様式 C-19

# 科学研究費補助金研究成果報告書

平成 21 年　4 月　1 日現在

研究種目：若手研究（B）
研究期間：2007 年～2008 年
課題番号：１９７００００３
研究課題名（和文）
数値データに対する階層化ルールと学習理論のデータマイニングへの応用
研究課題名（英文）
Layard structure rule for numeric data and application to data mining
研究代表者
全　眞嬉（JINHEE CHUN）
東北大学・大学院情報科学研究科・助教
研究者番号：８０４３１５５０

研究成果の概要：

本研究の目的はデータマイニングにおける現在の精度限界を打破するための知識抽出モデルの提案，理論研究，システムの構築である．
研究成果として提案した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．研究成果をまとめた論文が計算幾何分野のトップ国際会議ＡＣMSymposium on Computational Geometry2008(SoCG08)で発表した．さらに高く評価され Special Issue of DCG dedicated to SoCG'08 に招待され，ジャーナル論文を提出し 2009 年掲載予定である．

交付額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 2006 年度 | 1,300,000 | 0 | 1,300,000 |
| 2007 年度 | 1,300,000 | 390,000 | 1,690,000 |
| 年度 | | | |
| 年度 | | | |
| 年度 | | | |
| 総　計 | 2,600,000 | 390,000 | 2,990,000 |

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，知識発見，情報システム，計算理論，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は
・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

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．

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において，従来はデータの二値化を行ってから属性間相関関係を求めることが行われていた．しかしながら，この方法では，二値化誤差に起因する制度限界がある．本研究では数値データ集合を幾何学的に扱い，計算幾何学を用いて巨大数値データベースを効率的に処理する新しい最適化アルゴリズムを与え，上記の精度限界を超える精密な手法の提案を行う.さらに相関関係を幾何学的に可視化することにより，ユーザにとって知識発見過程が明示的であり判り易いシステム開発である．

## ３．研究の方法

・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．

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．本研究では数値データ処理問題を計算幾何の問題に置き換えて図形の階層化ルールとして扱う．

従来，計算幾何学での図形近似理論では主にハウスドルフ距離や最大垂直距離（L∞距離）などの最大距離を最小化する基準（最大値最小化基準）をもとに構築されている．しかし，最大値最小化基準では出力の任意性があり，測定エラーに弱いという弱点があり，応用によっては十分な近似が行われないという問題点があった．その点、例えばLp距離等の積分型の大域的距離は誤測定に対して頑健であり、多くの応用において図形の大域的性質をよりよく近似できる．

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．

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．

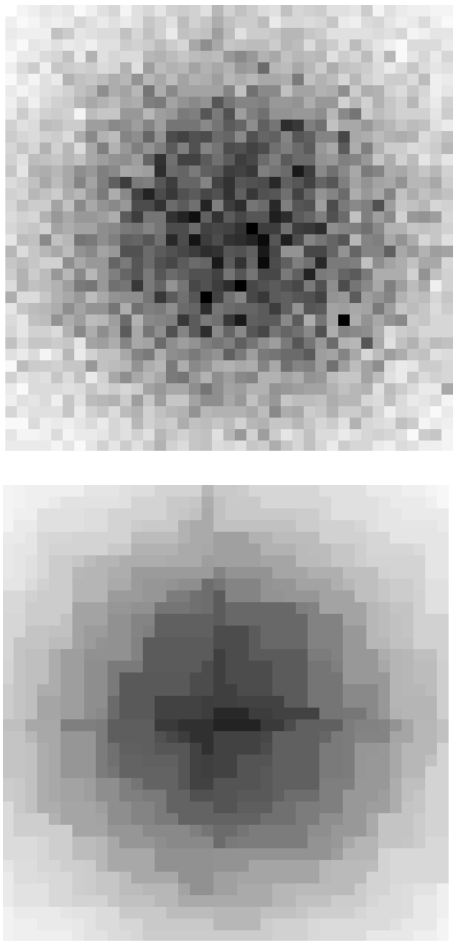

図１．関数として切り出す階層的セグメンテーション（２次元の例，上図は入力，下図は出力）

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を示す．d-次元グリッド上の中心点ｏから各グリッド点ｐへ向かうデジタル直線 dig(op)の集合の数学的に整合的な定義を与える．各デジタル直線は dig(op)は点ｏと点ｐの間のユークリッド線分$¥seg{op}$を近似し、すべてのデジタル直線の集合がユークリッド公理に類似した公理系を満たす． 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．

その研究結果を査読付き国際会議(雑誌論文[1,2,3,4,5])で発表した．

特に，雑誌論文[1]は計算幾何分野のトップレベルの国際会議である SoGC’08(Symposium on Computational Geometry)に採択され発表を行った．さらに高い評価を得て，国際ジャーナル Discrete & Computational Geometry に招待され 2009 年に掲載予定であるである．

