クラスタリング問題への適用に向けた
自己組織化マップの学習法改善に関する研究

2009年1月

加藤　聡

内容梗概

本論文は，Kohonenによって提案された自己組織化マップ（Self-Organizing Map : SOM）をクラスタリングに適用する問題に対し，そのクラスタリング性能の改善を目的とした2段階SOM（2-stage SOM），およびその改良手法である拡張2段階SOMを提案し，人工的な評価用データセットや，UCI Machine Learning (UCI ML）データベースから引用した実データを用いて，これらのデータセットに対する提案手法におけるクラスタリング性能ならびに，従来のクラスタリング手法との比較検討などに関して行った研究成果をまとめたものである．

第1章は，序論として，本研究に関連する従来の研究概要について述べ，本研究の目的ならびに，本研究を行うに至った背景及び，各章の概要を述べている．

第2章では，クラスタリング問題の概要について，種々のクラスタリング手法の分類を行い，特に，代表的なクラスタリングアルゴリズムであるk-means法や，階層的クラスタリング手法である最短距離法，最長距離法，ウォード法，また，グラフ理論を用いた手法について，それらの具体的なアルゴリズムと問題点について説明し，本論文で対象とする，SOMを用いたクラスタリングの位置付けを述べている．

第3章では，SOMを用いたクラスタリングの具体的手法を示し，その問題点を指摘した上で，提案手法である2段階SOMについて述べている．SOMの学習アルゴリズムでは，ある競合層セルが受けたコードベクトルの更新が，そのセルに隣接したセルにも影響するという，近傍学習と呼ばれる性質があり，これによって学習後のSOMの特徴マップにおけ

る位相保持写像が可能となる．位相保持写像は，SOMを用いたクラスタリングに対して重要な特性であるが，近傍学習によって，学習後にいずれのクラスタにも属さない不活性セルが発生し，これらの不活性セルが，学習後のクラスタ抽出に悪影響を及ぼす．そこで本章では，SOMの基本学習アルゴリズム（BSOM）と，近傍学習にしきい値作用を導入した学習アルゴリズム（THSOM）とを段階的に適用する手法を提案し，人工的に作成した入力データに対して，提案法では不活性セルの発生が抑制されていることを確認している．

第4章では，人工的に作成した評価用データセットに対する，2段階SOMのクラスタリング性能および問題点について論じている．本章において，提案手法である2段階SOMは，各クラスタにおけるデータの密度が一定の場合には，従来のSOMやk-means法などと比較して，クラスタリング時における誤分類率の改善が見られることを確認している．一方で，各クラスタにおけるデータの密度が一定でない場合，2段階SOMにおけるTHSOM過程におけるしきい値設定が困難となり，期待された通りの誤分類率の改善が得られないことを示している．

第5章では，前章において示された2段階SOMの問題点について，その改良手法である拡張2段階SOMの提案と，人工データおよびUCI MLデータベースから引用した種々の実データを用いた性能評価について述べている．拡張2段階SOMでは，そのTHSOM過程において，しきい値の尺度の算出方法を変更することにより，従来型2段階SOMでは困難であった，密度の異なるデータの正確なクラスタリングを実現している．このことを，人工データを用いた評価実験によって示し，また，UCI MLのIris，Breast-cancer wisconsin，Vowel，Thyroid glandデータセットによるクラスタリング実験を通して，拡張2段階SOMの有効性を確認している．

最後の第6章において全体の総括を行っている．

論 文 目 次

図目次

表目次

主要記号

第2章	
記号	定義・意味
D_{mahal}	マハラノビス距離
D_{euc}	ユークリッド距離
$\mathbf{x}$	入力データ
μ	k-means 法における代表ベクトル
$\mathbf{w}_i$	競合層のセル i が持つコードベクトル
T	SOM の最大学習回数
t	SOM の学習回数 $(0 < t \leq T)$
p_i	競合層上における，セル i から勝者セルまでの距離
$\Phi^{\mathrm{BSOM}}(p_i)$	BSOM の近傍関数
α_{ini}	学習率 α の初期値
σ_{ini}	近傍関数 Φ^{BSOM} のパラメータ σ の初期値
$\alpha(t)$	学習回数 t における学習率 α の値
$\sigma(t)$	学習回数 t における近傍関数 Φ^{BSOM} のパラメータ σ の値
第3章	
記号	定義・意味
Th	THSOM の しきい値
$n_{i,j}(t)$	学習回数 t までに，セル i とセル j が持つコードベクトルのユークリッド距離が，しきい値 Th を超えた回数
$\Phi^{\mathrm{THSOM}}(t)$	THSOM の近傍関数
η	2段階 SOM の THSOM 過程における学習率
V_i	SOM の学習過程において，セル i が「勝者セル」となった回数
D_i	セル i におけるコードベクトルの密度
V_{min}，V_{max}	競合層のすべてのセルにおける V_i の最小値，最大値
D_{min}，D_{max}	競合層のすべてのセルにおける D_i の最小値，最大値
$V_{\mathrm{N\_}i}$	V_i を，$[V_{i\_\mathrm{min}}, V_{i\_\mathrm{max}}]$ が [0.0 , 1.0] となるように正規化したもの
$D_{\mathrm{N\_}i}$	D_i を，$[D_{i\_\mathrm{min}}, D_{i\_\mathrm{max}}]$ が [0.0 , 1.0] となるように正規化したもの
L_i	セル i の活性度
L_{TH}	2段階 SOM の THSOM 過程において不活性セルを判定するための，活性度のしきい値

T_{TS1}	2段階SOMのTHSOM過程（前半部分）の学習回数
T_{TS2}	2段階SOMのTHSOM過程の学習回数
$\Phi^{\mathrm{2stg}}(t)$	2段階SOMのTHSOM過程における近傍関数
dW_i	競合層におけるセルiとセル$(i+1)$が持つコードベクトル同士のユークリッド距離
dW_{i_min}, dW_{i_max}	競合層のすべてのセルにおけるdW_iの最小値，最大値
dW'_i	dW_iを，$[dW_{i_\mathrm{min}}, dW_{i_\mathrm{max}}]$が[0.0,1.0]となるように正規化したもの
dW'_{true}	データ密度ヒストグラムにおいて，クラスタ境界に相当するピークの最小値
dW'_{false}	データ密度ヒストグラムにおいて，クラスタ境界に相当しないピークの最大値

第4章

記号	定義・意味
μ	クラスタリング対象データにおける個々の既知クラスタの平均ベクトル
ν	クラスタリングによって抽出された個々のクラスタの平均ベクトル
D_{match}	既知クラスタと，クラスタリングによって抽出されたクラスタとの平均ベクトル同士の距離の総和
P_{Err}	誤分類率(%)
dW'_{TH}	SOMの学習後にクラスタ抽出を行う際の，dW'_iのしきい値
L_i	2段階SOMにおけるセルiの活性度
L_{TH}	2段階SOMのTHSOM過程において不活性セルを判定するための，活性度のしきい値

第5章

記号	定義・意味
dV_i	セルiにおける学習時の勝利回数の，隣接セルとの変化量
dD_i	セルiにおけるコードベクトル密度の，隣接セルとの変化量
L_i	2段階SOMにおけるセルiの活性度
L_{TH}	2段階SOMのTHSOM過程において不活性セルを判定するための，活性度のしきい値
L'	拡張2段階SOMにおけるセルiの不活性度
L'_{TH}	拡張2段階SOMのTHSOM過程において不活性セルを判定するための，不活性度のしきい値
P_{Err}	誤分類率(%)
dW'_{TH}	SOMの学習後にクラスタ抽出を行う際の，dW'_iのしきい値

第1章

緒言

1.1 研究の背景

電子計算機の誕生から半世紀以上が経過した昨今，計算機システムにおけるCPUの処理性能および主記憶容量は現在もムーアの法則にしたがって向上を続け，それにともなって外部記憶ストレージの容量も増加の一途をたどっている．これに，近年のインターネットの普及による通信インフラの高速化・大容量化とが相まって，地球規模の情報の交換，あるいは蓄積が容易に可能となってきた．計算機システム内に大量に蓄積された情報は多種多様であり，これらのデータに内在する規則性や構造を見出すための，データ解析手法の研究が益々重要視されている．中でも，何らかの基準に基づく類似度にしたがって，対象データをいくつかのまとまり（クラスタ）にグループ化するクラスタリング[1][2]は，重要なデータ解析手法の一つとして，統計，パターン認識[3]，データマイニング[4]などの分野で盛んに研究されており，応用事例も非常に多い[5]．

クラスタリング手法は階層的手法と非階層的手法に大別される．非階層的クラスタリング手法の一つであるk-means法は，k個のクラスタ中心を，適当な初期状態から繰り返し計算によって求めるという簡潔なアルゴリズムであるため古くから研究され[6]，現在に至るも広く用いられている[7]．しかしながら，クラスタ数kが既知であることを前提とするため，kが未知の場合に適用するには，クラスタの併合や分割などの手法を取り入れる必

要があることや [8]，繰り返し計算によってクラスタ中心を収束させるため，結果が初期状態に大きく依存するといった問題点がある．そのため実際のデータ解析に用いる場合には，k の値や初期状態の選択などを，経験に基づく試行錯誤によって行わなければならない．さらに，k-means 法は k 個のクラスタ中心と，各クラスタ中心に割り当てられるデータとのユークリッド距離の総和を最小化する手法であるため，多次元空間内における超球状のクラスタ抽出を暗黙的な前提としている．したがって，この前提に当てはまらないクラスタの抽出は原理的に困難である．一方，階層的クラスタリング手法は，個々のデータをそれぞれ個別のクラスタとみなすことから出発し，あらかじめ定義された非類似度（クラスタ間の距離）尺度に基づいて，非類似度の小さいものから徐々にクラスタを併合して行く手法である．階層的クラスタリング手法には，非類似度の定義の違いによって，最短距離法 [9]，最長距離法 [10]，ウォード法 [11] など，いくつかのバリエーションがあるため，同一のデータに対しても適用する手法によってクラスタの併合過程が異なる．これによって，データの性質と適用する手法の組合わせによっては，k-means 法で抽出できないようなクラスタの抽出が可能な場合もある．しかしながら，あるデータに対してどの手法を適用すべきかについては，k-means 法の場合と同様に経験に基づく試行錯誤が必要である．さらに，階層的クラスタリングの各手法は，クラスタ同士の非類似度をまとめた非類似度行列をクラスタ併合の度に更新する必要がある．非類似度行列の大きさは，基本的にデータ数の二乗に比例するため，階層的クラスタリング手法は非類似度行列の更新にかかる計算量の観点から，大規模データへの適用が困難である [1]．

これらに対して，Kohonen の自己組織化マップ（Self-Organizing Map: SOM）[12] をクラスタリング問題に適用する研究が近年進められている．SOM は教師なし学習を行うニューラルネットワークの一種であり，SOM のネットワーク層（競合層と呼ばれる）ではセルと呼ばれるニューロンが1次元あるいは2次元格子状に配列している．SOM の学習は，入力データを競合層に繰り返し提示して，個々のセルにおける結合加重ベクトル（コードベクトルと呼ばれる）を更新していくことで行われる．学習後のコードベクトルの分布は入力

データの分布を反映したものになっており，さらに，競合層のセル配列の中で互いに隣接するセル同士は類似のコードベクトルを持つように学習が収束する性質がある．これは位相 (Topology) 保持写像と呼ばれ，SOM の大きな特長である．この性質を利用して，SOM は学習後のコードベクトルに入力データを対応付けることによって，多次元データの分布を1次元あるいは2次元の特徴マップとして可視化することができる [13] [14]．このことは，学習後のコードベクトルの状態に注目することで，データの存在が疎な部分，すなわちクラスタの境界を検出可能であることを示唆しており，寺島ら [15] は1次元SOM を用いたクラスタリング手法を，田中ら [16] は2次元SOM を用いたクラスタリング手法をそれぞれ提案している．特に寺島らは，k-means 法と比較した場合の，SOM によるクラスタリングの有効性を定性的に示している．クラスタリング手法としてSOM を捉えた場合，k-means 法などと比較して，初期状態への依存性が少なく，安定したクラスタリング結果を得られることが特徴である．

SOM をクラスタリング問題に適用する場合，SOM の学習性能がクラスタリング結果に影響を及ぼす．Kohonen が提案した基本学習アルゴリズムによる SOM（Basic-SOM: BSOM）[12] は，複数のクラスタからなる入力データを学習する場合，クラスタの境界部分を正しく学習できず，いずれのクラスタにも属さないコードベクトルを持った「不活性セル」が発生してしまう．この問題点を解決するために，青木らは，しきい値作用を導入した学習アルゴリズムによる SOM（Threshold-SOM: THSOM）[17] を提案している．しかしながら，THSOM は入力データの位相を保存することが困難になるという欠点があり，クラスタリング問題への適用は難しい．また，Uchino らは，学習時のセルの勝利回数に着目し，勝利回数の少ないセルを競合層から排除することで，不活性セルの発生を事実上抑制する手法を提案している [18]．この手法では，不活性セルの排除によって，対象データを表現するためのコードベクトル数が削減されてしまうため，SOM が持つベクトル量子化器としての特長 [19] が阻害されるおそれがある．

1.2 研究の目的と各章の概要

SOMのクラスタリング問題への適用は，前節で述べた通り，いくつかの問題点がある．しかしながら，SOMを用いれば，大規模なデータを少数のコードベクトルで近似でき，また，個々のクラスタが任意の分布形状を持っているような場合でも，その分布形状にコードベクトルの分布を近似させることができるなど，クラスタリング問題への適用を考えたときのSOMの有効性は大きい．したがって，本研究では，SOMの利点を残した上で，クラスタリング問題への適用に向けたSOMの学習アルゴリズムの改良を目的とする．

第2章では，非階層的クラスタリング手法であるk-means法や，階層的クラスタリング手法である最短距離法，最長距離法，ウォード法など，従来の代表的なクラスタリング手法との対比を通して，SOMを用いたクラスタリングの利点について述べる．k-means法は超球状のクラスタを暗に仮定しているため，任意形状のクラスタ抽出が困難であり，階層的クラスタリング手法は，非類似度行列が肥大化するために大規模データへの適用が困難である．これに対して，SOMは1つのクラスタを複数のコードベクトルによって近似することができるために，クラスタ形状に関する制約が緩和される．また，SOMが元来持っているベクトル量子化器としての性質[19]は，少数のコードベクトルで入力データの要約を得ることによって，大規模データへの適用が比較的容易に行えることを示唆している．本章ではまず，従来手法の具体的なアルゴリズムや問題点について概説した上で，SOMの基本学習アルゴリズムとその特長を示し，クラスタリング問題にSOMを適用する場合の有効性を論じる．

第3章では，クラスタリング問題に適用する際のSOMの問題点を述べ，その解決を目的としたSOMの学習法の改良を行う．SOMの基本学習アルゴリズムであるBSOMは，学習後にクラスタ間にコードベクトルが残留する性質がある．これらのコードベクトルを持ったセルは不活性セルと呼ばれ，学習後のクラスタ抽出時に障害となる．学習時に不活性セルが生じる問題は，青木らが提案したTHSOMによって解消されることが示されている[17]．

しかしながら，THSOMでは入力データの位相保持写像が著しく阻害されるという問題があり，クラスタリング問題への適用は非常に困難である．以上のことから，本章では，入力データに対してBSOMとTHSOMを段階的に適用するという，2段階の学習アルゴリズムを持つ「2段階SOM」を提案する[20] [21]．2段階SOMでは，BSOMの適用後に得られたコードベクトルをTHSOMの初期状態として用いることにより，入力データの位相保持写像を獲得し，かつ不活性セルの発生を抑制することを目的としている．この2段階SOMに対して，人工的に作成した入力データ用いた学習実験を行い，不活性セルの発生が抑制されていることを示す．

第4章では，人工的に作成した評価用データセットに対する，2段階SOMのクラスタリング性能の評価について述べる．クラスタリングの評価法には様々なものが考えられている[22]が，本研究ではラベル付きのデータを評価用データセットとして用いるため，ラベル情報を使わずにクラスタリングを行った結果と，ラベル情報に基づいてグループ化された結果とのマッチングをとり，その不一致の度合いを示す誤分類率によってクラスタリング結果を評価する．本章では，データに含まれるクラスタ数や，個々のクラスタにおけるデータ密度のばらつき，およびクラスタ形状などが異なるデータセットを作成し，k-means法，BSOM，2段階SOMおよび階層的クラスタリング手法（最短距離法，最長距離法，ウォード法）それぞれを適用した場合の誤分類率を比較することによって，提案手法である2段階SOMのクラスタリング性能を評価した[23]．

第5章では，前章において示された2段階SOMの問題点について，その改良手法である拡張2段階SOMの提案[24]と，人工データおよびUCI MLデータベースから引用した実データを用いた性能評価[25][26]について述べる．従来型の2段階SOMでは，クラスタ毎のデータ密度にばらつきがあるような場合に，直観とは異なるクラスタ抽出結果が得られるという，k-means法と同様の問題点がある．本章では，従来型2段階SOMのTHSOM適用過程において，各セルの活性度の算出方法を改善することによって問題点の解決を行っており，これによって2段階SOMが適用できるクラスタリング対象データの範囲の拡張を

図る．この拡張2段階SOMに対して，人工データを用いた予備実験によって学習アルゴリズム改善の有効性を示す．さらに，実データとしてUCI MLデータベースのデータセット[27]を用いたクラスタリング実験を行い，従来型2段階SOMでは正確なクラスタリングが困難であったデータセットに対して誤分類率の改善が見られることを示し，拡張2段階SOMの有効性を確認する．

最後の第6章において，本論文の各章で得られた知見をまとめ，今後の課題および展望について述べる．

第2章

クラスタリングの代表的手法とその問題点

クラスタリング（またはクラスタ分析）とは，分析対象となるデータを構成するサンプルの集合に対して，何らかの尺度に基づいて定義された類似度によって，互いに類似したサンプル同士をグループ化し，その分類結果によって，対象データに内在する特徴を見出す手法である．クラスタリングの対象となるデータは多種多様であり，分析の目的も様々であることから，クラスタリング手法にも，データの種類や分析の目的に応じて多くのものが存在している．本章では，従来の代表的なクラスタリング手法を，その目的や，クラスタを抽出する際のアプローチの違いに基づいて分類し，各手法の特徴と問題点について述べる．その上で，自己組織化マップ（SOM）について概説し，従来のクラスタリング手法における位置付けを述べる

2.1 クラスタリング手法の分類

前述した通り，クラスタリング手法には対象データの性質や分析目的によって様々なものが存在している．クラスタリング手法の分類については，古くは Williams と Lance によるもの [1] や，近年では Jain によるサーベイ [2] などが詳しい．これらに共通するのは，クラスタリング手法の分類において，「クラスタというものの考え方」，「対象データからのクラスタの見出し方」および「クラスタリングの目的」，などの視点を用いていることである．これらの視点すべてを考慮したクラスタリングの分類を図示することは難しいが，

WilliamsとLance，およびJainによる分類をまとめれば，クラスタリングの手法は図2.1に示すように分類される．この分類は，現在提案されている様々なクラスタリング手法を厳密に規定するものではないが，クラスタリング手法の体系的に整理する上では有用であると思われる．ここではまず，図2.1において示した様々な視点の意味について説明する．

排他的／非排他的

排他的あるいは非排他的とは，クラスタというものをどう考えるかについての，2つの異なる視点である．排他的クラスタの視点では，個々のサンプルは，ただ1つのクラスタに属するものとする．これに対して，サンプルが複数のクラスタに属することを許容するのが，非排他的クラスタの視点である．

外的基準なし／外的基準あり

外的規準とは，観測によって得られた情報とは別に，外部から与えられる情報であり，具体的には，クラスタリング対象データの各サンプルが属するべきクラスタに関する情報である．外的規準が与えられない場合，個々のサンプルが持つ属性（観測値）だけを用いて，サンプル同士の類似度などに基づいたクラスタリングが行なわれる．通常，単にクラスタリングと言えば外的規準が与えられない場合を指す．外的規準が与えられる場合の具体例としては，判別分析や最近傍識別，あるいはニューラルネットワークなどが挙げられるが，これらはクラスタリング問題ではなく，判別問題あるいはパターン認識問題として扱われることが多い．

階層的／非階層的

階層的クラスタリングとは，1つのクラスタをいくつかの部分クラスタに分割したり，あるいはその逆を行うことで，個々のクラスタ類似性に関する階層的な構造を求める手法である．これに対して，非階層的クラスタリングでは，クラスタ内では互いに類似し，逆にクラスタ間では類似していないようにサンプルを分類することを目的としている．

凝集的／分割的

階層的クラスタリング手法において，凝集的手法とは，個々のサンプルをそれぞれ個別のクラスタとみなすことから始め，最終的に対象データ全体が一つのクラスタとなるようにクラスタリングが進んで行く．これに対して，分割的手法とは，対象データの全サンプルを一つのクラスタとみなすことから始め，そこから徐々にクラスタの細分化を行っていく．ある手法が凝集的であるか分割的であるかは，単にアルゴリズムの実装上の問題であり，これらの違いが，階層的クラスタリング手法としての本質的な差異をもたらすことはない．

