知識獲得方法には処理速度に問題があり，大規模データマイニングに利用しにくい欠点がある．
・実用データマイニングで必須である数値データベースの取り扱いを考えると，数値デー

タの二値化誤差から生じる，正確性と学習汎用性のトレードオフに関する精度限界がある．

これらの問題を解決するために，2 次記憶上の巨大データベースを効率的に処理する最適化アルゴリズム研究が強く必要とされる．

２．研究の目的

データマイングの手法として「A ならば B である」といったデータベースの属性間の相関関係を求める結合ルールがあり，Apriori アルゴリズムが Agrawal らにより提案された．離散値属性間の結合ルールを求めるためのものであり，数値属性に対しては直接適用することはできない．なぜなら，数値属性は値に順序があるため，値を連続した区間で表すことが重要であるが，Apriori アルゴリズムではそのようなことは考慮していないからである．よって，数値属性を含む結合ルールを求めるには別のアルゴリズムを用いる必要がある．先行の研究では領域切り分けを用いて数値属性と離散値属性間の結合ルールの生成を行う手法が Fukuda らによって提案されている．この手法では「x∈R ならば B である」のようなルールを発見している．Fukuda らが提案した領域として切り出しを行うと，領域に入っているかいないかだけで判定を行うため，境界線上のデータと中央部のデータが同じ扱いをされるといった位置情報の損失(数値データの二値化誤差)がある．・　領域切り分けを用いて生成された結合ルールを学習に用いると，制限が強く，学習データに入っているノイズ（離れ値）を排除してしまうため，学習データ以外のデータに対して生成されるルールは不自然な領域ルールを生成するといった，過学習の問題点がある．また，属性を 3 つ以上与えると，計算量が多く計算が困難になる．すなわち，属性次元の制限の問題がある．

ここで，本来は切り取るべきものは多次元正規分布のような特徴をもったデータ分布であり，領域として切り取るより，性質の良い関数として捉える事が望ましい．データマイニング等によるデータ解析においては，入力データをシンプルな関数に近似して利用することが重要である．関数の近似においては様々なアプローチがあり，関数解析的な手法，学習による手法などとともに離散アルゴリズムを用いた計算幾何学的な最適近似のアプローチは広く研究されている．しかしながら，計算幾何学的な手法においては，従来の応用はパターンマッチ等であり，データ解析に用いる場合は定式化や最適化基準を適当なものに変更する必要がある．それに従ってアルゴリズム理論上の様々な困難は計算限界が生じ，それらを回避する必要がある．

３．研究の方法

本研究の最終的目的は，提案する数値データに対する階層化ルール理論のアルゴリズムの高速化と改良，学習において階層化ルールのエキスパートを用いたオンライン学習理論の研究を行い，過学習回避と予測精度を上げることである．

新しい領域族を階層最適化し，より高次元のルールの効率的なアルゴリズムの設計を行い，その結果を結合ルール生成だけでなく，データの視覚化及びデータマイニングへの幾何学的なアプローチにおいても有効に応用を行った．

自動的に抽出し表示された知識形態は，ユーザにより意志決定等の補助として用いられる．重要な条件は，抽出した知識形態がシンプルであり（単純性），正確にデータの特性を記述すること，知識としてサンプルに依存しない汎用性を持つ事さらにユーザにとって説得力があり，検証が容易であること（透明性）である．単純性と透明性の観点から，結合ルール及びそれを用いた決定木は有力な手法である．

本研究で提案する確率的な非決定性決定木構造を用いた階層構造は，現行の判定システムにおいて主流になっている決定論的な決定ルールに比較して，強いルールの影響を縮小する方法を適用する．本研究で提案する数値データに対する階層化ルールを用いることで拘束力の弱いルールで判定を行う，即ち非決定性を持たせた柔軟な決定システムの構築を行った．

