D-39 不均一サイズセルを用いた階層的クラスタリングの高速化

Speed-up of the Hierarchical Clustering Based on Cell Structure with Irregular Size

中村 朋健† 上土井 陽子‡ 吉田 典可‡

Tomotake NAKAMURA Yoko KAMIDOI Noriyoshi YOSHIDA

1 はじめに

与えられたデータにおいて類似したデータ要素を集めるプロセスのことをクラスタリングという．本研究では，疎な領域によって分けられる密な領域の要素の集合を集めるため，セルを密度情報を基に階層的に構築し，そのセル構造を用いたクラスタリングアルゴリズムを提案し，実験的に評価する．

2 STING

格子を用いる階層的手法である STING(STatistical INformation Grid-based method)[2] は，空間の領域を長方形のセルに分割する多重解像クラスタリング手法である．STING の階層構造を図 1に示し，クラスタリングアルゴリズムを以下に示す．

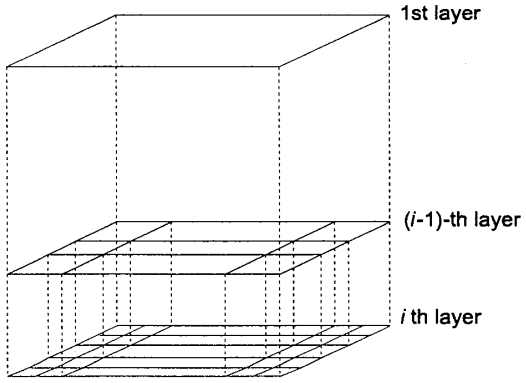

図 1: STING の階層構造

[STING アルゴリズム]

(1) 最下位レベルを決める．
(2) この層における各セルに対して，このセルが密であるというキュエリの妥当な信頼区間（または，大まかな確率）を計算する．
(3) (2) で定めた区間に各セルが入っているか，または，入っていないか分類する．
(4) この層が最下位レベルであるならば，(6) へ進む．そうでなければ，(5) へ進む．
(5) 1つレベルを下げる．下位レベル層において密なセルを形成するため (2) へ戻る．
(6) キュエリを満たす場合は，(8) へ進む．そうでなければ，(7) へ進む．
(7) 密なセルまで下がり処理を実行する．キュエリの要求を満たした結果を返し，(9) へ進む．
(8) 密なセルの領域を見つけ，密な領域に含まれるデータ要素の集合をクラスタとし，終了する．

†広島市立大学大学院 情報科学研究科 (Graduate School of Information Sciences, Hiroshima City University)

‡広島市立大学 情報科学部 (Faculty of Information Sciences, Hiroshima City University)

3 提案手法 1

STING はセル構造の最下位レベルに達するまで終了しない．本研究ではすべてのセルにおける最下位レベルへの降下を避けるため，ある一定の密度以上であれば最下位レベルまで下げずに密なセルと判断できるアルゴリズムを提案する．この手法ではセル分割数が少なく，セル分割に要する時間が少ない．提案手法 1 は以下の 2 つのフェーズでクラスタリングを行う．

3.1 フェーズ 1

分割を行い，結合に必要な情報を収集する．入力パラメータ PMin, PMax, MinC, PMed を用いた分割手法の概要を図 2に，アルゴリズムを以下に示す．

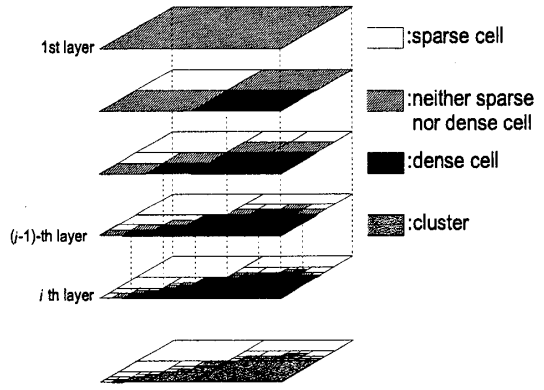

図 2: 階層構造

[提案手法 1 のフェーズ 1 アルゴリズム]