本論文が研究の対象とするクラスタリング手法は，基本的には，クラスタというものを排他的な視点で捉え，外的規準が与えられない非階層的な手法に位置付けられる．また，階層的クラスタリング手法についても，クラスタの階層構造において，個々のレベルでの具体的なクラスタ分割を求めることができ，それらのクラスタ分割結果は非階層的な手法におけるクラスタ分割結果と比較可能である．以上のことから，次節以降では，非階層的および階層的クラスタリングにおける，いくつかの具体的な手法について，それぞれの特徴および問題点なども交えて詳しく説明する．

2.2 非階層的クラスタリング法

2.2.1 k-means 法のアルゴリズム

k-means 法 [6] は，McQueen によって提案された非階層的クラスタリングの一手法であり，アルゴリズムが単純であることから，非階層的クラスタリングの代表的手法として古くから用いられている．

いま，クラスタリング対象となるデータの集合を X とし，個々のデータを $\mathbf{x}$ とすると，k-means 法のアルゴリズムは以下のように記述される．

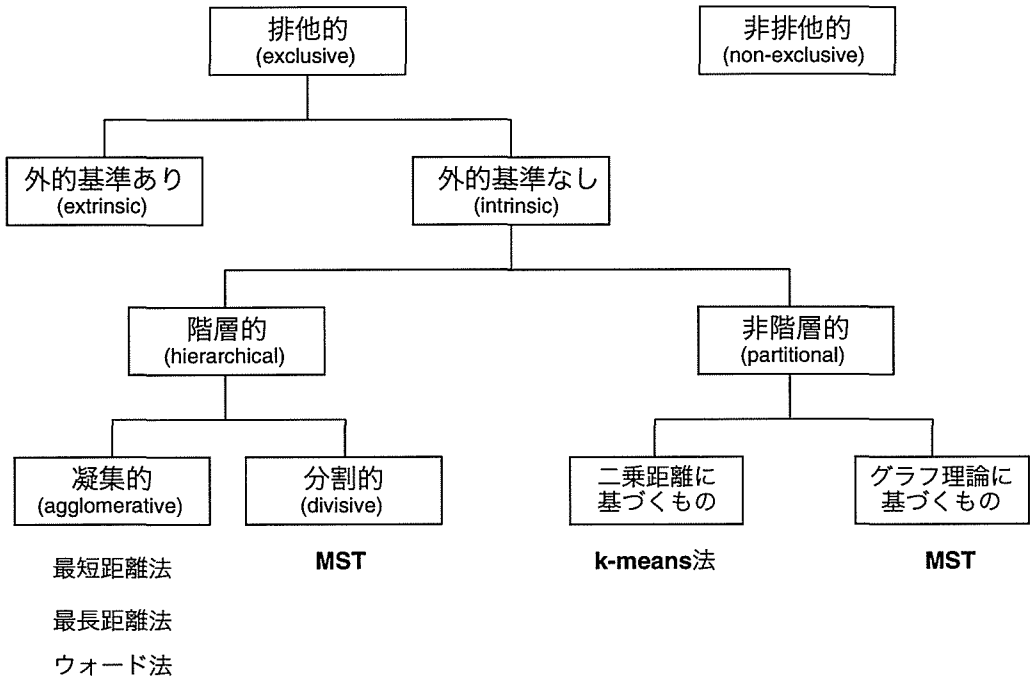


図 2.1 クラスタリング手法の分類

k-means 法のアルゴリズム

STEP1: k 個の初期代表点 $(\mu_1, \mu_2, ..., \mu_k)$ を生成する．具体的な生成方法としては，対象データが分布する範囲における一様乱数による方法や，対象データの中から k 個をランダムサンプリングする方法などがある．

STEP2: $\forall \mathbf{x} \in X$ を，$\min_i D(\mathbf{x}, \mu_i)$ となる代表点に割り当てる．ここで，$D(\cdot)$ は対象データと代表点との距離であり，ユークリッド距離が一般的に用いられる．この時点で，各代表点に対応付けられるクラスタが一応は得られたことになる．

STEP3: 得られたクラスタの中心を，各クラスタの新たな代表点とする．ここで,「クラスタ中心」は各クラスタの平均ベクトルとする．

STEP4: 代表点の更新量があらかじめ定められた値以下ならば，代表点の移動が収束したとみなしてクラスタリングを終了する．そうでない場合は，STEP2 に戻る．

2.2.2 クラスタリング例と k-means 法の問題点

k-means 法は，簡潔な手法であるが故に実際の応用において問題が生じる場合も多い．k-means 法の問題点は以下のようにまとめることができる．

クラスタ数 k が既知である

クラスタリング問題においては，そもそも対象データがいくつのクラスタから構成されているのかが分からない場合が多い．そのため，k-means 法は，対象データが仮に k 個のクラスタで構成されたとした場合に，どのようなクラスタ分割が得られるのかを知る手法であると解釈することもできる．

クラスタリング結果が初期状態に依存する

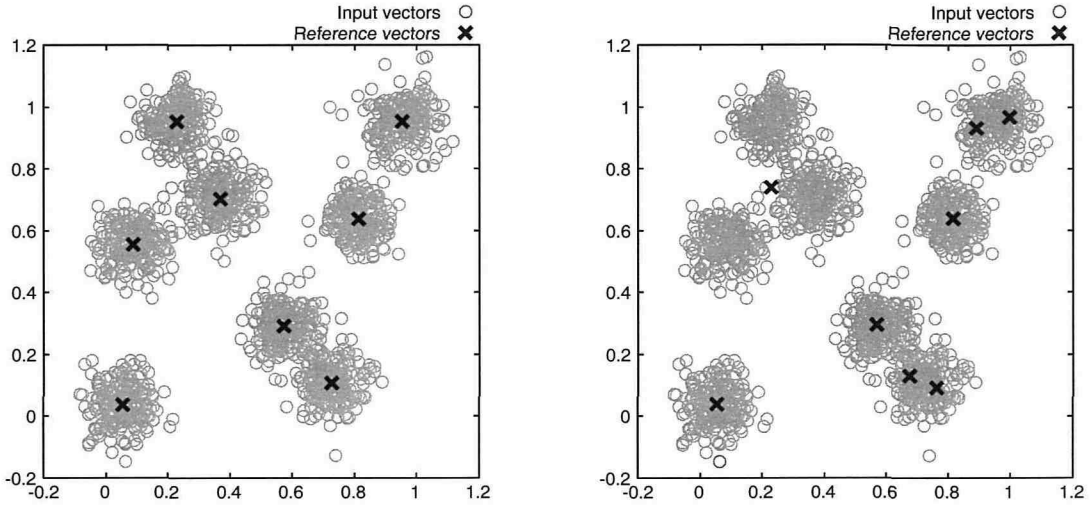
初期代表点を繰り返し計算によって徐々に収束させるため，クラスタ分割の結果が初期代表点の位置に大きく影響される．図 2.2(a) および 2.2(b) は，同じ対象データ（8 個の 2 次元正規分布で構成される）をそれぞれ異なる初期代表点によってクラスタリングした結果である．このように，k-means 法では，たとえ k の値を適切に設定できていたとしても，初期代表点の選び方によっては必ずしも望ましいクラスタリング結果が得られないことが分かる．

二乗距離に基づく

各代表点 μ_i への対象データ $\mathbf{x}$ の割り当てに際しては，二乗距離であるユークリッド距離

$$D_{\mathrm{euc}}(\mathbf{x}, \mu_i) = \sqrt{(\mathbf{x} - \mu_i)^{\mathrm{T}}(\mathbf{x} - \mu_i)} \tag{2.1}$$

が一般的に用いられる．このことは，k-means 法が，クラスタの形状が超球であること，すなわち対象データベクトルの各次元が等しい分散を持つことを暗に仮定しており，さらに，クラスタ同士の分散の違いなども考慮していないことを意味している．図 2.3 は，分散の



(a) クラスタリング成功例　　(b) クラスタリング失敗例

図 2.2　k-means 法における初期代表点によるクラスタ分割結果の違い

異なる 3 つのクラスタからなるデータを k-means 法によってクラスタリングした際の結果を示している．図 2.3 における参照ベクトルの位置から，本来 1 つのクラスタとして抽出されるべきものが 2 つのクラスタに分断されており（図 2.3 の左側），逆に 2 つのクラスタとすべきものが 1 つのクラスタとして抽出されている（図 2.3 の右側）ことが分かる．

この問題に対処するため，データベクトル各次元の分散を考慮したマハラノビス距離

$$D_{\mathrm{mahal}}(\mathbf{x}, \mu_i) = (\mathbf{x} - \mu_i)^{\mathrm{T}} \Sigma_i^{-1} (\mathbf{x} - \mu_i) \tag{2.2}$$

が用いられる場合もある．ここで，Σ_i^{-1} は，暫定的なクラスタ i に属するデータ群の分散共分散行列の逆行列である．マハラノビス距離を用いる場合，代表点が更新されるたびに，それらの代表点から得られる各クラスタにおける Σ_i^{-1} を更新する必要があるため，クラスタリングに要する計算量が増大してしまうという難点がある．さらに，データベクトルの次元数が多い場合，対象データのサンプル数が少ないと Σ_i^{-1} が正確に求まらなくなるという問題点もある．

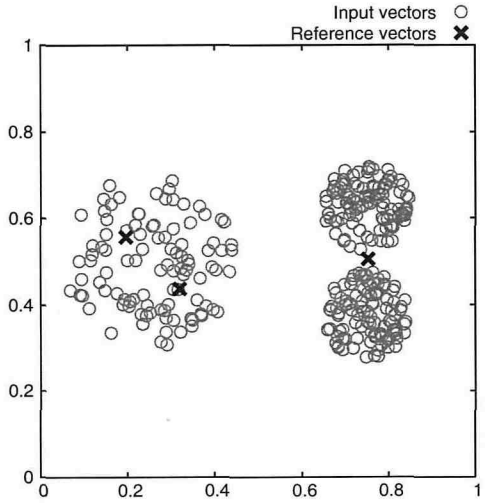


図 2.3 クラスタ内のデータ密度が異なる場合の k-means 法によるクラスタリング結果

2.3 階層的クラスタリング手法

階層的クラスタリング手法は，凝集的手法と分割的手法に大別されるが，一般的に階層的クラスタリングといえば凝集的手法を指す場合が多い．いずれの手法も，あらかじめ定められたクラスタ間の非類似度（クラスタ間距離）尺度に基づいて，徐々にクラスタを併合あるいは分割していく手法である．このとき，凝集的手法は個々のデータをそれぞれ個別のクラスタとみなすことから出発するのに対して，分割的手法ではすべての対象データを一つのクラスタとみなすことから出発する．クラスタの併合あるいは分割が，クラスタ間の非類似度の増加あるいは減少にともなって段階的に行なわれるため，階層的クラスタリングでは，図 2.4(a) に示す対象データから，図 2.4(b) に示す樹形図（デンドログラム）が得られ，デンドログラムからクラスタの併合・分割の過程を視覚的に捉えることができる．さらに，図 2.4(b) の点線によってデンドログラムを分割することによって，図 2.4(c) に示すような具体的なクラスタ分割を得ることもできる．

凝集的な階層的クラスタリングの具体的な手法にはいくつかのバリエーションがあり，代表的なものとして最短距離法 [9]，最長距離法 [10]，Ward 法 [11] が挙げられる．これらの

違いは，クラスタ併合の判定に用いられるクラスタ間距離尺度の定義の仕方であり，各手法において，任意の2クラスタ C_i，C_j におけるクラスタ間距離は，それぞれ以下のように定義される．

- 最短距離法（nearest neighbor method）

$$D(C_i, C_j) = \min_{\mathbf{x}_k \in C_i, \mathbf{x}_l \in C_j} D(\mathbf{x}_k, \mathbf{x}_l) \tag{2.3}$$

- 最長距離法（furthest neighbor method）

$$D(C_i, C_j) = \max_{\mathbf{x}_k \in C_i, \mathbf{x}_l \in C_j} D(\mathbf{x}_k, \mathbf{x}_l) \tag{2.4}$$

- ウォード法（Ward's method）

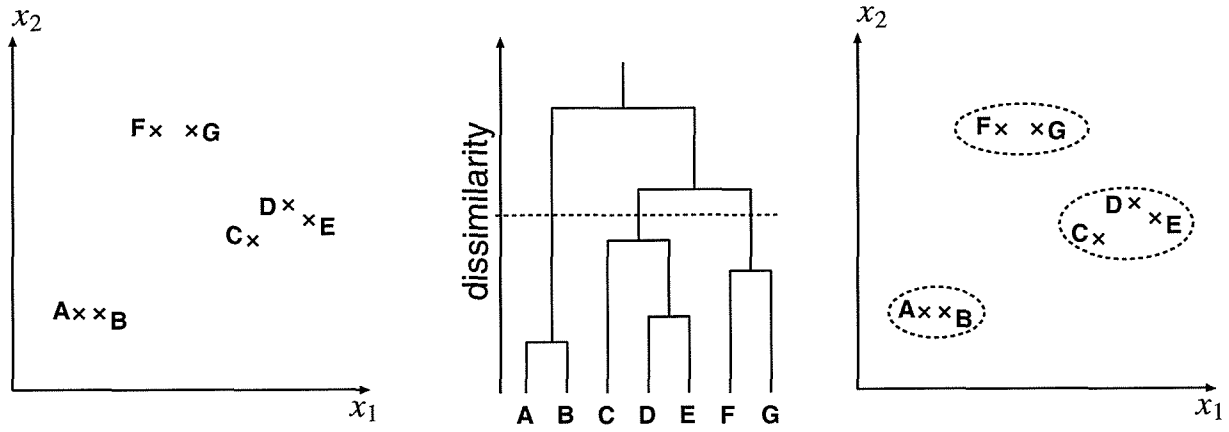
$$D(C_i, C_j) = E(C_i \cup C_j) - E(C_i) - E(C_j) \tag{2.5}$$

$$E(C) = \sum_{\mathbf{x} \in C} (D(\mathbf{x}, \mu_C)) \tag{2.6}$$

ただし，μ_Cはクラスタ C の平均ベクトル

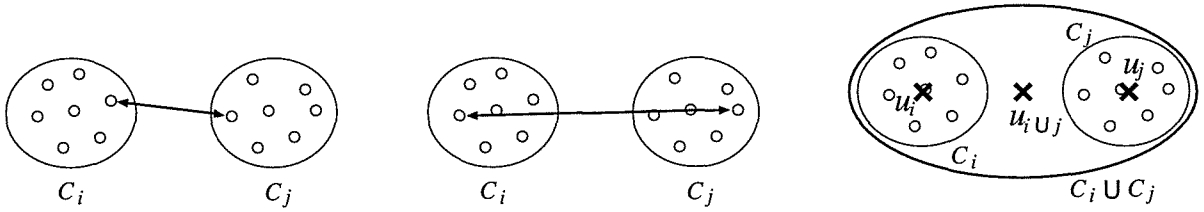
最短距離法では，2クラスタそれぞれに属するデータの中で，最も接近しているデータ同士の距離を，2クラスタ間の距離としている（図2.5(a)参照）．これに対して，最長距離法では，最も離れているデータ同士の距離を，2クラスタ間の距離とする（図2.5(b)参照）．また，ウォード法では，クラスタ併合前後における各データのクラスタ平均ベクトルからの二乗距離の総和の増分が，2クラスタ間の距離となる（図2.5(c)参照）．

これらのクラスタ間距離の定義式により，いま，3つのクラスタ C_1，C_2，C_3 があり，C_1 と C_2 が併合される場合を考えると，併合の前後における C_3 と C_1（あるいは C_2）とのクラスタ間距離が，最短距離法の場合はより短くなり，最長距離法やウォード法の場合はより長くなるという性質がある．このため，特に最短距離法では，局所的な範囲で連鎖的なクラスタ併合が起こりやすい（チェイニング効果と呼ばれる）という特徴がある．



(a) クラスタリング対象データ　(b) 樹形図（デンドログラム）　(c) クラスタ分割結果（3 クラスタ）

図 2.4　階層的クラスタリング



(a) 最短距離法　(b) 最長距離法　(c) ウォード法

図 2.5　階層的クラスタリングにおける非類似度尺度

また，階層的クラスタリングの各手法は，クラスタ併合を行うたびに，クラスタ同士の非類似度をまとめた非類似度行列を更新する必要がある．非類似度行列の大きさは，データ数の二乗に比例するため，階層的クラスタリングの計算量は基本的に $O(n^2)$ であり，ヒープを用いた高速アルゴリズムでは $O(n \log n)$ となることが知られている [28]．したがって，クラスタリング時に必要な計算時間の観点から見れば，階層的クラスタリング手法は，大規模データへの適用が困難であるといえる．

2.4 任意形状のクラスタ抽出を目的とした手法

2.2節で述べた k-means 法や，2.3節で述べた種々の階層的クラスタリング手法は，対象データの局所的な距離関係に基づいてクラスタを求めている．そのため，対象データの分布に何らかの形状的特徴が存在したとしても，その特徴を反映したクラスタ分割が得られるとは限らない．

対象データが任意の分布形状を持っているような場合に，分布の特徴を反映させたクラスタ分割を得るための手法の一つとして，グラフ理論を用いたものが挙げられる．その先駆は Zahn による，最小全域木（MST: Minimum Spanning Tree）を利用した手法である [29]．ここで，MST とは以下の条件を満たしたグラフである．

1. 巡回がない
2. すべてのノードが少なくとも1つのリンクを持つ
3. リンク長の総和が最小である

クラスタリング問題に MST を適用する場合，グラフを構成するノードは，クラスタリング対象の個々のサンプルデータに対応し，ノード同士の接続を表わすリンクのリンク長は，接続される2つのサンプルデータの非類似度となる．

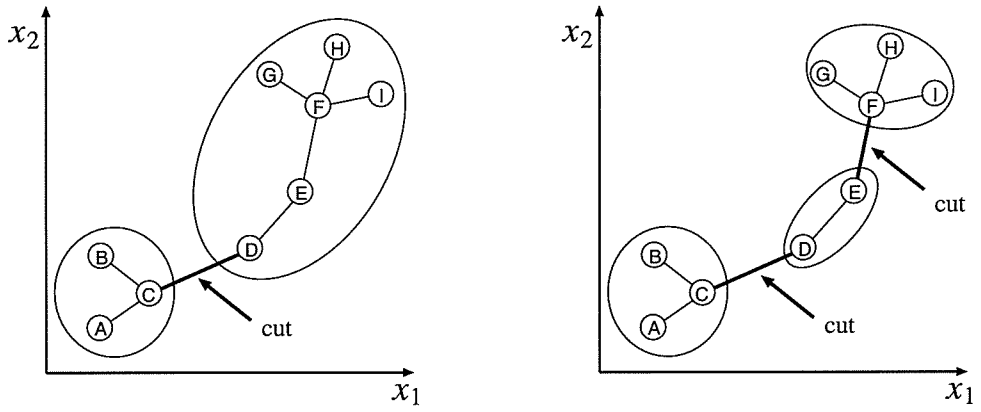
MST を用いたクラスタリングの手順

STEP1: クラスタリング対象データに対して算出される非類似度行列に基づいて，MST を求める．

STEP2: MST において，最大長（非類似度が最大）のリンクを検索し，そのリンクを切断することで，MST を2つに分割する．

STEP3: すべてのリンクが切断されるまで，STEP2 の操作を再帰的に繰り返す．

図2.6は，9個のデータに対して得られた MST による，クラスタリングの様子を示したものである．最大長のリンクから順次 **STEP2** を適用することにより，次第に細かくクラ



(a) 2 つのクラスタに分割する場合　　(b) 3 つのクラスタに分割する場合

図 2.6　MST によるクラスタリングの実行例

スタが分割されて行くことが分かる．このことから，MST を用いた手法は，切断する辺の長さを段階的に短くして行く，分割的な階層的クラスタリング手法として位置付けることもできる．実際に，2.3 節で述べた最短距離法によって得られたクラスタ分割は，MST の部分グラフとなることが知られており [30]，また，最長距離法によって得られたクラスタ分割は，MST の最大完全部分グラフとなることが知られている [31]．