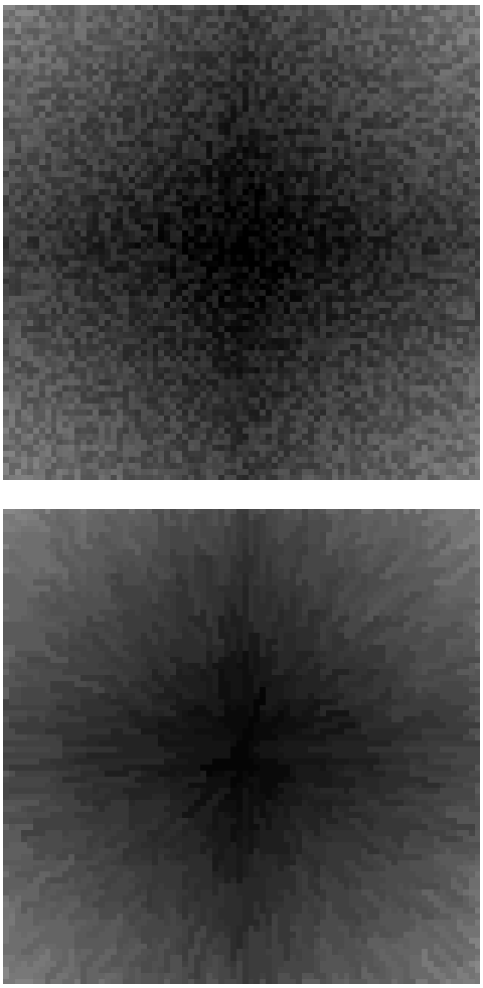

図２．デジタル線分を用いた近似の例（上図は入力，下図は出力）

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５件）

[1]Jinhee Chun, Matias Korman, Martin Nöllenburg and Takeshi Tokuyama, “Consistent digital rays”, Symposium on Computational Geometry (SoCG08),pp.355-364(2008) （査読有）

[2]Jinhee Chun, Matias Korman, Martin Nöllenburg and Takeshi Tokuyama, “Finding the maximum union of closures is NP-hard, even for trees”, The 11th Japan-Korea Joint Workshop on Algorithms and Computation(WAAC08), pp.139-144(2008) （査読有）

[3]Jinhee Chun, Matias Korman, Martin Nöllenburg and Takeshi Tokuyama, ″Digital Star Shapes and Their Applications″, Asian Association for Algorithms and Computation (AAAC'08), (2008) （査読有）

[4]Jinhee Chun, Matias Korman, Martin Nöllenburg and Takeshi Tokuyama, ″Consistent digital rays″, 24th European Workshop on Computational Geometry (EuroCG'08), pp.169-172(2008) （査読有）

[5]Jinhee Chun, Yuji Okada and Takeshi Tokuyama, ”Distance Trisector of Segments and Zone Diagram of Segments in a Plane”, The 4th International Symposium on Voronoi Diagrams in Science and Engineering (ISVD07), pp.66-73(2007) （査読有）

〔学会発表〕(計９件)

[1] 全眞嬉,“デジタル線分とその応用”, ワークショップ離散アルゴリズムの最先端, 東京工業大学, 2009年2月23日（招待講演）

[2]成田龍太, 全眞嬉, 徳山豪, “動画に対するコメントを利用した自動Web検索システム”, FIT2008 報科学技術フォーラム, 慶応大学, D-022, 2008年9月3日

[3]Jinhee Chun, Matias Korman, Martin Nöllenburg and Takeshi Tokuyama, “Consistent Digital Rays”, The 2 nd International Symposium on Information Electronics Systems(CERIES-GCOE08), 仙台, 2008年7月15日

[4]Jinhee Chun, Matias Korman, Martin Noellenburg, Takeshi Tokuyama, ″Consistent digital rays″, Symposium on Computational Geometry 2008 (SoCG'08), Maryland USA, 2008年6月 11日

[5]Jinhee Chun, Matias Korman, Martin Noellenburg, Takeshi Tokuyama, ″Digital Star Shapes and Their Applications″, Asian Association for Algorithms and Computation 2008 (AAAC'08), Hongkong, 2008年4月26日

[6]Jinhee Chun, Matias Korman, Martin Noellenburg, Takeshi Tokuyama, ″Consistent digital rays″, 24th European Workshop on Computational Geometry 2008 (EuroCG'08), Nancy France,2008年3月19日

[7]Jinhee Chun, Matias Korman, Martin Nöllenburg and Takeshi Tokuyama, ″Consistent digital rays″,電子情報通信学会コンピュテーション研究会, IBM基礎研究所, 2008年3月10日

[8]Jinhee Chun, Matias Korman, Martin Nöllenburg and Takeshi Tokuyama, ″Digital Star Shapes and Their Applications″, 情報処理学会 114 回アルゴリズム研究会, 豊橋, 2007年9月21日

[9]Jinhee Chun, Yuji Okada, Takeshi Tokuyama, ″Distance Trisector of Segments and Zone Diagram of Segments in a Plane″, International Symposium on Voronoi Diagrams in Science and Engineering(ISVD 2007), Wales UK, 2007年7月11日

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・助教
研究者番号：８０４３１５５０