４．研究成果

データマイニングにおいて，パターンマッチングなどでの計算幾何学手法の導入は与えられていた．しかしながら，数値データベースの幾何学的な相関の最適化を行うためにはアルゴリズム理論上の様々な困難性や計算限界が生じ，適切な定式化によりそれらの克服を行う必要がある．応募者は過去の研究において，最適階層構造を用いた結合ルールという知識の幾何学的表現法を提案し，国際的に高い評価を得た．

先行の研究では領域切り分けを用いて数値属性と離散値属性間の結合ルールの生成を行う手法が提案されたが、本来は切り取るべきものは多次元正規分布のような特徴をもったデータ分布であり，領域として切り取るより，性質の良い関数として捉える事が望ましい．データマイニング等によるデータ解析においては，入力データをシンプルな関数に近似して利用することが重要である．関数の近似においては様々なアプローチがあり，関数解析的な手法，学習による手法などとともに離散アルゴリズムを用いた計算幾何学

的な最適近似のアプローチは広く研究されている．しかしながら，計算幾何学的な手法においては，従来の応用はパターンマッチ等であり，データ解析に用いる場合は定式化や最適化基準を適当なものに変更する必要がある．それに従ってアルゴリズム理論上の様々な困難は計算限界が生じ，それらを回避する必要がある．

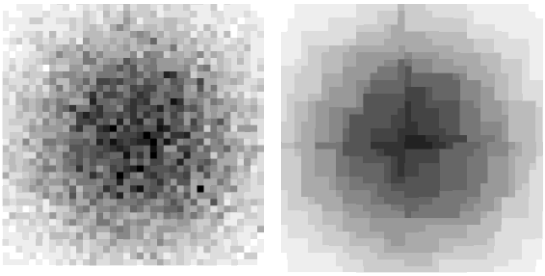

図１．関数として切り出す階層的セグメンテーション（２次元の例，左図は入力，右図は出力）

本研究では，階層的セグメンテーションの概念を用いて図１のような最適階層構造結合ルールの表現法を提案した．領域として切り出すのではなく，性質の良い関数として切り出しを行った．

デジタル線分族の場合の研究結果（雑誌論文[7]）は計算幾何分野のトップレベルの国際会議である SoGC で高い評価を得て，国際ジャーナル Discrete & Computational Geometry に招待され掲載された．

デジタル直線と対応するユークリッド線分の間の近似誤差は最大ハウスドルフ距離で評価し、n X n グリッド平面内での誤差に対し、漸近的に最適な　Θlog n の誤差限界を与えた．誤差限界の証明はディスクレパンシー理論とシンプルな構築アルゴリズムに基づいている．さらに、デジタル直線の単調性がなければ、誤差限界は O(1)に抑えられることを示した．図２はデジタル線分を利用した近似の入力（上）とその出力である．

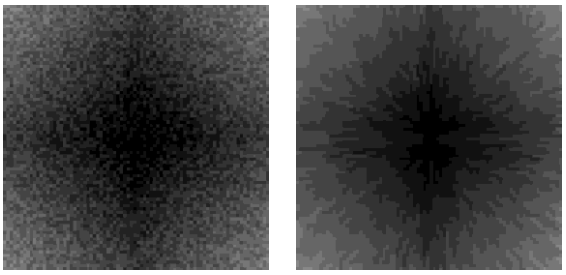

図２．デジタル線分を用いた近似の例（左図は入力，右図は出力）

本研究では，基本図形として長方形とタブローと呼ばれる図形を考え，この図形の非交差和領域を領域族とした最大重み領域問題を解くアルゴリズムを与えた．

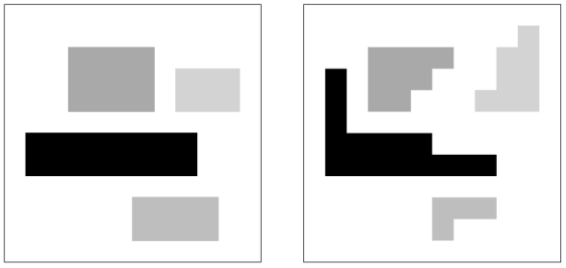

図３．図形の最適配置問題

タブローとは，ある点$p$を同じ隅に持つ長方形の和集合で表される領域のことである．これ以降では，この問題を各ピクセルが重みを持つ平面に，図３のように重みの総和が最大となる長方形やタブローを配置する問題として扱った．また，最適化の基準として，重みの総和を最大化する場合に加えて，最小値を最大化する場合についても考察した．