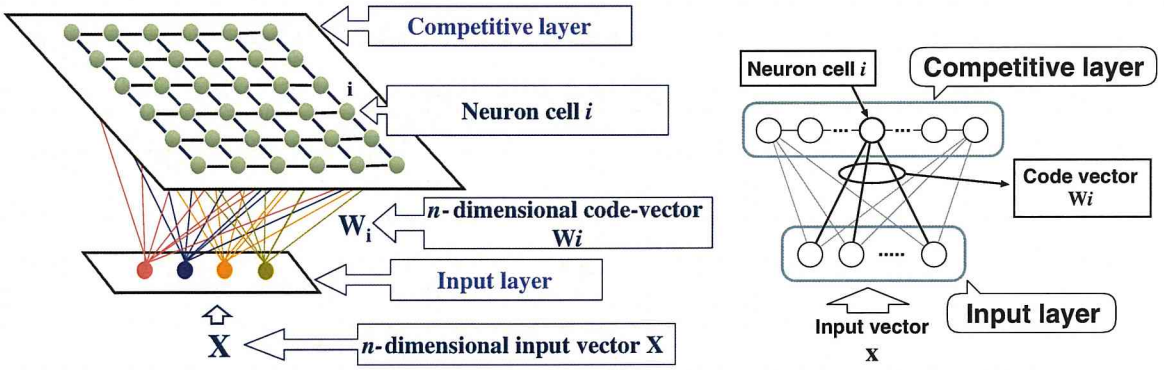
2.5　自己組織化マップ (SOM)

2.5.1　SOM の構造とその学習アルゴリズム

Kohonen によって 1980 年代に提案された自己組織化マップ（Self-Organizing Map:SOM）[12] は，教師なし学習を行うニューラルネットワークの一種であり，図 2.7 に示すように 2 層から構成されるネットワークである．第 1 層は入力層であり，N 次元の入力データを受け取るための入力セルが N 個配置されている．第 2 層は競合層と呼ばれ，入力層からの信号を受けるセルが 2 次元あるいは 1 次元格子状に配置されている．競合層のセルが 2 次元格子状に配置されたものは 2 次元 SOM，1 次元格子状に配置されたものは 1 次元 SOM と

呼ばれる．2次元SOMの場合，競合層の2次元格子には正方格子あるいは六角形格子が一般的に用いられる．競合層の各セルは入力層の全てのセルと重み付きの結合をしているため，各セルはN次元の重みベクトルを持っている．これらの重みベクトルはコードベクトルと呼ばれる．

SOMの基本動作は，N次元入力データ空間から競合層のセルが配置された低次元の離散空間への写像を行うことである．具体的には，入力データに最もよく一致するコードベクトルを持つセルを“勝者セル”とし，そのセルと入力データを対応付けることによって，N次元の入力データを，低次元の離散空間である競合層に写像する．ここで，入力データが写像された競合層は，その入力データの特徴マップと呼ばれる．SOMにおける入力データの学習とは，N次元空間中の入力データの分布の状態が最も良く競合層への写像に反映されるように，コードベクトルを更新することである．



(a) 2 次元 SOM　　(b) 1 次元 SOM

図 2.7　SOM の構造

Kohonen が提案した SOM の基本学習アルゴリズムを以下に示す．本論文では，この学習アルゴリズムに基づく SOM を BSOM（Basic SOM）と呼ぶ．

SOM の基本学習アルゴリズム（BSOM）

STEP1: コードベクトル $\mathbf{w}_i(i = 1, 2, \ldots, N)$ をランダムに（乱数を用いて）初期化する．
STEP2: 入力層に入力ベクトルを提示する．
STEP3: 入力ベクトルに最も類似した（距離の近い）コードベクトルを持つセルを検索し，これを「勝者セル」c とする．
STEP4: コードベクトル $\mathbf{w}_i(i = 1, 2, \ldots, N)$ の更新を次式を用いて行う．

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \alpha(t)\, \Phi^{\mathrm{BSOM}}(p_i)\, (\mathbf{x} - \mathbf{w}_i(t)) \tag{2.7}$$

STEP5: $t = t + 1$ として STEP2 へ戻る．最大学習回数 T まで学習が完了すれば終了する．

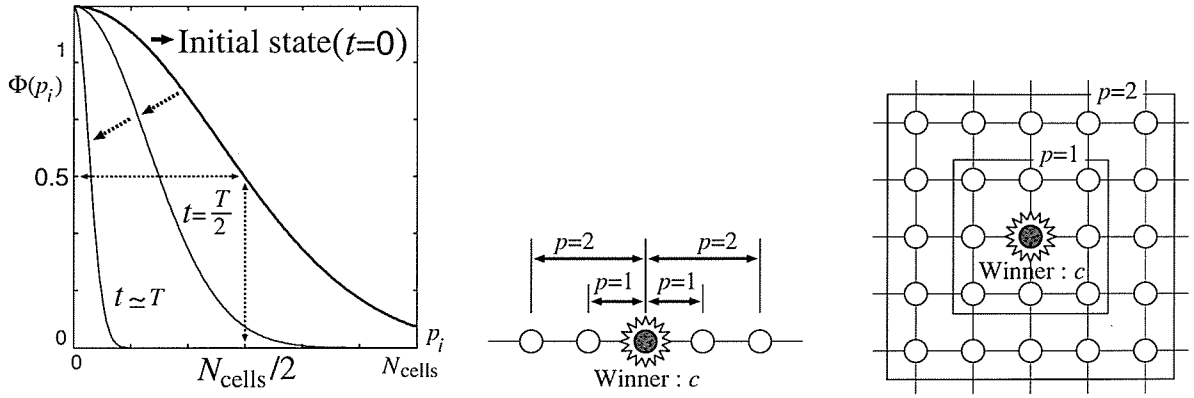
ここで，$\alpha(t)$ は学習回数 t における学習率であり，初期値 α_{ini} から始まり，あらかじめ与えられた最大学習回数 T で最小となるように，t の増加に伴って単調に減少する．また，$\Phi^{\mathrm{BSOM}}(p_i)$ は，図2.8(b) あるいは図2.8(c) に示すような，勝者セル c の第 p 近傍 $(p = 1, 2, \cdots)$ に位置するセルに対して，徐々に学習率を小さくする近傍関数である．ここで，p_i は競合層上でのセル i から勝者セル c までの距離である．

近傍関数の具体的な形としては，式 (2.9) に示すようなガウス型の関数（図 2.8(a) 参照）が一般的に用いられる．式 (2.9) における $\sigma(t)$ は，競合層上での近傍のサイズを定義する時変のパラメータであり，式 (2.7) における $\alpha(t)$ と同様，初期値 σ_{ini} から始まり，学習が進むにつれて単調に減少する．

$$\alpha(t) = \alpha_{\mathrm{ini}} \left(1 - \frac{t}{T}\right) \tag{2.8}$$

$$\Phi^{\mathrm{BSOM}}(p_i) = \exp\left(-\frac{{p_i}^2}{\sigma^2(t)}\right) \tag{2.9}$$

$$\sigma(t) = \sigma_{\mathrm{ini}} \left(1 - \frac{t}{T}\right) \tag{2.10}$$



(a) 近傍関数$\Phi(p)$とσの関係

(b) 勝者セルcと近傍セルとの距離p（1 次元 SOM の場合）

(c) 勝者セルcと近傍セルとの距離p（2 次元 SOM の場合）

図 2.8 競合層における「近傍」の定義と近傍関数の概形

2.5.2 k-means 法に対する SOM の有効性

SOM は競合学習アルゴリズムの一つであり，競合学習によるクラスタリング手法は，ある条件の下ではk-means 法と等価であることが知られている [7].

k-means 法は，k 個の代表ベクトルを$\mathbf{w}_c(c = 1, 2, ..., k)$，$\mathbf{w}_c$ に対応するクラスタをS_c，N 個の入力ベクトルを$\mathbf{x}_i(i = 1, 2, ..., N)$としたときに，以下に示されるような，入力ベクトルと代表ベクトルとの二乗距離の総和Rの最小値を探索する問題に帰着される.

$$R = \sum_{c=1}^{k} \sum_{\mathbf{x}_i \in S_c} \| \mathbf{x}_i - \mathbf{w}_c \|^2 \tag{2.11}$$

また，k-means 法では，次式によって代表ベクトル$\mathbf{w}_c$が更新される.

$$\mathbf{w}_c^{\text{new}} = \mathbf{w}_c^{\text{old}} + \frac{1}{N_c} \sum_{\mathbf{x}_i \in S_c} (\mathbf{x}_i - \mathbf{w}_c^{\text{old}}) \tag{2.12}$$

ここで，N_cは代表ベクトル$\mathbf{w}_c^{\text{old}}$に対応付けられた入力データの個数である.

これに対して SOM では，あるベクトル$\mathbf{x}_i$が入力されたときに，式 (2.7) によってコードベクトルが更新されて行く．ここで，式 (2.7) において近傍関数Φをなくし，αを 1 に固定

し，さらに各セルについて，自分が勝者となった N_c 個の入力データ群に対し，一括して $\mathbf{w}_c$ の更新を行なえば，SOM の学習アルゴリズムは式 (2.12) で示される k-means 法と等価となる．すなわち，SOM の学習アルゴリズムは，k-means 法における参照ベクトルの更新を逐次的に行ない，かつ近傍関数の導入によって，複数の参照ベクトルが同時に更新されるものであるといえる．SOM の学習アルゴリズムの特徴は，式 (2.7) における係数 α や近傍関数 Φ により，k-means 法とは異なり，入力データの学習において，コードベクトルの更新が緩やかに行なわれることである．

2.5.3 SOM における学習の特徴

前節で述べたように，競合学習アルゴリズムの一つである SOM は，k-means 法と本質的に類似している．したがって，SOM の学習アルゴリズムによるコードベクトルの更新は，個々のコードベクトルを k-means 法における代表ベクトルとみなしたときに，入力データとコードベクトルとの二乗距離の総和（式 (2.11) 参照）が最小となるように収束していく．このことは，学習後のコードベクトルの分布が，学習時に与えられた入力データの分布を近似したものになっていることを意味しており，SOM の学習アルゴリズム適用後に得られたコードベクトル群の分布に見られる大きな特徴の一つである．図 2.9(b) および図 2.9(c) は，図 2.9(a) を入力データとして学習した際の，2 次元 SOM および 1 次元 SOM の学習結果を示している．いずれの場合も，学習後のコードベクトルは，入力データが密集している領域に集中して分布しており，コードベクトルの分布が入力データの分布の様子を反映していることが確認できる．

SOM のもう一つの特徴は，入力データの局所的な位相 (topology) が，学習後に得られたコードベクトルの分布に基づいた特徴マップにおいても保存されていることである．ここで，「位相が保存されている」とは，入力データ空間内における個々の入力データの位置関係と，個々の入力データにそれぞれ最も近いコードベクトルを持つセルの，競合層上での位置関係が対応していることである．この性質は，入力データから競合層セルへの位相保

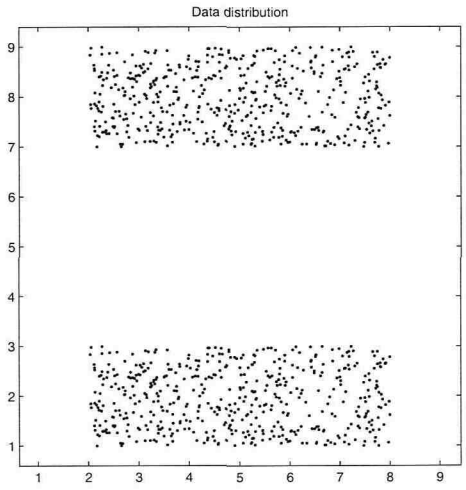
持写像 (topological mapping) と呼ばれる．図2.9(b) および図2.9(c) では，競合層でのセルの隣接関係にしたがってコードベクトル同士を線分で結んでいる．図2.9(b) から，2次元SOMでは入力データが持つ2次元空間中での「上下左右」の位置関係が，競合層におけるセル配列の位置関係と対応していることが確認できる．また，図2.9(c) に示すように，1次元SOMの場合は，セル配列の「左右」の位置関係が，入力データ分布の「上下」あるいは「左右」の位置関係と局所的に対応していることが確認できる．

この位相保持写像は，式(2.7) における近傍関数 Φ^{BSOM} によって実現されている．式(2.9) に示すように，SOMの学習の初期段階では，Φ^{BSOM} における近傍領域の拡がりを決定するパラメータ $\sigma(t)$ の値が大きいため，コードベクトルの更新は勝者セルだけにとどまらず，ほとんどすべてのセルのコードベクトルが，勝者セルのコードベクトルと同じ方向に更新される．これによって，入力データの大域的な位相がコードベクトルの分布に反映される．学習の進行にともなって $\sigma(t)$ の値が減少して行くため，Φ^{BSOM} における近傍領域の拡がりは徐々に縮小して行く．これによって，学習過程の終盤では，コードベクトルの更新は勝者セルおよびその直近のセルに限定されるようになり，入力データの局所的な位相がコードベクトルの分布に反映されるようになる．

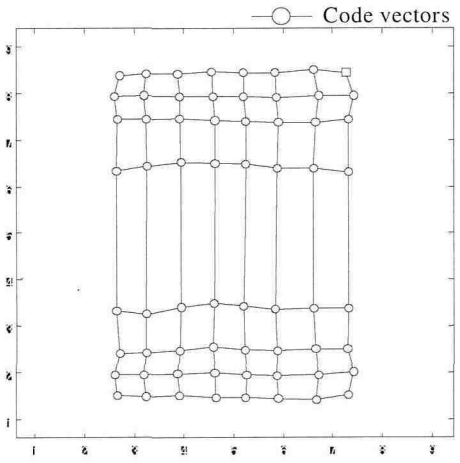
2.6 本章のまとめ

本章では，クラスタリング問題の概要について，種々のクラスタリング手法の分類を行った．特に，代表的なクラスタリングアルゴリズムである k-means 法や，階層的クラスタリング手法である最短距離法，最長距離法，ウォード法，また，グラフ理論を用いた手法については，それらの具体的なアルゴリズムと問題点を説明した．

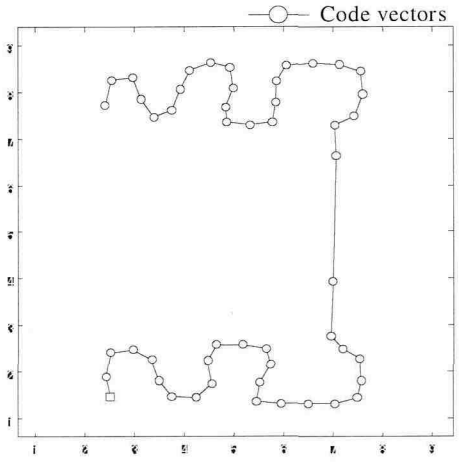
k-means 法や，階層的クラスタリング手法では，個々のサンプル同士の非類似度を求める際にユークリッド距離を用いる場合が多く，そのため，形状的特徴をもつクラスタの抽出が困難である．一方，MSTなどのグラフ理論を用いた手法や，最短距離法のような一部の階層的手法では，クラスタの形状的特徴に対応することが可能であるが，大規模なデー



(a) 入力データ（2 つの 2 次元一様分布）



(b) 2 次元 SOM における学習結果（コードベクトルを入力データの空間にプロット）



(c) 1 次元 SOM における学習結果（コードベクトルを入力データの空間にプロット）

図 2.9 SOM における「データ密度の反映」と「位相保持写像」

タに適用する場合に，非類似度行列を更新する際の計算量が増大するという問題がある．

これに対して，自己組織化マップ（Self-Organizing Map: SOM）は，大規模なデータを少数のコードベクトルで近似でき，また，クラスタが形状的特徴を持っているような場合でも，クラスタの分布形状にコードベクトルの分布を反映させることができる．したがって，SOMをクラスタリング問題に適用すれば，任意形状クラスタの抽出や，大規模データへの適用における計算時間の削減などを同時に実現できると期待される．

次章では，SOMの基本学習アルゴリズムによって得られた特徴マップからのクラスタ抽出法，およびその際に明らかとなるSOMの問題点について述べ，クラスタリング問題に向けたSOMの学習アルゴリズムの改良とその効果について述べる．

第3章

クラスタリング問題への適用に向けた自己組織化マップの改良

3.1 SOMを用いたクラスタリング

3.1.1 学習後の特徴マップを用いたクラスタ抽出法

SOMの学習後に得られる特徴マップでは，競合層の格子上で隣接するセル間のコードベクトルが類似しており，さらに，入力データ空間でのデータの疎密が，学習後のコードベクトルの分布に反映されるという性質があることを2.5.3節で述べた．これらの性質を利用することにより，隣接セル間のコードベクトルが大きく異なる部分をクラスタ境界として検出することで，入力データに内在しているクラスタ群を抽出することが可能となる．このとき，1次元SOMを用いたクラスタリングの具体的な流れは，以下に示すように考えることができる．

1. 特徴マップの作成

   SOMにクラスタリング対象データを学習させ，図2.9(c)に示すようなコードベクトルの並び（マップ）を得る．

2. 特徴マップの解析

   (a) 各セルi $(i = 1, 2, \cdots, m-1)$におけるデータの密度を，セルiとセル$i+1$それ

ぞれのコードベクトル間のユークリッド距離 dW_i として次式で求める．

$$dW_i = \| \mathbf{w}_i - \mathbf{w}_{i+1} \| \tag{3.1}$$

(b) 各セル i $(i = 1, 2, \cdots, m-1)$ におけるデータ密度 dW_i を，その最大値 $dW_{i\_\max}$ と最小値 $dW_{i\_\min}$ に基づいて0〜1に正規化し，これを dW_i' とする．

$$dW_i' = \frac{dW_i - dW_{i\_\min}}{dW_{i\_\max} - dW_{i\_\min}} \tag{3.2}$$

(c) 横軸にセルの番号，縦軸に dW_i' の値をプロットしたグラフを作成する．本論文ではこれ以降，この dW_i' のグラフを「データ密度ヒストグラム」と呼ぶ．このヒストグラムの山の部分に位置するセル i と，セル $i+1$ の間がクラスタの境界であると考えられる．

3. ラベル付け

データ密度ヒストグラムに基づいて競合層を分割し，分割された各セル群ごとに適当なラベルを付ける．

2. のマップ解析までを行うことでクラスタ境界の検出が完了し，3. のラベル付けが終了した時点で，SOMを用いたクラスタリングが完了する．

なお，競合層上で隣接するセル同士のコードベクトル間距離の大小によってクラスタの境界を検出する場合，競合層のセル群が4方向の隣接関係を持つ2次元SOM（図2.7(a)参照）を用いた場合と比較して，2方向の隣接関係しか持たない1次元SOM（図2.7(b)参照）は，クラスタ境界の検出における特徴マップの解析が非常に容易である．したがって本論文では，1次元SOMを用いたクラスタリング手法について今後の議論を進めるものとする．

3.1.2 SOMによるクラスタリングの問題点

BSOMの学習過程において，コードベクトルの更新は勝者セルだけではなく，その近傍に位置するセルでも行なわれる．そのため，BSOMは入力データの位相を保存した学習を

行うことができる．しかし，入力データが複数のクラスタを含んでいる場合，図3.1に示すように，コードベクトルの一部は近傍学習によって複数のクラスタに同じ頻度で移動しようとする．その結果，学習終了後において，入力データがほとんど存在しない領域にコードベクトルが留まっているという状況が起こる．これらのコードベクトルは，入力データの特徴マップ作成に寄与することが少ないため，本論文では，これらのコードベクトルを持ったセルを「不活性セル」と呼ぶ．不活性セルの発生は，SOMの学習アルゴリズムの特徴である近傍学習によって起こる．そのため，例えば図3.2に示すように，最大学習回数など，SOMの学習パラメータを調節することによって不活性セルの発生を抑えることは困難である．

不活性セルが学習後のクラスタ抽出に及ぼす具体的な影響について，人工的に作成した入力データを用いて詳しく説明する．図3.3(a)および図3.4(a)は，作成した2種類の入力データを図示したものであり，それぞれ3個あるいは4個のクラスタから構成されるものである．これらの入力データをBSOMによって学習させた後のコードベクトルの分布と，そこから3.1節で述べた手法によって得られたデータ密度ヒストグラムを，それぞれ図3.3および図3.4に示す．

図3.3(b)および図3.4(b)を見ると，いずれの入力データの場合もクラスタ間にコードベクトルが残留し，不活性セルが発生していることが分かる．このとき，3クラスタデータ（図3.3(a)参照）のように，クラスタ間の境界が明確であるような場合は，図3.3(c)に示すように，データ密度ヒストグラムにおいてクラスタ境界部分で明瞭なピークが確認できる．一方，4クラスタデータ（図3.4(a)参照）のように，非常に接近したクラスタが存在するような場合には，図3.4(c)に示すように，データ密度ヒストグラムの該当部分には非常に曖昧なピークしか形成されない．