(1) 入力データが PMin 以上かつ，PMax より小さければ (2) へ進む．PMax 以上であれば，入力データ全体が 1 つの密なセルであると判断し，セルの周りの点を保持しておく．PMin より小さければ，入力データ全体が疎なセルと判断する．
(2) 疎・密が判断されていないすべてのセルを 4 分割する．分割されるセルがなければ終了する．
(3) セルの大きさが MinC より小さければ (5) へ進む．
(4) 分割後，新たに構成された各セルに対して密度が PMax 以上であれば，密なセルと判断し，セルの周りの点を保持しておく．密度が PMin より小さければ，疎なセルと判断する．(2) へ戻る．
(5) 各セルに対しセルの密度が PMed 以上であれば，密なセルと判断し，セルの周りの点を保持しておく．PMed より小さければ疎なセルと判断し，終了する．

3.2 フェーズ 2

密なセルにおいて，共通な点を持つセル同士を結合し，セルの連結成分 1 つに属するすべてのデータ要素の集合を一つのクラスタとする．概要を図 3に示す．

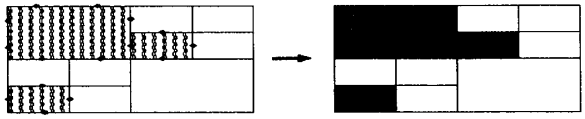

図 3: 提案手法 1 のフェーズ 2 の概要

3.3 シミュレーション実験

提案手法1をC言語を用いてSUN Ultra60 Model1450上に実現し，シミュレーション実験を行った．入力データとして2次元描画データを用いた．

図4に示すデータ要素数66,129個の入力データを用いて，最下位レベルまで分割するときと分割しないときのフェーズ1に要する時間の比較を行う．最下位レベルまでセルを同一な大きさに分割したときの解を均一分割解とし，明らかに密または疎と判断されるセルはそれ以上分割しないことで，セルを不均一な大きさに分割したときの解を不均一分割解とした．両解は計算時間とクラスタリング結果を総合的に考え主観的に一番良いと判断したときの解とする．

入力データに対する均一分割解の出力結果を図5に，不均一分割解の出力結果を図6に示す．両出力結果において，同じクラスタに属する点は同色，同型で表し，ノイズはノイズのみを抽出し同色，同型で表している．

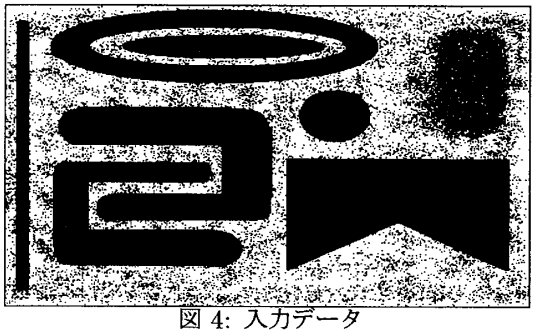

図 4: 入力データ

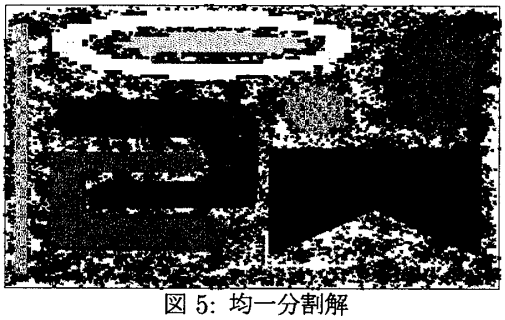

図 5: 均一分割解

図 6: 不均一分割解

両解の計算時間を表1に示す．

表 1: 入力データに対する計算時間 [s]