本研究では隅位置情報が与えられる場合の長方形とタブローの最適配置問題を解くアルゴリズムを提案した．

特に長方形の場合，隅位置情報を入力として与えないとき，最小値最大化問題も総和最大化問題も共に NP 困難である．

この２つの問題に対して，隅位置情報を与えることにより，最小値最大化問題は多項式時間で解くことができ，総和最大化問題は FPT であることを示した．従って，どちらの問題も効率的なヒューリスティックを設計することができると考えられる．

空間計算量に関しては，この問題を繰り返し解いたとしても，1 回解くのに必要な領域以上を使うことはない．よって，最小値最大化問題も総和最大化問題も $O(kn^2)$ の領域を用いて，アルゴリズムを反復できることを示した．

本結果は国際会議 COCOON2010（学会発表[4]）で発表し高い評価を得て国際ジャーナル DMAA （ Mathematics, Algorithms and Applications）に招待され掲載された（雑誌論文[1]）．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 7 件）

[1] Shinya Anzai, Jinhee Chun, Ryosei

Kasai, Matias Korman and Takeshi Tokuyama, “Effect of Corner Information in Simultaneous Placement of K Rectangles and Tableaux ”, Discrete Mathematics, Algorithms and Applications (DMAA), Vol.02, No. 4, pp.527-537(2010)（査読有）, Invited Paper

[2] Shinya Anzai, Jinhee Chun, Ryosei Kasai, Matias Korman and Takeshi Tokuyama, “Effect of Corner Information in Simultaneous Placement of K Rectangles and Tableaux”, Proceedings of the 16th International Computing and Combinatorics Conference (COCOON’10), Lecture Notes in Computer Science6196, pp.235–243 (2010)（査読有）

[3] Jinhee Chun, Yuji Okada and Takeshi Tokuyama, “ Distance Trisector of a Segment and a Point”, Interdisciplinary Information Sciences, Vol. 16, No. 1, pp.119-126 (2010)（査読有）

[4] Jinhee Chun, Ryosei Kasai, Matias Korman and Takeshi Tokuyama , “Algorithms for Computing the Maximum Weight Region Decomposable into Elementary Shapes”, Proceedings of the 20th Annual International Symposium on Algorithms and Computation (ISAAC’09), LNCS 5878, pp.1166-1174(2009)（査読有）

[5] Jinhee Chun, Ryosei Kasai, Matias Korman and Takeshi Tokuyama , “Algorithms for Optimal Segmentation of Regions Decomposable into Basic Shapes”, Proceedings of the 11th Japan-Korea Joint Workshop on Algorithms and Computation(WAAC’09),pp.138-145(2009)（査読有）

[6]全眞嬉, 葛西亮生, マティアス コルマン, 徳山豪, “基単調領域の非交差和領域の最適イメージ切り出しアルゴリズム”, 第 8 回 FIT 情報科学技術フォーラム 2009, Vol.8,RA-006, 37-42 項, 2009 年（査読有）

[7] Jinhee Chun, Matias Korman, Martin Nöllenburg and Takeshi Tokuyama, “Consistent digital rays”, Discrete and Computational Geometry , Vol.42, No.3, pp.359-378 (2009)（査読有）, Invited Paper

〔学会発表〕（計 13 件）

[1]安斎進也, 全眞嬉, 葛西亮生, コルマンマティアス, 徳山豪, “タブローの最適配置問題”, 情報処理学会アルゴリズム研究会, 2011 年 3 月 7 日, 沖縄