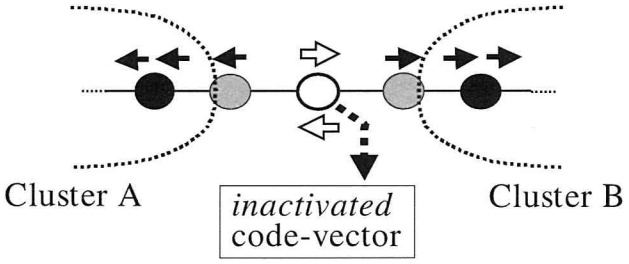
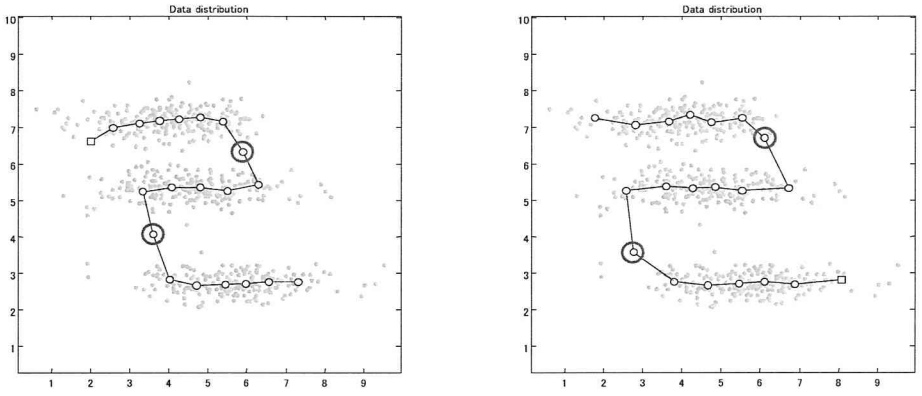


図 3.1　2つのクラスタの中間に位置するコードベクトルの学習中の挙動



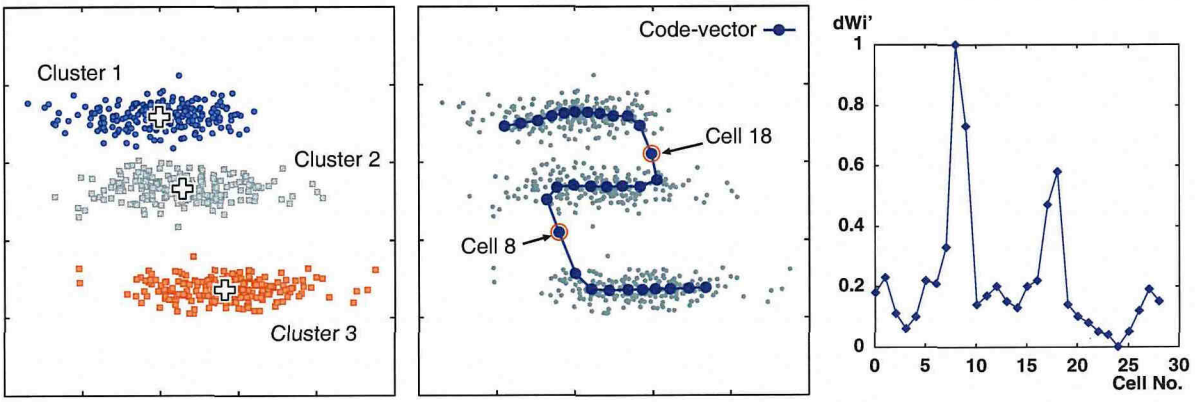
(a) $T_{\max}$=50,000 回　　(b) $T_{\max}$=1,000,000 回

図 3.2　SOM の最大学習回数と「不活性セル」の発生

3.2　しきい値SOM（THSOM）

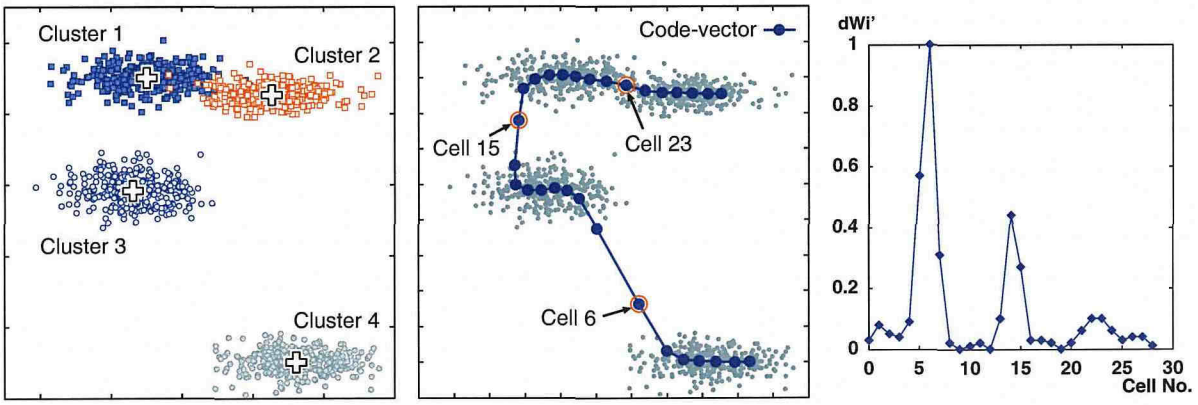
3.2.1　THSOM による不活性セル発生の抑制

SOM をクラスタリングに適用する場合，入力データの位相保持写像の形成が重要であることは，競合層のセルの並びに沿ってマップ解析を行ない，クラスタ境界の推定を行うという，SOM によるクラスタリングの基本的な方針からも明らかである．位相保持写像の形成は，SOM の学習アルゴリズムにおける近傍学習によって実現されているが，その近傍学習によって不活性セルが発生し，マップ形成後のデータ密度ヒストグラムの作成時に悪影



(a) 入力データ（3 クラスタ） (b) 学習後のコードベクトルの分布 (c) 特徴マップの解析結果（データ密度ヒストグラム）

図 3.3 学習終了後のコードベクトル分布とマップ解析結果（3 クラスタデータ）



(a) 入力データ（4 クラスタ） (b) 学習後のコードベクトルの分布 (c) 特徴マップの解析結果（データ密度ヒストグラム）

図 3.4 学習終了後のコードベクトル分布とマップ解析結果（4 クラスタデータ）

響を及ぼしてしまう.

この問題に対して，青木らは，コードベクトルの更新時にしきい値作用を導入した，しきい値SOM(Threshold SOM，以下THSOM)を提案している[17]. 不活性セルの発生を抑えるには，SOMの学習アルゴリズムにおける近傍学習に何らかの制限を加えれば良い. THSOMでは，BSOMのコードベクトル更新式である，式(2.7)における近傍関数Φ^{BSOM}の形式を変更することで近傍学習を制限しており，具体的には，以下に示す形式で関数$\Phi_i^{\mathrm{THSOM}}(t)$が定義されている.

$$\Phi_i^{\mathrm{THSOM}}(t) = \begin{cases} 1 & \text{if cell } i \text{ is “winner”} \\ \exp\left(-\frac{n_{i,j}(t)^2}{\sigma^2}\right) & \text{if cell } i \text{ places around “winner”} \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

ここで，iはセルの番号，tは学習回数を表す. 式(3.3)では，式(2.9)の場合とは異なり，勝者セルおよび競合層内で勝者セルに距離1で隣接するセルに対してのみ$\Phi_i^{\mathrm{THSOM}}(t) > 0$となる. すなわち，THSOMでは勝者セルおよび勝者セルに距離1で隣接するセルのみがコードベクトル更新の対象となり，それ以外のセルでは強制的に$\Phi_i^{\mathrm{THSOM}}(t) = 0$となる. また，$n_{i,j}(t)$はあるセル$i$とその隣接セル$j$との間に設けられたカウンタであり，次式で示す通り，セルiが勝者セルとなり，かつ$\mathbf{w}_i(t)$と$\mathbf{w}_j(t)$のユークリッド距離が，しきい値Thより大きいときに増加する(ただし，$n_{i,j} = n_{j,i}$とする).

$$n_{i,j}(t) = \begin{cases} n_{i,j}(t-1) + 1 & \text{if cell } i \text{ is “winner” and} \\ & \quad \| \mathbf{w}_i(t) - \mathbf{w}_j(t) \| > Th \\ n_{i,j}(t-1) & \text{otherwise} \end{cases} \tag{3.4}$$

すなわち，学習の過程で互いに隣接するセルiおよびセルjのコードベクトルが大きく離れた場合，式(3.4)によってカウンタ値$n_{i,j}$が学習の進行に伴って増加する. そのため，セ

ルiに隣接するセルjが勝者セルになったとしても，式(3.3)においてΦ_i^{THSOM}が非常に小さくなる．したがって，実質的にセルiとセルjの隣接関係が切断され，その結果として，不活性セルの発生を抑えることができる．

3.2.2 THSOMの問題点

SOMの学習においては，事前に対象データの分布に関する情報が与えられることはない．そのため，初期のコードベクトル生成については，コードベクトルの各々の次元に対して，対象データにおいて対応する次元が取り得る値の範囲を調べ，その範囲の一様乱数によって初期値を設定することが一般的である．また，コードベクトルの初期化においては，対象データの位相はまったく考慮されていないため，学習の初期状態における競合層では，対象データに対する位相保持写像は獲得されていない．このような初期状態に対し，BSOMにおいて最終的に位相保持写像が獲得されるのは，コードベクトルの更新時における近傍の範囲（図2.8(a)の近傍関数$\Phi^{\mathrm{BSOM}}(p)$の裾野の広さ）が，学習の初期段階では非常に広いことによる．

これに対してTHSOMでは，式(3.3)に示す通り，コードベクトルの更新は勝者セルと勝者セルに直接に隣接するセルでしか行なわれない．このことは，THSOMはBSOMと比較して，対象データに対する位相保持写像を獲得する能力が極めて低いことを意味している．そのためTHSOMでは，一様乱数によるコードベクトルの初期化を行うと，SOMをクラスタリングに用いる際の大きな特長である位相保持写像の獲得が，著しく阻害されるという問題点がある[20]．したがって，THSOMによって位相保持写像を実現するには，コードベクトルの初期状態を，何らかの方法で対象データの大域的な位相を保持したものにしなければならない．

3.3 2段階SOM

3.3.1 2段階SOMの概要

SOMをクラスタリングに用いる場合，入力データの位相を保存した学習を行うことが極めて重要であり，さらに，精度の良いクラスタリングを行うためには，不活性セルが発生しないことが望ましい．したがって，上記で述べたBSOMとTHSOMの長所だけを取り入れた学習方法を検討することにより，クラスタリングに適したSOMを構築することができると考えられる．THSOMの問題点は，3.2.2節でも述べた通り，学習時におけるコードベクトルの初期設定に留意しなければ，学習後の特徴マップにおいて位相保持写像が行なわれないことである．これに対しBSOMでは，コードベクトルの初期状態に対する制約はTHSOMと比較して弱く，入力データ空間中で一様に分布するコードベクトルを初期状態として与えたとしても，入力データの位相を保存した特徴マップを得ることができる．

そこで，3.2.2節で述べた問題点を解決する方法として，本論文では，BSOMの学習過程によって得られたコードベクトルの分布を，THSOMの学習過程におけるコードベクトル分布の初期状態として利用する学習手法を提案する．提案手法における学習過程の概略は，以下に示すように非常にシンプルであり，本論文では，この提案手法を「2段階SOM」と呼ぶ．

1. BSOMの学習アルゴリズムを適用する
2. 学習パラメータ α_{ini}，σ_{ini}，T の値をリセットする
3. THSOMの学習アルゴリズムを適用する

2段階SOMの学習過程におけるコードベクトルの収束の様子は，模式的には図3.5のように示すことができる．まず，BSOMによる学習過程では，コードベクトルは入力データの分布を反映し，かつ位相を保存した状態に収束する．この時点では，入力データが存在しない，すなわちクラスタの境界と思われる領域にコードベクトルが残留し，不活性セルが生じている．その後，THSOMによる学習過程において，THSOMのしきい値作用によっ

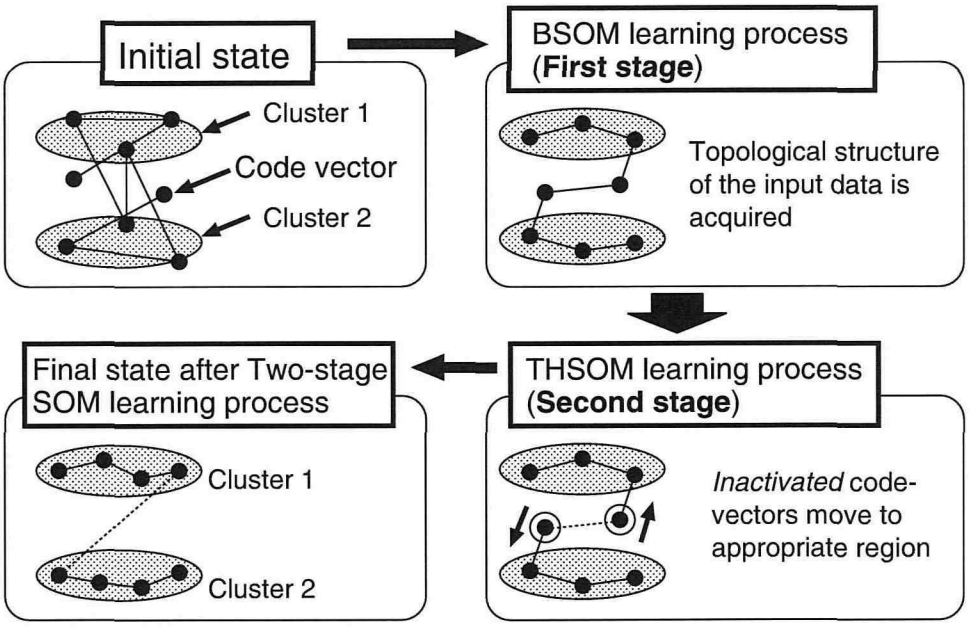


図 3.5 2段階SOMの学習の流れ

て，不活性セルのコードベクトルを適切なクラスタに移動させる．このような2段階の学習過程を経ることにより，2段階SOMではクラスタリングに適した学習結果を得ることが期待できる．

3.3.2 2段階SOMにおけるTHSOM過程

本論文では，2段階SOMのTHSOM過程において，式(3.3)で示される近傍関数Φ^{THSOM}の定義を以下のように変更する．

$$\Phi_i^{\mathrm{2stg}}(t) = \begin{cases} 1 & \text{if cell } i \text{ is “winner”} \\ \eta \times F(t,i) & \text{if cell } i \text{ places around “winner”} \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

ここで，ηは勝者セルに距離1で隣接するセルに対するコードベクトル更新率$(0.0 < \eta \leq 1.0)$

である．また，$F(t,i)$ は，THSOM 過程をコントロールする関数であり，以下のように定義される．

$$F(t,i) = \begin{cases} 1 & \text{if } 0 < t \leq T_{\text{TS1}} \text{ or } (T_{\text{TS1}} < t \leq T_{\text{TS2}} \text{ and } L_i > L_{\text{Th}}) \\ 0 & \text{if } T_{\text{TS1}} < t \leq T_{\text{TS2}} \text{ and } L_i \leq L_{\text{Th}} \end{cases} \tag{3.6}$$

式 (3.6) における L_i は，セル i の活性度を表しており，$t = T_{\text{TS1}}$ 時におけるセル i とその隣接セル間のコードベクトルの距離 D_i と，期間 $0 \leq t \leq T_{\text{TS1}}$ におけるセル i の勝利回数 V_i を組み合わせ，以下のように算出される．

$$L_i = \frac{V_{\text{N\_}i}}{D_{\text{N\_}i}} \tag{3.7}$$

$$V_{\text{N\_}i} = \frac{V_i - V_{\min}}{V_{\max} - V_{\min}} \tag{3.8}$$

$$D_{\text{N\_}i} = \frac{D_i - D_{\min}}{D_{\max} - D_{\min}} \tag{3.9}$$

式 (3.9) における D_i は，セル i におけるコードベクトルの密度を表していると解釈することができ，クラスタ間に孤立した不活性セルと，不活性セルに隣接するセルが類似の値をとらないようにするために，具体的には以下のように算出される．

$$D_i = \min(\| \mathbf{w}_i - \mathbf{w}_{i+1} \|, \| \mathbf{w}_i - \mathbf{w}_{i-1} \|) \tag{3.10}$$

また，$L_i = V_{\text{N\_}i}/D_{\text{N\_}i}$ という形で活性度を定義するのは，不活性セルでは D_i が大きく V_i が小さいという特徴があるためである．

式 (3.5) および式 (3.6) から分かるように，THSOM 過程の学習期間には $0 \leq t \leq T_{\text{TS1}}$ と $T_{\text{TS1}} < t \leq T_{\text{TS2}}$ の 2 つの区分が存在する．期間 $0 \leq t \leq T_{\text{TS1}}$ は，不活性セルの判別を容易にするために，クラスタ間に残存するコードベクトルを孤立させるためのものである．この期間では，勝者セルおよび勝者セルに隣接するセルに限り $\Phi_i^{\text{2stg}}(t) = 1$ あるいは 1 に近い値 $(= \eta)$ とし，不活性セルの持つコードベクトルを，そのコードベクトルを挟む 2 つのクラスタの双方に引き寄せる．その結果，クラスタ間に不活性セルのコードベクトルが 1 つだ

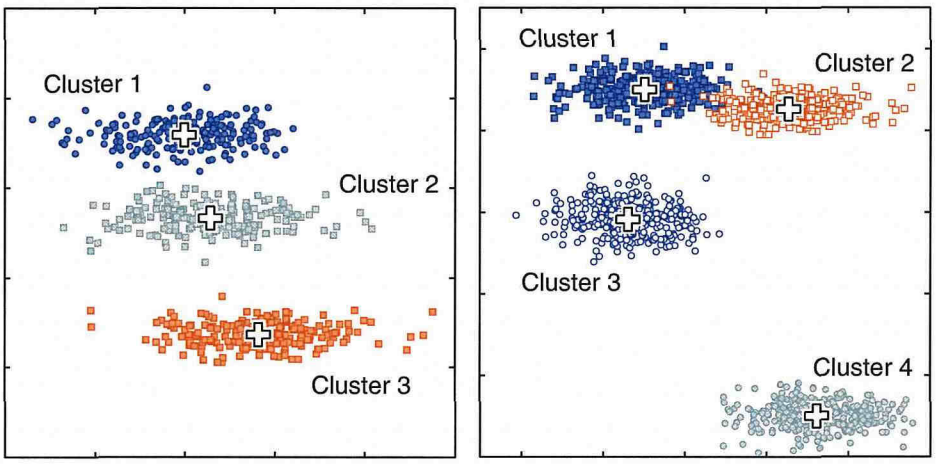
け残される傾向が強くなり，不活性セルの存在が強調される．その後の期間$T_{\mathrm{TS1}} < t \leq T_{\mathrm{TS2}}$において，しきい値による不活性セルの判別が行われる．$t = T_{\mathrm{TS1}}$の時点において，セルiにおける活性度L_iがしきい値（$= L_{\mathrm{Th}}$）以下の場合に，セルiに隣接するセル$i+1$が仮に勝者セルであったとしても$\Phi_i^{\mathrm{2stg}} = 0$とすることで，$T_{\mathrm{TS1}} < t \leq T_{\mathrm{TS2}}$の期間では，セル$i$とセル$i+1$の隣接関係が切断される．

3.4 2段階SOMの学習実験と従来SOMとの比較

3.4.1 実験方法

BSOM，THSOM，2段階SOMそれぞれの学習特性を評価するため，2次元の人工的な入力データを用いた数値実験を行った．この実験の目的は，BSOM，THSOM，および2段階SOMそれぞれにおける学習後のコードベクトル分布を比較することによって，BSOMにおける不活性セルの発生や，THSOMにおいて位相保持写像が得られないなどの問題点を，2段階SOMが解消していることを確認することである．さらに，2段階SOMのTHSOM過程において，セルの勝利回数を考慮した活性度L_iの導入や，不活性セルの強調を意図した学習過程の2分割など，本論文において実装したTHSOMの効果を確認することも実験目的の一つである．

BSOMおよび2段階SOMで共通の学習パラメータとして，競合層のセルの構造は1次元配列で，セル数は30個とし，コードベクトルの初期状態は全てのクラスタを覆う範囲の一様分布とした．その他の学習パラメータとしては，BSOMでは$\alpha_{\mathrm{ini}} = 0.25$，$\sigma_{\mathrm{ini}} = 18.0$，$T = 12,000$回とした．THSOMでは$\alpha_{\mathrm{ini}} = 0.25$，$\sigma_{\mathrm{ini}} = 3.0$，$T = 12,000$回とし，しきい値設定については，3クラスタデータでは$Th = 2.0$，4クラスタデータでは$Th = 3.0$，とした．また，2段階SOMにおけるTHSOM過程においては，$\eta = 0.9$，$T_{\mathrm{TS1}} = 12000$，$T_{\mathrm{TS2}} = 24000$とし，期間$0 \leq t \leq T_{\mathrm{TS1}}$では$\alpha_{\mathrm{ini}} = 0.05$，$\sigma_{\mathrm{ini}} = 3.0$，また，期間$T_{\mathrm{TS1}} < t \leq T_{\mathrm{TS2}}$では$\alpha_{\mathrm{ini}} = 0.05$，$\sigma_{\mathrm{ini}} = 3.0$，$L_{\mathrm{TH}} = 0.3$とした．