	フェーズ1	フェーズ2	合計
均一分割解	0.59	164.16	164.75
不均一分割解	0.48	34.48	34.96

実験結果からセル不均一分割によりフェーズ1で短縮された時間はわずかであることが分かる．フェーズ2に要する時間はフェーズ1に要する時間に比べ非常に大きかった．そのため，フェーズ1での時間の差はクラスタリング時間にあまり影響しなかった．フェーズ2に要する時間が1秒程度になれば，フェーズ1での時間の差が計算時間の改良に寄与すると考えられる．

4 提案手法2

提案手法2では，提案手法1の問題点であるフェーズ2の計算時間を短縮することを考える．提案手法1のフェーズ2を改良するためには，大きさの不均一なセルの隣接関係が容易に分かるリストを作成し，それを基に結合を行う．

提案手法2は以下の2つのフェーズでクラスタリングを行う．

4.1 フェーズ1

提案手法1のフェーズ1とほぼ同じであるが，ここでは密なセルの周りの点を保持せず，結合手続きのためにセルの隣接関係が分かるリスト表を各階層毎に更新しながら作成する．作成するリスト表は各セルに対し隣接するセルが分かるリストの集合である．ただし，隣接するセルがそのセルよりも大きさが小さいときはリスト表に出力しない．具体的なリスト表の作成例を図7に示す．ここで，各セル内の3桁で表された数字をセル番号と呼び，左上が1，右上が2，左下が3，そして右下が4とし，上位桁からつける．

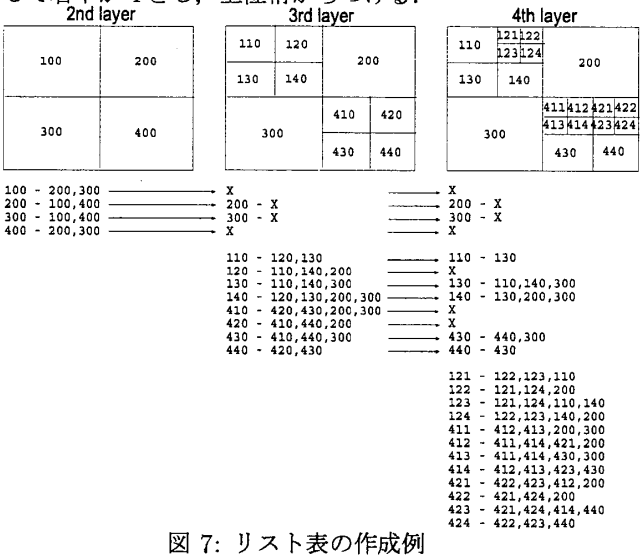

図 7: リスト表の作成例

4.2 フェーズ2

フェーズ1で作成したリストから疎なセルを取り除き，密なセルのみの隣接関係が分かるリスト表に書き換える．あるリストと他のリストに同一のセル番号が存在するならば，それらのリストを結合させる．一方のリストに存在するすべてのセル番号が他のリストすべてに存在しなかったら，それだけがクラスタを形成するものと考える．

5 考察

すべてのセルを最下位レベルまで分割しないことにより，フェーズ1での分割手続き，および，フェーズ2での結合手続きの計算時間は短縮されると予測できる．

提案手法1と2の計算複雑さを比較すると，セルの周りの点を保持する場合と，リスト表を作成する場合の計算複雑さは共に $O(N + cn)$ である．ここで，N は分割されたセル数，n はデータ要素数，そして c は層数とする．また，提案手法1に比べて提案手法2の結合に要する計算複雑さは，提案手法1では $O((MN)^2)$ に対し提案手法2では $O(N)$ で行えると予測でき，大幅な計算時間の削減を期待できる．ここで，M は一つのセルに対する周りの点の最大数とする．

参考文献

[1] J. Han and M. Kamber: "Data Mining: Concepts and Techniques," Academic Press, pp. 335-376, 2001.

[2] W. Wang, J. Yang and R. Muntz: "STING:A statistical information grid approach to spatial data mining," Technical Report, Department of Computer Science University of California, pp.1-15, 1997. (in Proc. 1997 Int. Conf. Very Large Database, pp.186-195,1997)