[2]安斎進也, 全眞嬉, 葛西亮生, コルマンマティアス, 徳山豪, “タブローの最適配置問題”, 冬のＬＡシンポジウム, 2011 年 2 月 2 日, 京都

[3]Jinhee Chun, Natsuda Kaothanthong, Takeshi Tokuyama , “Image Recognition And Retrieval By Using Distance Information”, 冬のＬＡシンポジウム, 2011 年 2 月 1 日, 京都

[4]Shinya Anzai, Jinhee Chun, Ryosei Kasai , Matias Korman and Takeshi Tokuyama, “Effect of Corner Information in Simultaneous Placement of K Rectangles and Tableaux”, 16th Annual International Conference on Computing and Combinatorics (COCOON 2010), 2010 年 7 月 19 日, ベトナム Nha Trang

[5] Shinya Anzai, Jinhee Chun, Ryosei Kasai, Matias Korman and Takeshi Tokuyama, “Algorithms for Image Segmentation Based on Geometric Approach”, Asian Association for Algorithms and Computation 2010 (AAAC'10), 韓国, 2010 年 4 月 18 日(査読有)

[6] 安斎進也, 全眞嬉, 葛西亮生, コルマンマティアス・徳山豪, “長方形やタブローの同時配置における隅位置情報の効果”, 電子情報通信学会コンピュテーション研究会, IEICE Technical Report, vol. 110, no. 12, COMP2010-5, pp. 33-38, 2010 年 4 月 22 日, 草津

[7] 安齋進也, 全眞嬉, 葛西亮生, コルマン

マティアス,徳山 豪,“画像切り出しに対するアルゴリズムの提案”,冬のＬＡシンポジウム，2010 年 2 月 2 日，京都

[8] Jinhee Chun, Ryosei Kasai, Matias Korman and Takeshi Tokuyama, “On simultaneous optimal segmentation of digital objects”, 電子情報通信学会コンピュテーション研究会, IEICE Technical Report, vol. 109, no. 235, COMP2009-32, pp. 1-8, 2009 年 10 月 16 日，仙台

[9] 全 眞嬉, 葛西 亮生, コルマン マティアス, 徳山 豪, “基単調領域の非交差和領域の最適イメージ切り出しアルゴリズム”, 第 8 回 FIT 情報科学技術フォーラム 2009, Vol.8,RA-006, 37-42 項, 2009 年 9 月 3 日, 仙台（査読有）

[10] 櫻庭敦之, 成田龍太, 全眞嬉, 徳山豪,“ジャストインタイムウェブ広告におけるタクソノミ自動生成手法”, 第 8 回 FIT 情報科学技術フォーラム 2009, D-026, 185-186 項，2009 年 9 月 3 日，仙台

[11]葛西 亮生, コルマン マティアス, 全 眞嬉, 徳山 豪,“基本図形に分割可能な領域の最適切り出しアルゴリズム”,LA シンポジウム 2009，2009 年 7 月 22 日，松島

[12]櫻庭敦之, 成田龍太, 全眞嬉, 徳山豪,“ジャストインタイムウェブ広告におけるタクソノミ自動生成手法”,LA シンポジウム 2009，2009 年 7 月 22 日，松島

[13]Ryosei Kasai, Jinhee Chun, Matias Korman, Takeshi Tokuyama, “Algorithms for optimal segmentation of regions decomposable into basic shapes”, コンピューテション研究会, IEICE Technical Report,Vol.109, No.108, pp.23-30, 2009 年 6 月 29 日，札幌

６．研究組織

(1)研究代表者

全　眞嬉（CHUN JINHEE）

東北大学・大学院情報科学研究科・准教授

研究者番号：８０４３１５５０

(2)研究分担者

（　　）

研究者番号：

(3)連携研究者

（　　）

研究者番号：