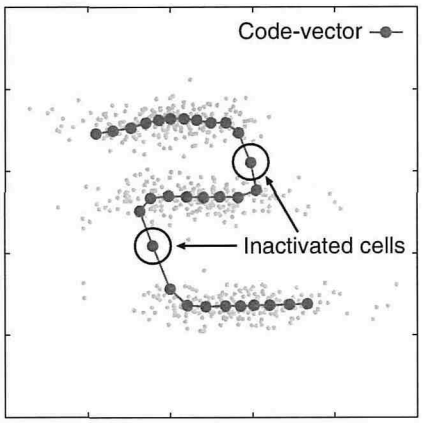
(a) 3 クラスタの入力データ　　(b) 4 クラスタの入力データ

図 3.6　実験対象データ（2 次元正規分布データ）

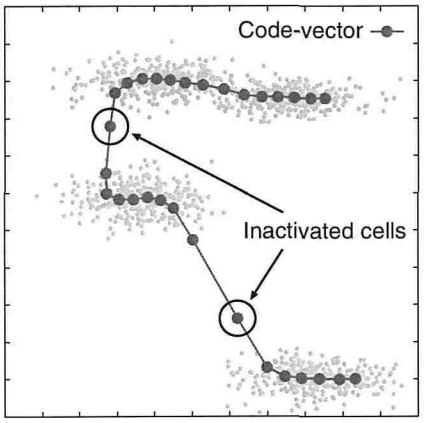
3.4.2　実験 1：学習後のコードベクトルの比較

BSOM および THSOM の学習時の問題点と，それらが 2 段階 SOM において解消されていることを確認するために，図 3.6(a) に示すように，x, y 方向の分散が異なる 2 次元正規分布データ 200 点からなるクラスタ 3 つで構成された入力データを学習させた．図中の＋印は各クラスタの中心ベクトルを表す．このときの BSOM，THSOM，2 段階 SOM それぞれの学習結果を図 3.7 および図 3.8 に示す．図中の●印はコードベクトルを表している．また，コードベクトル同士を結合する線は，それぞれのコードベクトルを持つセルの競合層での結合に対応しており，競合層におけるコードベクトル同士の隣接関係を表している．

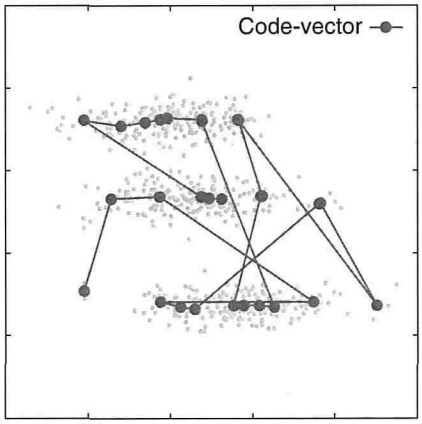
図 3.7(a) および図 3.7(b) では，コードベクトル同士の結合と，競合層におけるセルの 1 次元配列の結合が一致しており，BSOM では入力データの位相保持写像が獲得できていることが分かる．同時に，どちらのクラスタにも属さないコードベクトルを持った不活性セルが複数発生していることも確認できる（図 3.7(a) および図 3.7(b) の実線で囲んだ部分）．一方，THSOM では，図 3.7(c) および図 3.7(d) に示すように，不活性セルの発生は認めら



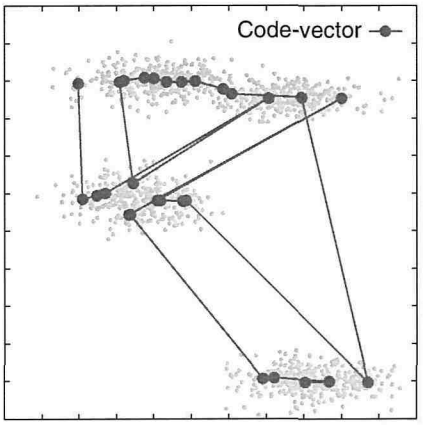
(a) BSOM の学習結果（3 クラスタデータ）



(b) BSOM の学習結果（4 クラスタデータ）

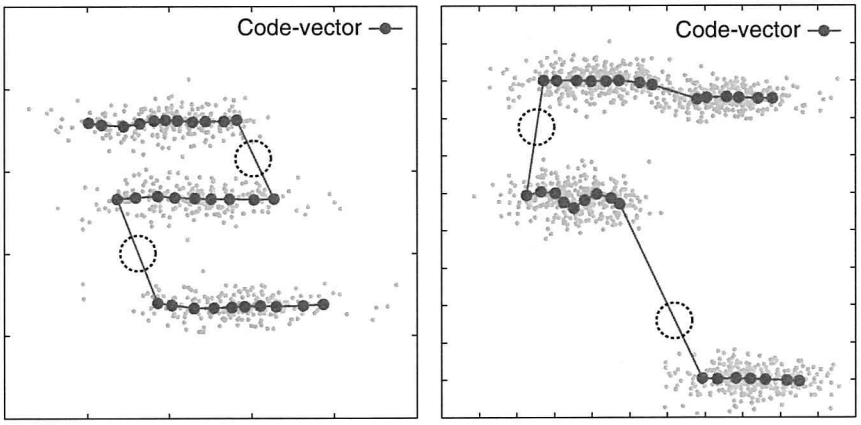


(c) THSOM の学習結果（3 クラスタデータ）



(d) THSOM の学習結果（4 クラスタデータ）

図 3.7　2 次元正規分布データに対する BSOM および THSOM 単独の学習結果



(a) 3 クラスタデータの場合　　(b) 4 クラスタデータの場合

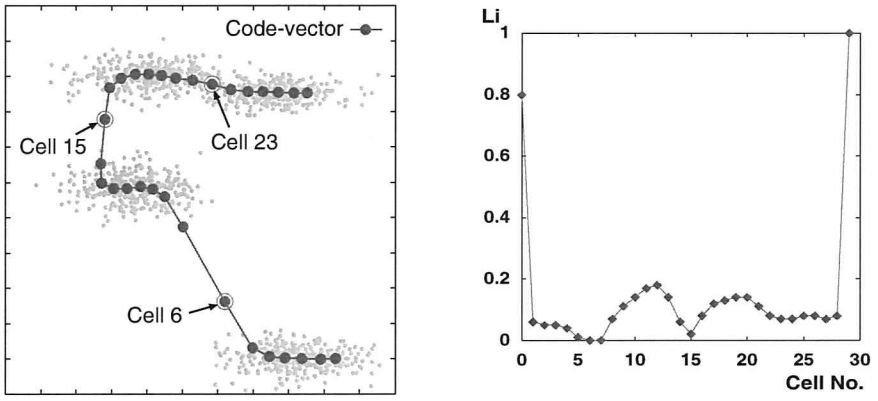
図 3.8　2 次元正規分布データに対する 2 段階 SOM の学習結果

れないが，入力データの位相保持写像が獲得できていないことが分かる．この THSOM の学習結果から，データ密度ヒストグラムのピーク部の検出によって個々のクラスタを正しく抽出することは不可能である．これらに対して 2 段階 SOM による学習結果では，図 3.8 に示すように，BSOM 過程で得られた入力データの位相保持写像が乱されることなく，さらにクラスタ間距離にばらつきがある場合においても，THSOM 過程によってすべてのコードベクトルがいずれかのクラスタに確実に移動しており，不活性セルが発生していないことが確認できる（図 3.8(a) および図 3.8(b) の破線で囲んだ部分）．

3.4.3　実験 2：2 段階 SOM における THSOM 過程の効果

次に，2 段階 SOM において実装した THSOM 過程の効果を確認するための学習実験を行った．入力データとしては，図 3.6(b) に示すような，4 つのクラスタから構成される入力データを用いた．

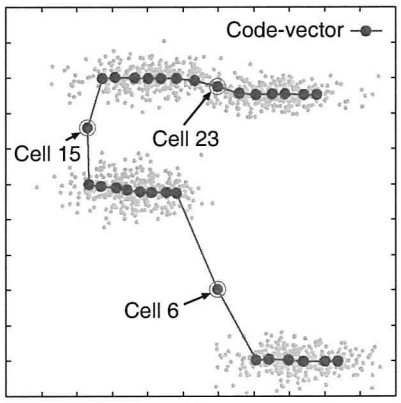
図 3.9(a) および図 3.9(b) は，それぞれ 2 段階 SOM における BSOM 過程の学習結果（図中の●印は学習後のコードベクトルを表す）および，学習後のコードベクトルから算出さ



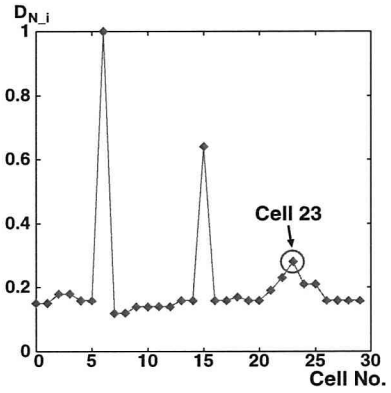
(a) BSOM 学習過程終了後のコードベクトル　　(b) セル i の活性度 L_i のグラフ

図 3.9　2段階SOMにおけるBSOM過程の学習結果と活性度 L_i のグラフ

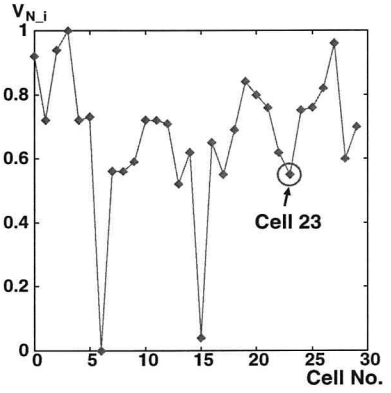
れた，セル i における活性度 L_i のグラフを示している．図3.9(a) におけるセル23のように，接近したクラスタの境界付近に位置するセル群では，図3.9(b) に示すように，L_i のグラフが曖昧なピークしか示さない．このような場合に，セル23を不活性セルとして検出することは困難である．一方，2段階SOMにおけるTHSOMの前半過程（期間 $0 \leq t \leq T_{\mathrm{TS1}}$）の学習結果と，学習後のコードベクトルから算出されたコードベクトル間距離 $D_{\mathrm{N\_}i}$，および学習時の勝利回数 $V_{\mathrm{N\_}i}$ のグラフをそれぞれ図3.10(a)〜図3.10(c) に示す．$D_{\mathrm{N\_}i}$，$V_{\mathrm{N\_}i}$ いずれの場合も，非常に接近したクラスタ間に存在するセル23の部分では曖昧なピークしか示さないが，$D_{\mathrm{N\_}i}$ と $V_{\mathrm{N\_}i}$ を組み合わせた $L_i(= V_{\mathrm{N\_}i}/D_{\mathrm{N\_}i})$ のグラフ（図3.11(a) 参照）では，より明確なピークがセル23の部分に現れている，すなわち，THSOMの前半過程終了後のコードベクトルから得られた L_i のグラフでは，クラスタ間距離の影響が緩和されていることが分かる．図3.11(a) に示す L_i のグラフから，学習期間 $T_{\mathrm{TS1}} < t \leq T_{\mathrm{TS2}}$ において，不活性セルとみなす活性度 L_i のしきい値 L_{TH} を0.35とすることで，セル6-7，15-16，23-24の隣接関係が切断されることが分かる（図3.10(a) の×印）．これに対して，BSOMの学習過程によって得られた L_i のグラフ（図3.9(b) 参照）では，セル23-24の隣接関係を切断す



(a) THSOM 学習過程 ($0 \leq t \leq T_{\mathrm{TS1}}$) 終了後のコードベクトル

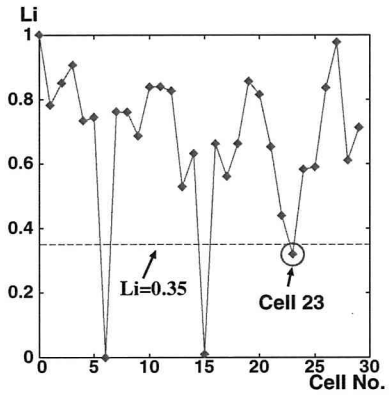


(b) セル i のコードベクトル間距離 D_i のグラフ

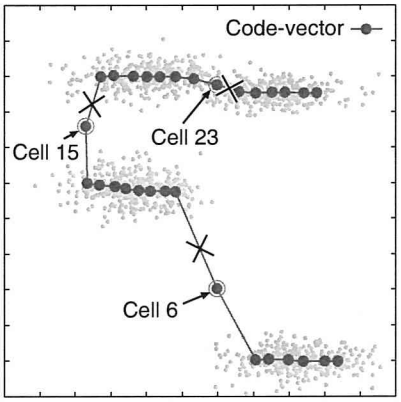


(c) セル i の勝利回数 V_i のグラフ

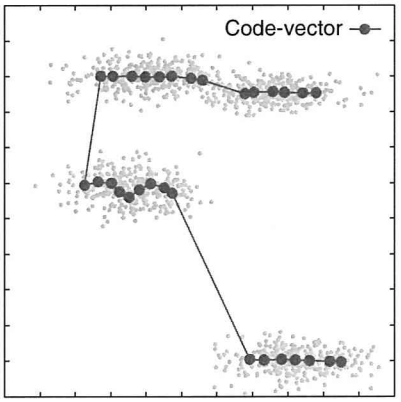
図 3.10 2 段階 SOM における THSOM($0 \leq t \leq T_{\mathrm{TS1}}$) 適用後の学習結果とコードベクトル間距離 D_i および勝利回数 V_i のグラフ



(a) セル i の活性度 L_i のグラフ



(b) THSOM 後半過程における隣接関係の切断位置



(c) 2 段階 SOM 全学習過程終了後のコードベクトル

図 3.11 活性度 L_i のグラフと THSOM($T_{\mathrm{TS1}} < t \leq T_{\mathrm{TS2}}$) 適用後の学習結果

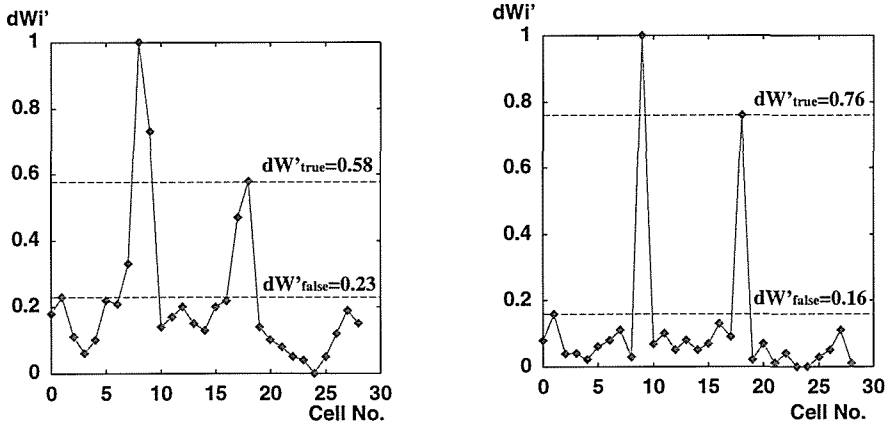
るかどうかの判断は難しい．以上より，3.3.2 節で導入した THSOM の改良によって，2 段階 SOM の THSOM 過程における，期間 $0 \leq t \leq T_{\mathrm{TS1}}$ の学習過程では，BSOM と比較して，より良好な活性度 L_i のグラフが得られることが分かる．

3.5 クラスタ抽出時における2段階SOMの有効性

3.5.1 データ密度ヒストグラムに見られる2段階SOMの利点

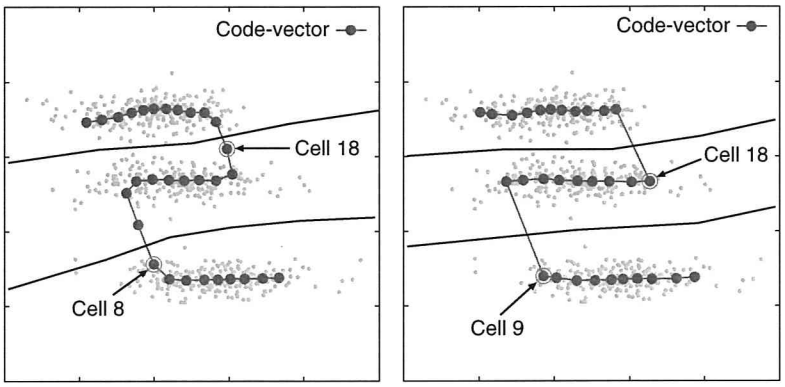
図3.12(a)および(b)は，BSOMおよび2段階SOMによって得られた，3クラスタデータに対する学習結果（図3.7(a)および図3.8(a)参照）から，データ密度dW_i'に基づくデータ密度ヒストグラムを作成した結果である．図3.12(a)および(b)において，$dW_i' = 0.50$程度に設定すれば，それぞれ図3.13(a)，(b)に示すような3つのクラスタを抽出することができる．

また，図3.12(a)，(b)において，クラスタ境界を表すピークの最小値をdW'_{true}，そうでないピークの最大値をdW'_{false}とすれば，BSOMの場合$dW'_{\text{true}} = 0.58$，$dW'_{\text{false}} = 0.23$であるのに対し，2段階SOMでは$dW'_{\text{true}} = 0.76$，$dW'_{\text{false}} = 0.16$であった．したがって，2段階SOMを用いたクラスタリングでは，BSOMを用いた場合に比べて，クラスタ境界を正しく抽出できるdW_i'のしきい値の範囲を，より広く確保することができる．このことから，dW_i'のしきい値の自動設定を考えた場合に，2段階SOMを用いれば，より安定したクラスタリング結果が得られると考えられる．



(a) BSOMの場合　　(b) 2段階SOMの場合

図3.12　3クラスタデータにおけるデータ密度dW_i'のヒストグラム

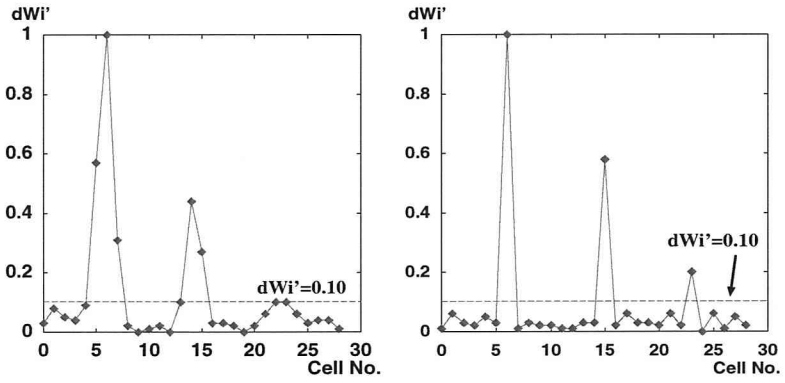


(a) BSOM の場合　　(b) 2 段階 SOM の場合

図 3.13　3 クラスタからなる入力データの学習結果とクラスタリング結果

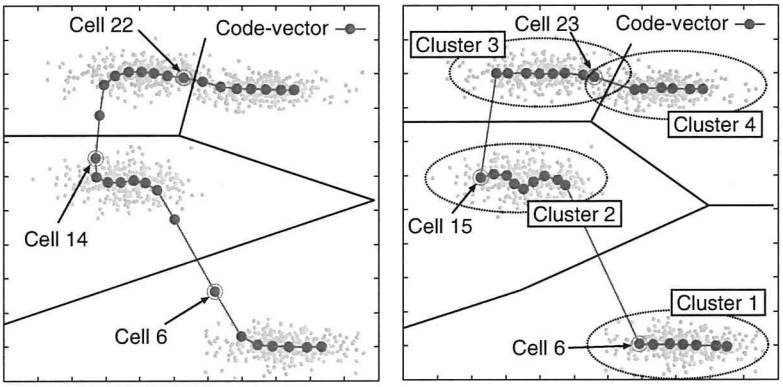
3.5.2　2 段階 SOM を用いた階層的なクラスタ抽出

2 段階 SOM を用いたクラスタリングの利点は，BSOM と比べ，dW_i' のヒストグラムのピークがより明確に形成されることである．クラスタリング問題に対する 2 段階 SOM の有効性を詳しく調べるために，BSOM と 2 段階 SOM の場合についてクラスタリングの追加実験を行なった．入力データは，3.4.3 節の学習実験において，図 3.6(b) で示したものを用いた．この入力データは，1 組のクラスタ間が非常に接近している特徴がある．入力データの学習終了後，マップ解析によって得られた dW_i' のヒストグラムをそれぞれ図 3.14(a), (b) に示す．これらのヒストグラムにおいて，$dW_i' = 0.10$ と設定すれば，BSOM，2 段階 SOM いずれを用いた場合も，図 3.15(a)，(b) に示すような 4 つのクラスタを抽出することができる．このとき，BSOM の場合では $dW'_{\mathrm{true}} = 0.10$，$dW'_{\mathrm{false}} = 0.08$ であるのに対し，2 段階 SOM の場合では $dW'_{\mathrm{true}} = 0.20$，$dW'_{\mathrm{false}} = 0.06$ であった．したがって，クラスタ同士が非常に接近しているような場合においても，2 段階 SOM を用いれば，より正確なクラスタ抽出が可能であることが分かった．さらに，このような 2 段階 SOM の特長を生かせば，dW_i のしきい値設定を変化させることにより，図 3.16 に示すような階層的なクラスタ抽出



(a) BSOM の場合　　(b) 2 段階 SOM の場合

図 3.14　4 クラスタデータにおけるデータ密度 dW_i' のヒストグラム



(a) BSOM の場合　　(b) 2 段階 SOM の場合

図 3.15　4 クラスタからなる入力データの学習結果とクラスタリング結果

(図中の “Cluster No.” は，図 3.15(b) の Cluster1〜4 に対応する）も，BSOM の場合と比較して容易かつ正確に行なうことができる．

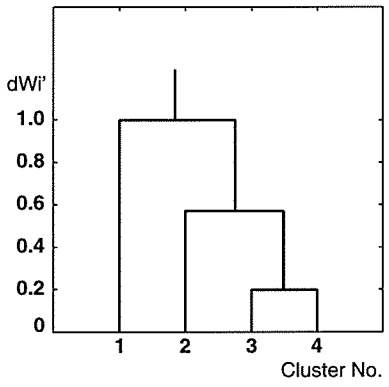


図 3.16 図8(b)に基づく階層的なクラスタ抽出（樹形図）

3.6 本章のまとめ

本章では，SOMの基本学習アルゴリズム（BSOM）と，SOMを用いたクラスタリングの具体的手法を示し，クラスタリング問題への適用における，BSOMの問題点を示した．BSOMの学習アルゴリズムは，ある競合層セルが受けたコードベクトルの更新が，そのセルに隣接したセルにも影響するという，近傍学習と呼ばれる性質がある．この性質によって，学習後のSOMの特徴マップにおける位相保持写像が可能となる．位相保持写像は，SOMを用いたクラスタリングに対して重要な特性であるが，一方で，近傍学習によって，学習後にいずれのクラスタにも属さない不活性セルが発生し，これらの不活性セルが，学習後のクラスタ抽出に悪影響を及ぼす．そこで本章では，BSOMと，近傍学習にしきい値作用を導入した学習アルゴリズム（THSOM）とを段階的に適用する2段階SOMと呼ぶ手法を提案し，人工的に作成した入力データに対して，2段階SOMでは不活性セルの発生が抑制されることを，学習実験によって示した．

次章では，本章で提案した2段階SOMを実際にクラスタリング問題に適用し，その性能評価について詳しく議論する．

第4章

人工データを対象とした2段階SOMのクラスタリング実験

本章では，人工的に作成したクラスタリング対象データを用いて，2段階SOMによるクラスタリングの基礎的な性能評価を行う．その際，k-means法，従来の基本SOM（BSOM），および階層的クラスタリング手法（最短距離法，最長距離法，ウォード法）を比較対象とし，これらの手法に対する，2段階SOMによるクラスタリングの有効性を示す．さらに，本実験によって明らかとなった2段階SOMの問題点について述べる．

4.1 実験方法

4.1.1 クラスタリング対象データ

通常のクラスタリング問題を考えるとき，その対象データには，個々のサンプルがそれぞれどのクラスタに属するかを表わすラベル情報は与えられていない．そのため，対象データに含まれるクラスタ数は未知である．しかしながら，本章では種々のクラスタリング手法の性能を定量的に評価することが目的である．したがって，本章の実験対象となるデータに含まれるクラスタ群については，個々のサンプルはそれぞれ対応するクラスタ番号によってラベル付けされているものとする．

本章におけるクラスタリング実験で用いられる対象データについては，2段階SOMのクラスタリング性能を多面的に評価するため，対象データにおける個々のクラスタが，大き

く分けて以下に示す3パターンの分布特徴を持つように，人工的なデータセットを作成するものとした．

1. 各クラスタが等分散の正規分布であるもの
2. 各クラスタにおけるデータの密度が異なるもの
3. 各クラスタが任意の分布形状を持つもの

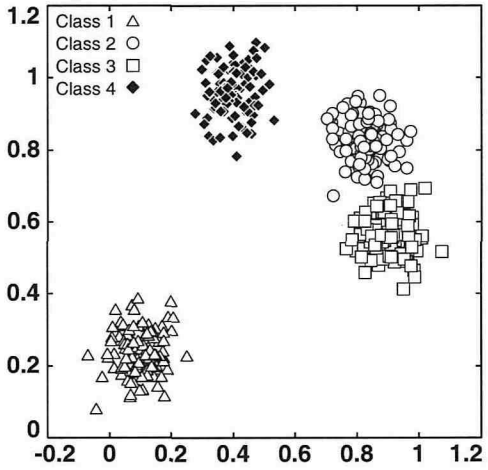
本章では以降，1. の特徴を持つデータセットを「正規分布型データ」，2. および3. の特徴を持つデータセットを「非正規分布型データ」と呼ぶ．

正規分布型データについては，4個あるいは8個のクラスタ中心（正規分布の平均ベクトルμに相当する）を一様乱数によって定め，4クラスタ，8クラスタそれぞれの場合について，表4.1に示すパラメータに基づいた，4個あるいは8個の2次元正規分布で構成されるデータセットを，5パターンずつ作成した．図4.1(a)および図4.1(b)に，4クラスタおよび8クラスタで構成されるデータセットの例を示す．

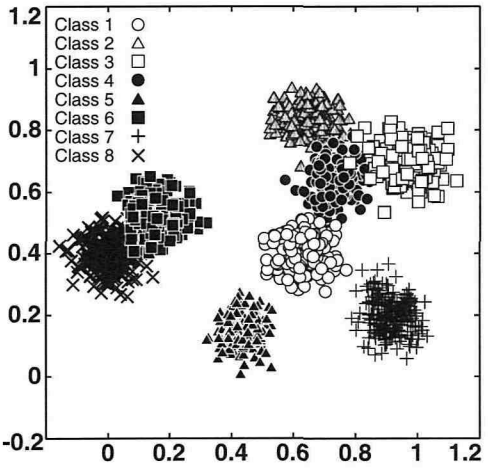
非正規分布型データについて，各クラスタにおけるデータの密度が異なる場合については，密度が異なる円形の2次元一様分布のクラスタ3個から構成されており，サンプル数は300個（100個×3クラスタ）である．また，任意の分布形状を持つ場合については，非等方的な2次元正規分布に対して2次の非線形変換を施してクラスタの形状を歪ませ，歪んだ形状のクラスタ2つから構成されるデータセットを作成した．このデータセットのサンプル数は1000個（500個×2クラスタ）である．図4.1(c)および図4.1(d)に，密度の異なるクラスタを持つデータセット，および任意の分布形状のクラスタを持つデータセットを示す．

4.1.2 クラスタリング結果の評価方法

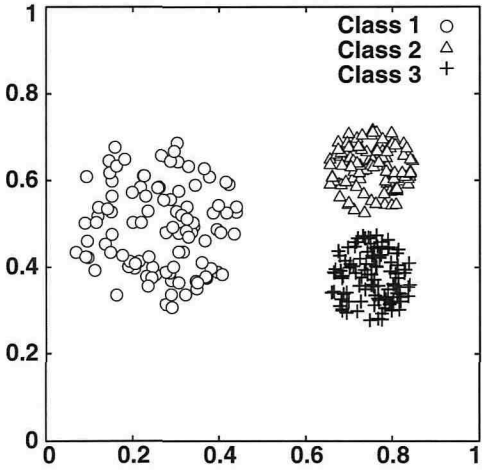
各手法におけるクラスタリング結果を定量的に評価するために，本実験では，クラスタリング終了後に個々のデータに割り振られたラベルと，それらのデータにあらかじめ付加されていたラベルとを照合したときの誤分類率P_{Err}を式(4.1)によって求め，クラスタリン



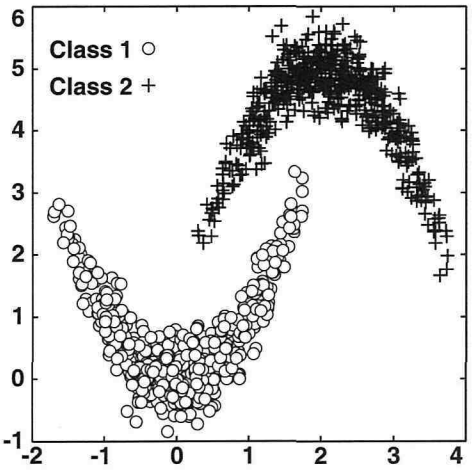
(a) 正規分布型 (4 クラスタ)



(b) 正規分布型 (8 クラスタ)



(c) 非正規分布型, "Type 1" (3 クラスタ)



(d) 非正規分布型, "Type 2" (2 クラスタ)

図 4.1 クラスタリング対象データの例

表 4.1 クラスタリング対象データ（正規分布型）のパラメータ

	各クラスタの中心点 μ_{each}	クラスタ毎の標準偏差 σ_{each}	総サンプル数 N_{total}
4クラスタデータ	一様乱数により4点を生成	0.06	400 (100 × 4)
8クラスタデータ	一様乱数により8点を生成	0.05	1600 (200 × 8)

グ性能を示す指標とした．

$$P_{\text{Err}} = \frac{N_{\text{Err}}}{N_{\text{total}}} \times 100 \tag{4.1}$$

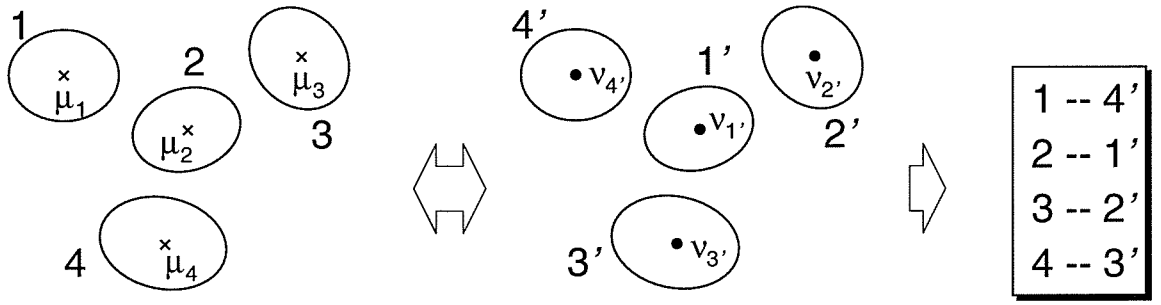
ここで，N_{Err} は誤分類数であり，N_{total} はクラスタリング対象データの総サンプル数である．また，クラスタリングによって抽出された個々のクラスタには，適当なラベルが自動的に割り当てられている．そのため，正解ラベルに基づくクラスタ番号と，クラスタリングによって自動的に割り当てられたクラスタ番号とのマッチングを行う必要がある．本実験では，図4.2に示すように，対象データにおける各クラスタの平均ベクトル群 μ_i と，クラスタリングによって抽出された各クラスタの平均ベクトル群 $\nu_{j'}$ との間の，すべての μ_i と $\nu_{j'}$ の組合せから，式(4.2)に示すように，平均ベクトル同士の距離の総和が最小である組合せを検索することで，クラスタ番号のマッチングを行った．

$$D_{\text{match}} = \min_{\{i\}} \sum_{j'} |\mu_{\{i\}} - \nu_{j'}| \quad , \quad \{i\} \in \text{Perm}(N) \tag{4.2}$$

ここで，$\text{Perm}(N)$ は，1〜N までの整数からなる順列組合せの数列を表わす．

4.1.3 各手法のクラスタリング実行時の設定

従来のクラスタリング手法のうち，本論文では非階層的手法として k-means 法を，また，階層的クラスタリング手法として最短距離法，最長距離法およびウォード法を比較対象とした．本実験で用いるクラスタリング対象データはラベル付きであるため，クラスタ数は既知である．したがって，k-means 法における代表ベクトル数 k は，対象データの既知の



(a) 正解ラベルに基づくクラスタ番号　(b) 抽出されたクラスタ群に付けられたラベル　(c) マッチング結果

図 4.2 抽出されたクラスタと正解クラスタとのマッチング法

クラスタ数に設定した．また，階層的クラスタリングの各手法については，対象データの非類似度を求めるための距離尺度としてユークリッド平方距離を用い，既知のクラスタ数となるまでデータの併合を繰り返すこととした．

k-means 法，BSOM および 2 段階 SOM では，クラスタリング開始時の参照ベクトルあるいはコードベクトルの初期状態によって，クラスタリング結果にばらつきが生じる．そのため，一つの入力データセットに対して，k-means 法では k 個（2 クラスタのとき $k = 2$，3 クラスタのとき $k = 3$）の初期クラスタ中心を，SOM を用いた手法では各セルが持つコードベクトルの初期状態をそれぞれランダムに変化させ，複数回の試行を行って各試行において算出された誤分類率を平均した．具体的な試行回数については，正規分布型データセットの場合，4 クラスタあるいは 8 クラスタについて，それぞれ作成した 5 パターンのデータセットに対し，1 パターンにつき 5 回の試行を行ない，計 25 回（5 パターン × 5 回）の試行を行った．非正規分布型データセットについては，それぞれについて 100 回の試行を行った．なお，階層的クラスタリングの各手法については，初期状態が一意に定まるため，試行回数は 1 回である．

BSOM，および 2 段階 SOM の BSOM 過程の学習回数は，学習サンプル数の 200 倍とし，

表 4.2 BSOMおよび2段階SOMによるクラスタリング実行時のパラメータ設定

BSOM				
	正規分布 4 クラスタ	正規分布 8 クラスタ	非正規分布 Type 1	非正規分布 Type 2
競合層のセル数 N_{cell}	20	40	40	40
dW_i' の しきい値 dW'_{TH}	0.18	0.19	0.31	1.00
2段階SOM				
	正規分布 4 クラスタ	正規分布 8 クラスタ	非正規分布 Type 1	非正規分布 Type 2
競合層のセル数 N_{cell}	20	40	40	40
L_i の しきい値 L_{TH}	0.16	0.19	0.10	0.00
dW_i' の しきい値 dW'_{TH}	0.33	0.35	0.35	1.00

"非正規分布 Type 1": クラスタごとの密度が異なるデータセット（図 4.1(c) 参照）
"非正規分布 Type 2": 任意の分布形状のクラスタを持つデータセット（図 4.1(d) 参照）

2段階SOMのTHSOM過程においては，T_{TS1} および T_{TS2} を，それぞれ学習サンプル数の200倍および400倍とした．さらに，SOMを用いた場合の競合層のセル数 N_{cell}，学習後にクラスタ抽出を行う際に用いるデータ密度ヒストグラムのしきい値 dW'_{TH}，および，2段階SOMのTHSOM過程において，不活性セル判定のための，活性度 L_i のグラフのしきい値 L_{TH} については，表4.2に示すように定めた．ここで，競合層セル数についてはいずれも経験的に定めたものである．また，dW'_{TH} および L_{TH} については，各試行においてそれぞれ既知数のクラスタを抽出するように，実験用プログラムが自動的に設定する．そのため，表4.2では，それぞれのデータセットに対する100試行における平均値を示している．

4.2 実験結果

4.2.1 正規分布型データに対するクラスタリング性能の比較

正規分布型の，4クラスタおよび8クラスタのデータセットに対して，k-means法，BSOMおよび2段階SOMを適用したときの誤分類率を表4.3に示す．表中の値は，25回（5パターンの入力データ×5試行）の平均値である．また，本実験では，誤分類率が5％を上回っ

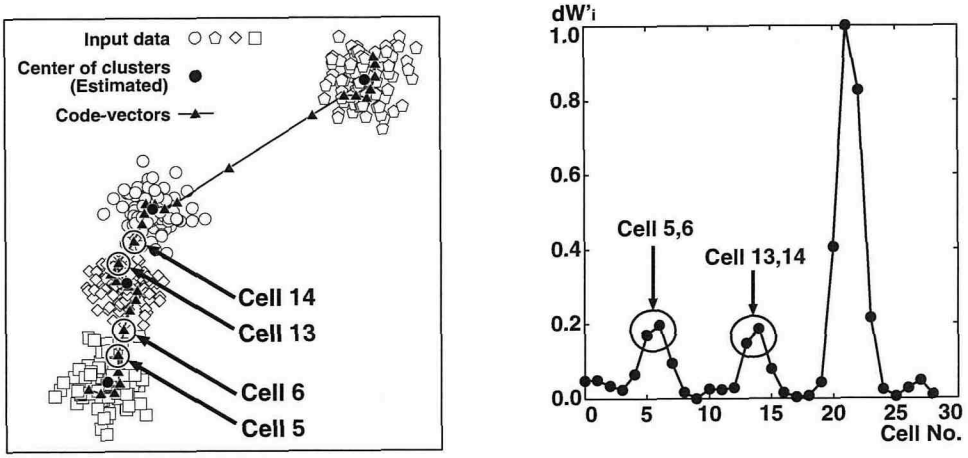
表 4.3 正規分布型データに対する BSOM，2 段階 SOM，k-means 法の誤分類率 $P_{\mathrm{Err}}(\%)$ の比較

	2 段階 SOM	Basic SOM	k-means 法
正規分布型 4 クラスタデータ	0.72 (0/25)	4.5 (2/25)	8.4 (5/25)
正規分布型 8 クラスタデータ	1.3 (0/25)	2.4 (1/25)	20.3 (17/25)

た場合に，クラスタリングに失敗したとみなしており，括弧内の数値は 25 回の試行のうちクラスタリングに失敗した回数を示している．

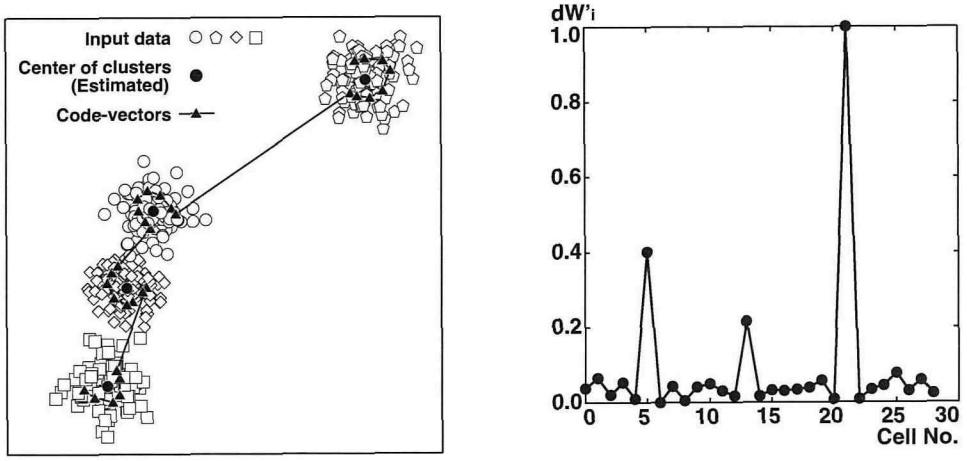
k-means 法では，代表ベクトルの更新において，入力データの大局的な構造に関する知識は与えられない．したがって，クラスタリングの初期段階で代表ベクトルが局所最適解に収束し，クラスタリングに失敗してしまうケースが多い．この傾向は，クラスタ数すなわち k-means 法における代表ベクトル数の増加にともなって顕著となる．そのため，表 4.3 に示すように，BSOM や 2 段階 SOM を用いた場合と比較して誤分類率の平均値が高くなっており，特に 8 クラスタデータの場合に，誤分類率の平均値が高くなっている．一方，SOM の場合，各セルのコードベクトルは，学習アルゴリズムにおける近傍関数の作用によって，入力データの大局的な構造を反映しながら徐々に拡散して行く．そのため，BSOM および 2 段階 SOM では，k-means 法に対して良好な結果を得ており，特に 2 段階 SOM は，誤分類率およびクラスタリング失敗回数ともに，BSOM や k-means 法と比較して最も低い値となっている．

BSOM および 2 段階 SOM について，クラスタの抽出が正しく行なわれた場合の，学習後のコードベクトルおよび推定されたクラスタ中心と，データ密度 dW_i' によるデータ密度ヒストグラムを，それぞれ図 4.3，図 4.4 に示す．図 4.3(a) および図 4.4(a) において，▲は学習終了後の BSOM あるいは 2 段階 SOM におけるコードベクトルを示している．BSOM，2 段階 SOM それぞれにおける学習後のコードベクトルの分布から，データ密度ヒストグラ



(a) クラスタリング結果　　(b) データ密度ヒストグラム

図 4.3　BSOM によるクラスタリング結果とデータ密度ヒストグラム



(a) クラスタリング結果　　(b) データ密度ヒストグラム

図 4.4　2 段階 SOM によるクラスタリング結果とデータ密度ヒストグラム

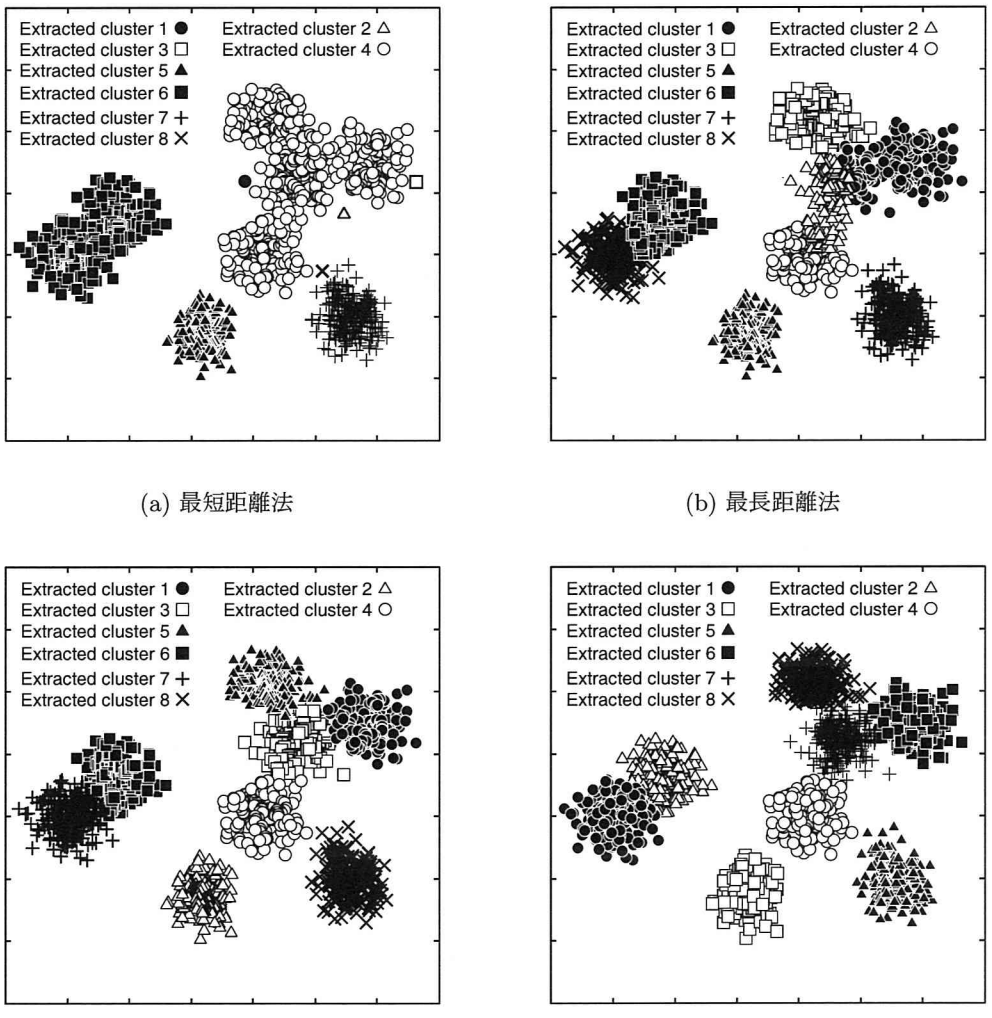
ムを求めた結果が図4.3(b)および図4.4(b)である．2段階SOMでは，THSOMによる2段階目の学習過程の効果によって，不活性セルの発生が抑えられており，その結果，データ密度ヒストグラムではクラスタ境界を示すピークが明確に現れていることが分かる（図4.4のセル5,13および21）．これに対してBSOMの場合では，図4.3(a)で示すように，クラスタの境界付近に残留しているコードベクトルを持った不活性セル（図4.3のセル5,6および13,14）により，推定されるクラスタ境界にズレを生じる．そのため，クラスタ境界付近のデータが正しく分類されず，このことが誤分類数の増加につながっていると考えられる．

次に，2段階SOMと，階層的クラスタリング手法である最短距離法，最長距離法およびウォード法におけるクラスタリング性能を比較する．表4.4は，それぞれの手法における誤分類率の値をまとめたものである．いずれのデータセットにおいても，階層的クラスタリング手法の中ではウォード法が最も低い誤分類率を示しているが，2段階SOMはウォード法と比較して，さらに低い誤分類率を示すことが確認できる．これらに対して，最短距離法では誤分類率が極端に高くなっている．図4.1(a)および図4.1(b)からも分かるように，使用した人工データには，複数のクラスタが互いに接近している領域が存在している．そのため，最短距離法では2.3節で述べたチェイニング効果によって，誤分類率が非常に高くなってしまったと考えられる．このことを確認するために，8クラスタで構成される正規分布型データに対して，2段階SOMを適用した場合のクラスタ抽出結果，および階層的クラスタリング手法において，既定のクラスタ数（=8）でクラスタの併合を終了したとき

表 4.4　正規分布型データに対する2段階SOMと階層的クラスタリング手法の誤分類率 $P_{\mathrm{Err}}(\%)$ の比較

	2段階SOM	最短距離法	最長距離法	ウォード法
正規分布型 4クラスタデータ	0.72	29.5	1.3	0.9
正規分布型 8クラスタデータ	1.3	47.8	5.1	1.6

のクラスタ併合状態を図4.5に示す．最短距離法ではチェイニング効果によって正しく分離されていないクラスタが確認できる（図4.5(a)参照）．また，最長距離法においても，図4.5(b)に示すように，一部のクラスタが正しく分離されていない．これらに対し，ウォード法（図4.5(c)参照）や2段階SOMを用いた手法（図4.5(d)参照）では，入力データにおけるクラスタ群とほぼ合致した状態のクラスタ群が得られていることが分かる．



(a) 最短距離法　(b) 最長距離法

(c) ウォード法　(d) 2段階SOM

図 4.5　階層的クラスタリングおよび2段階SOMによるクラスタ併合状態（正規分布型，8クラスタ）

表 4.5 非正規分布型データに対する 2 段階 SOM と BSOM，k-means 法の誤分類率 $P_{\mathrm{Err}}(\%)$ の比較

	2 段階 SOM	Basic SOM	k-means 法
非正規分布型 Type 1	13.8 (37)	40.2 (85)	24.6 (48)
非正規分布型 Type 2	0.38 (0)	0.48 (0)	2.4 (0)

非正規分布型 Type 1: クラスタごとの密度が異なるデータセット（図 4.1(c)）
非正規分布型 Type 2: 任意の分布形状のクラスタを持つデータセット（図 4.1(d)）

4.2.2 非正規分布型データに対するクラスタリング性能の比較

表 4.5 は，2 種類の非正規分布型データそれぞれに対する，2 段階 SOM，BSOM および k-means 法の，100 試行における平均誤分類率をまとめたものである．また，括弧内の数字は，100 試行のうち，誤分類率が 20%を超えたため，正しいクラスタ分割が得られていないと判断された回数である．

密度の異なる入力データ（図 4.1(c) 参照）に対して，k-means 法は誤分類率が非常に高いことが確認できる．このとき，k-means 法における 100 試行すべてのクラスタ分割結果を確認したところ，図 4.1(c) に示すような正しいクラスタ分割が得られない場合が，100 試行中 48 回発生していた．また，2 段階 SOM では，k-means 法に比べて誤分類率の減少が認められるが，正規分布型データの場合のような大幅な減少には至っておらず，100 試行中 37 回は正しいクラスタ分割が得られていなかった．

次に，階層的クラスタリング手法においては，入力データの特徴によって，正しいクラスタ分割が得られる手法が異なる結果となった．最短距離法では，チェイニング効果によって，図 4.1(c) における Class2 と Class3 のように非常に接近したクラスタがある場合には，これらが一つのクラスタとみなされてしまう反面，図 4.1(d) のような入力データに対しては，正しいクラスタ分割が得られる場合が多い．また，最長距離法およびウォード法については，k-means 法と同様に，クラスタ間距離の定義においてクラスタ形状が考慮されて

表 4.6 非正規分布型データに対する2段階SOMと階層的クラスタリング手法の誤分類率 $P_{\mathrm{Err}}(\%)$ の比較

	2段階SOM	最短距離法	最長距離法	ウォード法
非正規分布型 Type 1	13.8	33.7	0	0
非正規分布型 Type 2	0.38	0	3.2	3.7

非正規分布型 Type 1: クラスタごとの密度が異なるデータセット（図 4.1(c)）
非正規分布型 Type 2: 任意の分布形状のクラスタを持つデータセット（図 4.1(d)）

いないため，図 4.1(d) のような入力データに対して，クラスタ境界部分で誤分類が生じてしまう．そのため，k-means 法と同程度の誤分類率となっている．

4.2.3 2段階SOMの問題点

本節では，非正規分布型データのうち，個々のクラスタにおけるデータの密度が異なるデータセットに対して，2段階SOMにおける誤分類率が大きく改善されない原因について考察する．

SOMの学習過程において，競合層の各セルが持つコードベクトルは，入力データの分布を反映するように更新される．ここで，入力データの個々のクラスタにおけるデータの密度が，クラスタごとに大きく異なる場合，密度の高いクラスタに位置するコードベクトル群と，密度の低いクラスタに位置するコードベクトル群とでは，コードベクトル間距離の全体的な値が大きく異なってしまう．

一方，2段階SOMのTHSOM過程では，BSOM過程で生じた不活性セルを，しきい値によって判別するために，コードベクトル間距離 D_i と，各セルの勝利回数 V_i に基づいた，式 (3.7) で定義される，セル i の活性度 L_i が用いられる．このとき，個々のクラスタに位置するコードベクトル群の間で，コードベクトル間距離 D_i の平均値が大きく異なっていると，それにともなって L_i の値もクラスタごとに大きく異なってしまう．したがって，各セルにおける L_i 値の大小では，個々のセルが不活性セルに該当するかどうかを正確に判定す

ることが困難となる．

THSOM過程における不活性セルの判定は，2段階SOMの学習終了時でのコードベクトルの分布に影響を与える．個々のクラスタにおけるデータの密度が異なる場合，2段階SOMのTHSOM過程における不活性セルの判定が正しく機能せず，その結果として，学習後のクラスタ抽出が正しく行なわれなかったものと考えられる．

4.3 本章のまとめ

本章では，第3章において提案した2段階SOMについて，人工的に作成した様々なデータセットを用いて，そのクラスタリング性能を，BSOM，k-means法，および階層的クラスタリング手法である最短距離法，最長距離法，ウォード法と比較した．一連の実験結果から確認された，2段階SOMに関する所見を以下にまとめる．

まず，正規分布型データに対する実験結果から，k-means法では，クラスタ数の増加にともなってクラスタリングに失敗する頻度が増加することが確認された．これに対し，SOMを用いた場合では，入力データの学習時に，各セルのコードベクトルは入力データの大局的な構造を反映しながら徐々に更新されるため，クラスタ数の増加に関わらず安定したクラスタリング結果が得られることが確認された．特に，2段階SOMは，BSOMと比較して，入力データの学習時に不活性セルが生じないため，より正確にクラスタの境界を推定することができ，結果として，クラスタリング終了後の誤分類数を安定して低く抑えることができる．

次に，非正規分布型データに対する実験では，任意の分布形状のクラスタを持つデータセットの場合，2段階SOMは，k-means法や最長距離法，ウォード法と比較して誤分類率の低下が認められたが，BSOMに対する誤分類率は同程度であった．また，密度が異なるクラスタで構成されるデータセットの場合，2段階SOMはk-means法，最短距離法に対して誤分類率の低下が確認でき，BSOMに対する大幅な誤分類率の改善も認められた．しかしながら，改善の程度は高くなく，更なる改良が必要なことが示唆された．

次章では，4.2.3節における考察をもとに，2段階SOMの学習手法の改良法を提案し，その改良の効果について議論する．

第5章

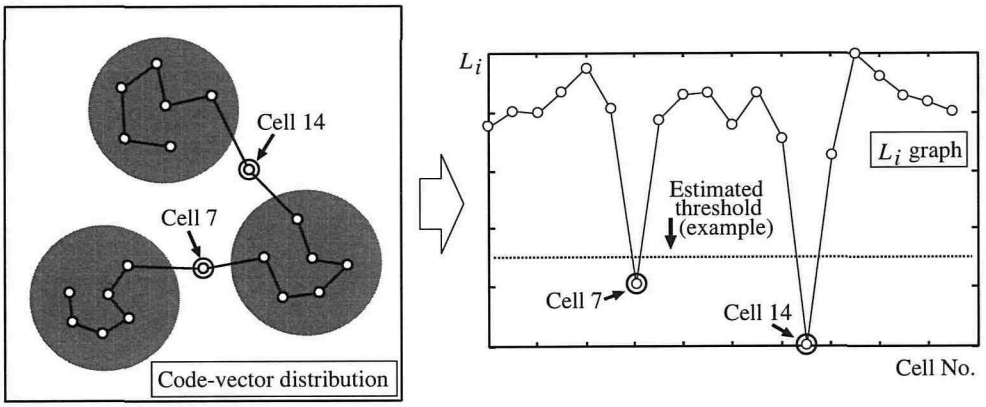
2段階SOMの拡張と実データを対象とした評価実験

本章では，4.2.3節において述べた2段階SOMの問題点について，その原因を分析した上で，2段階SOMのTHSOM過程における学習手法の拡張を行う．この「拡張2段階SOM」に対して，実データを対象としたクラスタリング実験を行ない，拡張2段階SOMの有効性について議論する．

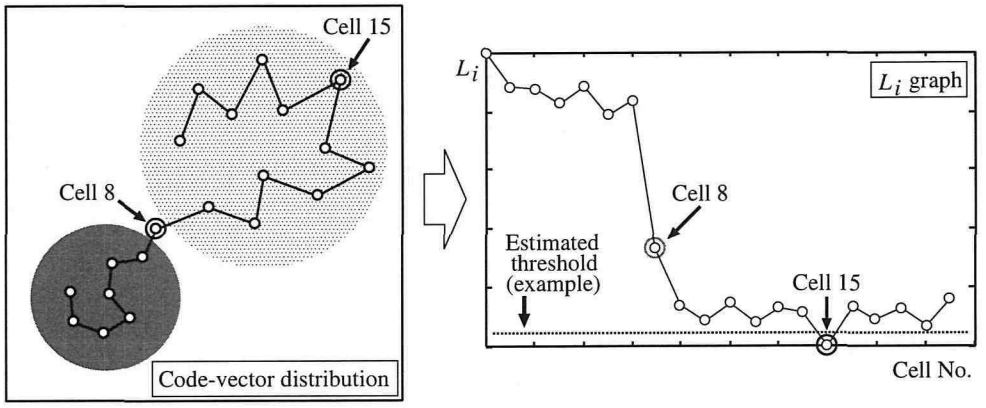
5.1 2段階SOMの問題点と拡張2段階SOMの提案

5.1.1 問題点の定性的な分析

2段階SOMのTHSOM過程では，不活性セルを判別するためのしきい値の尺度として用いられる活性度L_iは，セルiにおける隣接セルとのコードベクトル間距離D_i，および学習時のセルiの勝利回数V_iを，それぞれ$[0,1]$に正規化したV_{N_i}およびD_{N_i}によって，$L_i = V_{\mathrm{N}_i}/D_{\mathrm{N}_i}$と定義した（式(3.7)参照）．また，SOMの学習によって，コードベクトル群は入力データの密度を反映した分布に収束して行くため，D_iおよびV_iの値は，セルiが持つコードベクトルが位置する周辺の入力データの密度に依存している．ここで，個々のクラスタにおけるデータの密度が大きく異なる場合を仮定すると，それぞれのクラスタ内に分布するコードベクトルを持つセルの集団において，D_iやV_i，ひいてはL_iの平均的な大きさが変化すると考えられる．結果的に，L_i値の大小に基づくしきい値設定法では，不



(a) 各クラスタにおけるデータの密度がほぼ同一の場合



(b) 各クラスタにおけるデータの密度が大きく異なる場合

図 5.1 2段階SOMのTHSOM過程に用いられる活性度 L_i の問題点（概念図）

活性セルの判別が正しく行なわれない恐れがある．

図5.1は，このことを図式的に説明するものである．図5.1(a)の左側に示すように，3つのクラスタ内におけるデータの密度がほぼ同一の場合，コードベクトルは各クラスタに同程度の密度で分布する．このようなコードベクトルの分布からは図5.1(a)の右側に示すような L_i のグラフが得られるため，L_i 値の大小に基づいたしきい値設定に問題は生じない．これに対して，図5.1(b)の左側に示すような入力データの分布の場合，密度の低いクラス

タ部分ではコードベクトルの分布もまばらなため，これらのコードベクトルを持つセルに対するL_iが全体的に低い値となってしまう．したがって，図5.1(b)の右側に示すように，L_i値の大小によるしきい値設定では，不活性セルの判別が正しく機能しないと考えられる．

5.1.2　2段階SOMの拡張

L_iの算出元であるD_iおよびV_iは，セルiがクラスタ境界付近に位置するときに，その値が大きく変化すると考えられる．したがって本節では，D_i値およびV_i値の変化量に基づいたしきい値尺度を用いる拡張手法を提案する．具体的には，式(5.1)〜(5.3)に示すように，競合層のセルiにおける$D_{\mathrm{N_}i}$および$V_{\mathrm{N_}i}$それぞれについて，隣接セルとの変化量を個別に求め，これらの積に基づくものとしてL_iの定義を式(5.1)に示すように変更する．この拡張手法を，以降では拡張2段階SOMと呼ぶ．

$$L'_i = dV_i \times dD_i \tag{5.1}$$

$$dD_i = |D_{\mathrm{N\_}i} - D_{\mathrm{N\_}(i-1)}| + |D_{\mathrm{N\_}i} - D_{\mathrm{N\_}(i+1)}| \tag{5.2}$$

$$dV_i = |V_{\mathrm{N\_}i} - V_{\mathrm{N\_}(i-1)}| + |V_{\mathrm{N\_}i} - V_{\mathrm{N\_}(i+1)}| \tag{5.3}$$

$$V_{\mathrm{N\_}i} = \frac{V_i - V_{\min}}{V_{\max} - V_{\min}} \tag{5.4}$$

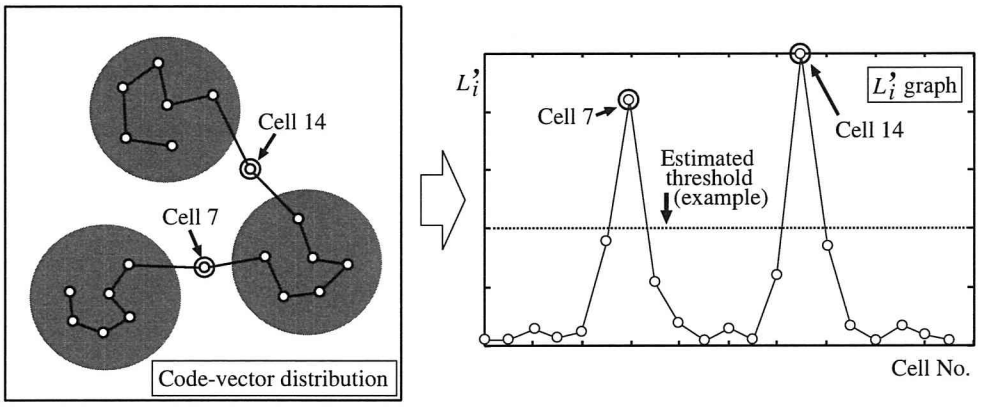
$$D_{\mathrm{N\_}i} = \frac{D_i - D_{\min}}{D_{\max} - D_{\min}} \tag{5.5}$$

L'_iのグラフは，セルiがクラスタ境界付近に位置するときに，上向きのピークを示すと予想される（図5.2(a)右側および図5.2(b)右側参照）．したがって，L'_iはセルiの不活性度を表していると考えることができる．また，この変更により，拡張2段階SOMのTHSOM過程では，式(3.6)が以下のように変更される．

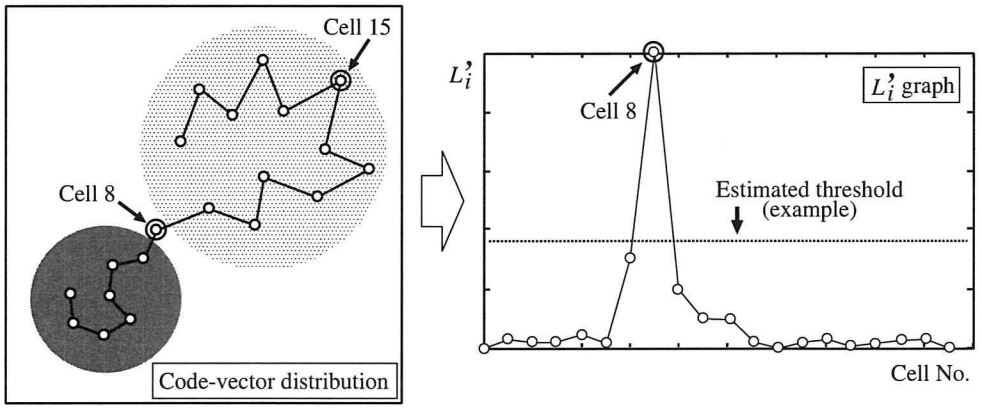
$$F(t,i) = \begin{cases} 1 & \text{if } 0 < t \leq T_{\mathrm{TS1}} \text{ or } (T_{\mathrm{TS1}} < t \leq T_{\mathrm{TS2}} \text{ and } L'_i < L'_{\mathrm{Th}}) \\ 0 & \text{if } T_{\mathrm{TS1}} < t \leq T_{\mathrm{TS2}} \text{ and } L'_i \geq L'_{\mathrm{Th}} \end{cases} \tag{5.6}$$

なお，クラスタごとのデータの密度に大きな差がない場合においても，クラスタ境界ではL'_i値が大きく変化すると考えられる(図5.2(a)右側参照)．このことから，拡張2段階

SOM は，従来型 2 段階 SOM が正常にクラスタ境界を検出できるデータセットに対して，同様にクラスタ境界を検出可能であると考えられる．



(a) 各クラスタにおけるデータの密度がほぼ同一の場合



(b) 各クラスタにおけるデータの密度が大きく異なる場合

図 5.2 拡張 2 段階 SOM におけ不活性度 L_i' のグラフ（概念図）

表 5.1 2段階SOMによるクラスタリング実行時のパラメータ設定（人工データを対象とした場合）

従来型 2 段階 SOM		
	非正規分布型 Type 1	非正規分布型 Type 2
競合層のセル数 N_{cell}	10	45
L_i の しきい値 L_{TH}	0.10	0.00
dW_i' の しきい値 dW'_{TH}	0.35	1.00
拡張 2 段階 SOM		
	非正規分布型 Type 1	非正規分布型 Type 2
競合層のセル数 N_{cell}	10	45
L_i' の しきい値 L'_{TH}	0.15	1.00
dW_i' の しきい値 dW'_{TH}	0.45	1.00

“非正規分布型 Type 1”: クラスタごとの密度が異なるデータセット（図 4.1(c)）
“非正規分布型 Type 2”: 任意の分布形状のクラスタを持つデータセット（図 4.1(d)）

5.2 人工データによる予備実験

5.2.1 実験方法

拡張2段階SOMの効果を検証するため，第4章において用いた非正規分布型データによる予備的なクラスタリング実験を行ない，従来型2段階SOMと拡張2段階SOMにおけるクラスタリング結果を比較する．その際，BSOM過程の学習回数 T_{BS} は入力データ数の200倍とし，THSOM過程については T_{TS1} および T_{TS2} を，それぞれ入力データ数の200倍および400倍とした．さらに，THSOM過程では $\eta = 0.9$ とした．競合層のセル数，THSOM過程における L_i あるいは L_i' のしきい値，ならびに学習後のクラスタ抽出における dW_i' のしきい値については，表5.1にまとめた通りである．

5.2.2 学習後のコードベクトルの分布

まず，5.1節で述べた2段階SOMの拡張により，学習終了後のコードベクトルの分布における従来の2段階SOMの問題点が解消されていることを確認する。

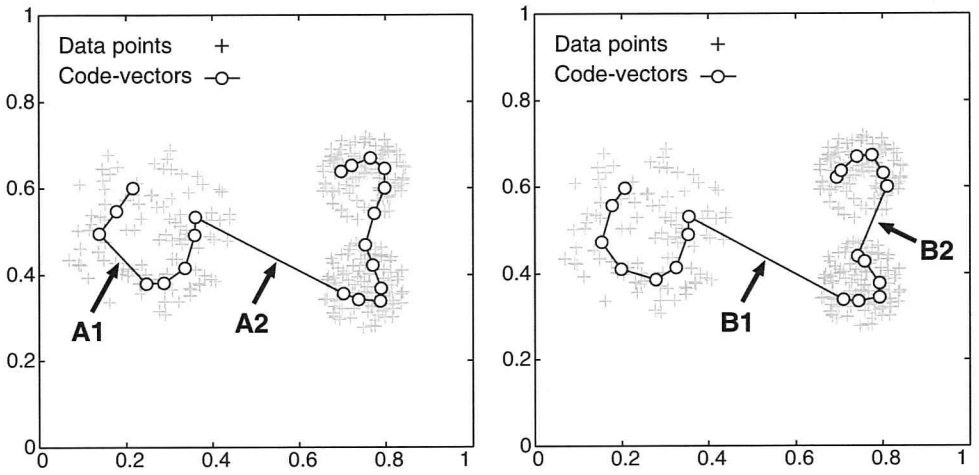
第4章における図4.1(c)のデータセットを，従来の2段階SOM，拡張2段階SOMそれぞれに学習させたときの，学習後のコードベクトルの分布を示したものを図5.3に示す。従来型2段階SOMおよび拡張2段階SOMともに，クラスタ間にコードベクトルの残留は認められないが，従来型2段階SOMでは，密度の低いクラスタ領域の一部（図5.3(a)中の矢印A1）がクラスタ境界であるとみなされ，該当部分でコードベクトル同士が互いに引き離されてしまっている。これは，5.1節で述べた理由により，しきい値設定による不活性セルの推定が正しく行われなかったためであると考えられる。これに対して拡張2段階SOMでは，クラスタごとの密度の違いに影響されることなく，図5.3(b)中の矢印B1およびB2で示すように，各クラスタの境界部分でコードベクトル同士が引き離されていることが分かる。

また，図4.1(d)のように，クラスタ形状は歪曲しているが，クラスタ毎のデータ密度がほぼ等しいようなデータに対しては，図5.4(a)および図5.4(b)に示すように，従来型2段階SOMおよび拡張2段階SOMともに，クラスタの境界部分でコードベクトル同士が大きく引き離され，良好な学習結果を示していることが確認できる。

5.2.3 クラスタリング実験結果

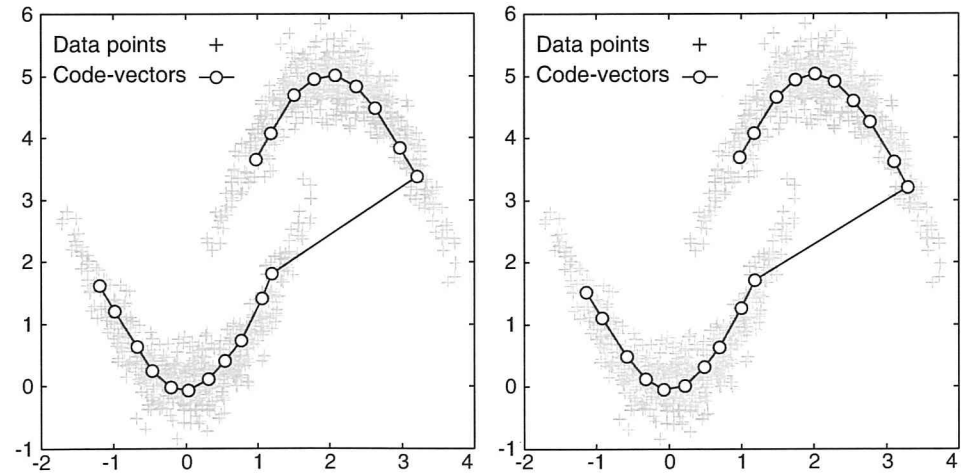
表5.2は，それぞれの入力データに対する各手法の誤分類率をまとめたものである．k-means法および2段階SOMにおける誤分類率は，100試行の平均値であり，括弧内の数字は，100試行のうち，誤分類率が20%を超えたため，正しいクラスタ分割が得られていないと判断された回数である．なお，k-means法および，階層的クラスタリングの3手法における誤分類率は，比較のために表4.5および表4.6の該当部分を再掲している．

密度の異なる入力データ（“Type 1”）に対して，拡張2段階SOMは100試行すべてにおいて正しいクラスタ分割が得られている．そのため，従来型2段階SOMやk-means法と比較して，誤分類率が大きく減少しており，最長距離法やウォード法と比較してもほぼ同等の誤分類率を示していることが確認できる．次に，任意の分布形状のクラスタを持つ



(a) 従来型 2 段階 SOM (b) 拡張 2 段階 SOM

図 5.3 図 4.1(c) のデータセットに対する，学習後のコードベクトルの分布



(a) 従来型 2 段階 SOM (b) 拡張 2 段階 SOM

図 5.4 図 4.1(d) のデータセットに対する，学習後のコードベクトルの分布

表 5.2 人工データを用いた従来型および拡張2段階SOMのクラスタリング実験結果

	k-means 法	従来型 2段階 SOM	拡張 2段階 SOM	最短距離法	最長距離法	ウォード法
Dataset Type 1	24.6 (48)	14.4 (27)	**0.38** (0)	33.7	0	0
Dataset Type 2	2.4 (0)	0.38 (0)	**0.37** (0)	0	3.2	3.7

"非正規分布型 Type 1": クラスタごとの密度が異なるデータセット（図 4.1(c)）
"非正規分布型 Type 2": 任意の分布形状のクラスタを持つデータセット（図 4.1(d)）

入力データ（"Type 2"）に対しては，従来型2段階SOMと拡張2段階SOMともに，ほぼ同程度の誤分類率を示している．

以上のことから，従来型2段階SOMがクラスタリングを正しく行なえないデータに対して，拡張2段階SOMはクラスタリング結果が大きく改善できること，ならびに，従来型2段階SOMが正しくクラスタリングを行うデータに対しては，拡張2段階SOMも同様のクラスタリング結果を示すことが確認できた．

5.3 実データによるクラスタリング実験

拡張2段階SOMによるクラスタリングの有効性をさらに確認するため，実データを用いた性能評価を行う．対象として用いる実データには様々なものが考えられるが，評価実験の追認性を考慮して，本論文では機械学習アルゴリズムの評価用データセットである，UCI Machine Learning データベース [27] を利用した．

5.3.1 UCI Machine Learning データベース

UCI Machine Learning (UCI ML) データベースは，カリフォルニア大学アーバイン校が公開しているもので，その目的は，人工知能の研究課題の一つである機械学習（Machine Learning）において提案される様々な学習アルゴリズムに対して，共通のサンプルデータによって性能評価を行うことである．機械学習の応用分野は多岐にわたるため，データベー

スには数量データ，カテゴリカルデータ，あるいはそれらを混合したものなど，非常に多くのデータセットが収録されている．これらのデータセットは，いずれも実環境から採取されたデータであるため，新たに提案された機械学習アルゴリズムの応用可能性などを評価するのに適している．

本論文では，クラスタ数やデータの分布などを考慮して，UCI ML データベースの中から，Iris，Breast-cancer Wisconsin（以下 BCW），Vowel-context（以下 Vowel），Thyroid gland（以下 Thyroid）の 4 つのデータセットをクラスタリング対象として使用した．これらのデータセットにおける，サンプル数，クラス数，サンプルデータの次元数を表 5.3 に示す．また，各データセットの概要については以下の通りである．

Iris ･･･ Edgar Anderson によって採取されたアヤメの分類データである．Fisher の文献 [32] において，判別分析のサンプルデータとして用いられた後，現在に至るまでベンチマーク用データとしてよく用いられている．3 種類のアヤメ（Setosa，Virginica，Versicolor）それぞれ 50 サンプルを，がく片や花弁の長さなど 4 つの属性で表わしている．

BCW ･･･ ウィスコンシン大学によって収集された，乳がんの検診データである．699 個の腫瘍の組織サンプルを，細胞集団の凝集度や，細胞のサイズあるいは形状の一様性など，10 種類の属性で表わしており，個々のサンプルに対して陰性（Benign）あるいは陽性（Malignant）のラベルが付加されている．本論文では，欠損値を持つサンプルを除いた 683 個のデータを使用する．

Vowel ･･･ 英語を母国語とする 15 名の話者が，10 種類の母音をそれぞれ 6 回ずつ発声したものを録音し，録音された母音の音声を，線形予測分析によって 10 次元の特徴量で表わしたものである．本論文では，今回の実験に合わせて男性話者 8 名による 3 母音（/i/，/A/，/O/）のデータを抽出し，データセットを再構築した．

Thyroid ･･･ 甲状腺（Thyroid gland）の状態を示す検診データを集めたもので，個々の

表 5.3 UCIデータベースから引用した各データセットの諸元

	Iris	BCW	Vowel※	Thyroid
サンプル数	150	683	126	215
属性数（次元数）	4	10	10	5
クラス数	3	2	3	3

※ 3母音のデータを抽出して使用

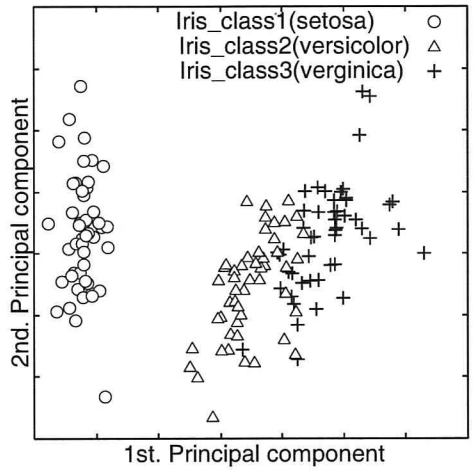
サンプルは，放射線免疫測定法によって計測された甲状腺刺激ホルモンの分泌量などを示す情報で構成されている．これらのサンプル群は，甲状腺の状態によって，euthyroidism (normal)，hypothyroidism，および hyperthyroidism の3グループに分類されている．

ここで，各データセットの分布を可視化するために，それぞれに対して主成分分析を施し，得られた第1主成分および第2主成分の固有ベクトルで構成される2次元空間にデータをプロットしたもの，および，第2主成分までの累積寄与率を図5.5に示している．Irisは verginica と versicolor の2クラスタが非常に接近しており，BCWやThyroidは各クラスタにおけるデータの密度が大きく異なっている (図5.5(b) および図5.5(c) 参照)．また，Vowelは2つのクラスタ “/i/” および “/O/” の分布が大きく歪曲していることが分かる (図5.5(d) 参照)．

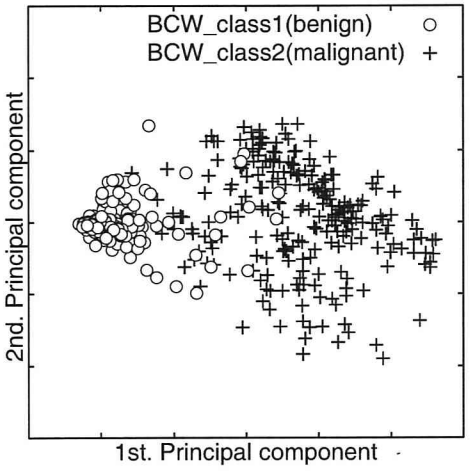
5.3.2 実験方法

これらの対象データを，k-means法，従来型および拡張2段階SOM，そして3種類の階層的クラスタリング手法（最短距離法，最長距離法，ウォード法）それぞれを用いてクラスタリングし，各手法における誤分類率を比較した．

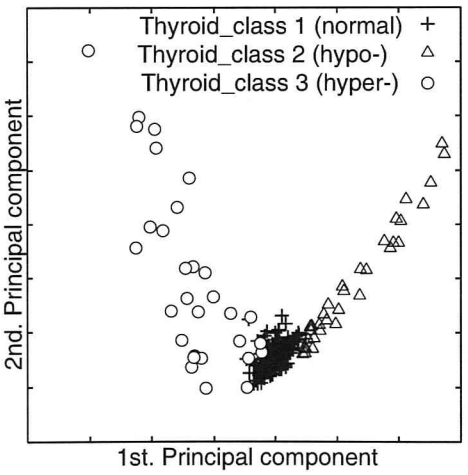
k-means法および，階層的クラスタリングの各手法における実行時の設定については，第4章と同様である．また，従来型および拡張2段階SOMにおける，学習回数およびTHSOM過程でのηの値については，5.2節と同様とした．競合層のセル数，THSOM過程における



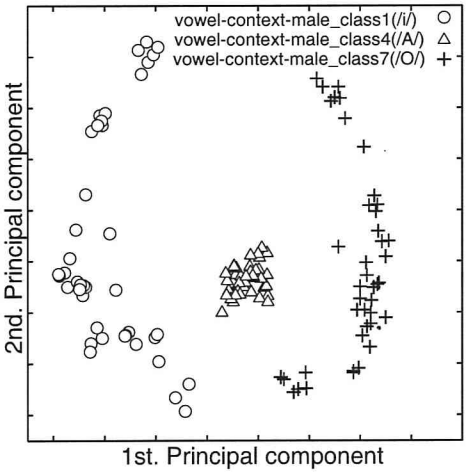
(a) Iris(第 2 主成分までの累積寄与率 = 0.958)



(b) BCW(第 2 主成分までの累積寄与率 = 0.762)



(c) Thyroid(第 2 主成分までの累積寄与率 = 0.741)



(d) Vowel(第 2 主成分までの累積寄与率 = 0.708)

図 5.5 主成分分析によって可視化した UCI の各データセットの分布

表5.4 2段階SOMによるクラスタリング実行時のパラメータ設定（UCIデータを対象とした場合）

従来型2段階SOM				
	BCW	Iris	Vowel	Thyroid
競合層のセル数 N_{cell}	10	35	20	10
L_i の しきい値 L_{TH}	0.0	0.12	0.004	0.007
dW_i' の しきい値 dW'_{TH}	1.0	0.27	0.81	0.89
拡張2段階SOM				
	BCW	Iris	Vowel	Thyroid
競合層のセル数 N_{cell}	10	35	20	10
L_i' の しきい値 L'_{TH}	1.0	0.19	0.85	0.81
dW_i' の しきい値 dW'_{TH}	1.0	0.28	0.81	0.76

L_i あるいは L_i' のしきい値，ならびに学習後のクラスタ抽出における dW_i' のしきい値については，表5.4にまとめた通りである．

5.3.3 実験結果および考察

表5.5は，それぞれのデータセットに対する各クラスタリング手法の誤分類率をまとめたものである．k-means法および2段階SOMにおける誤分類率の値は100試行の平均値であり，括弧内の数字は，100試行のうち，誤分類率が20%を超えているため正しいクラスタ分割が得られなかったと考えられる回数である．また，拡張2段階SOMにおける誤分類率の数値の中で太字で示されているものは，従来型2段階SOMと比較して，誤分類率が有意に低下（有意水準5%）したと認められたものである．

BCW，Thyroidデータセットのように，それぞれのクラスタにおいてデータの分布密度が異なっているような場合，従来型2段階SOMと比較して，拡張2段階SOMの誤分類率が有意に低下していることが示されており，Thyroidに関してはk-means法に対する改善も認められる．一方，Iris，Vowelデータセットに関しては，従来型2段階SOMと拡張2段階SOMにおける誤分類率に有意差があるとはいえず，いずれもk-means法に対して良好な結果を示した．このことから，従来型2段階SOMによって良好なクラスタリングを行な

表 5.5 UCI データセットに対する各クラスタリング手法の誤分類率 $P_{\mathrm{Err}}(\%)$ の比較

	k-means 法	従来型 2 段階 SOM	拡張 2 段階 SOM	最短距離法	最長距離法	ウォード法
BCW	3.83 (0)	3.91 (0)	**3.67** (0)	34.0	9.44	3.15
Thyroid	16.5 (17)	17.8 (0)	**13.7** (0)	29.8	20.9	12.6
Iris	14.0 (8)	10.5 (9)	**9.59** (0)	32.0	16.0	10.7
Vowel	20.8 (42)	0 (0)	0.10 (0)	0	42.9	42.9

えるようなデータに対しては，拡張 2 段階 SOM によっても同等のクラスタリング結果が得られると考えられる．

次に，最短距離法，最長距離法，ウォード法それぞれにおける各データセットの誤分類率を比較すると，Vowel データのようにクラスタの分布が歪曲している場合には最短距離法が最も良く，その他のデータセットについてはウォード法が最も良好であるといえる．これに対して 2 段階 SOM では，いずれのデータセットに対しても，他の階層的クラスタリング手法と比較して同等以上のクラスタリング性能を示している．

これらの実験結果は，5.2.3 節で述べた，人工データを用いた場合の実験結果における所見と一致しており，したがって，階層的クラスタリング手法や k-means 法，従来型 2 段階 SOM と比較して，拡張 2 段階 SOM を用いたクラスタリングの有効性が確認できたと考えられる．

5.4 クラスタリングの安定性

ここでは，表 5.2 および表 5.5 に示した誤分類率とは別の観点として，競合層のセル数を変化させたときの，クラスタリング性能の安定性について述べる．一般に，SOM を用いたクラスタリングでは，対象データに対して最適と思われる競合層セル数を経験的に設定し

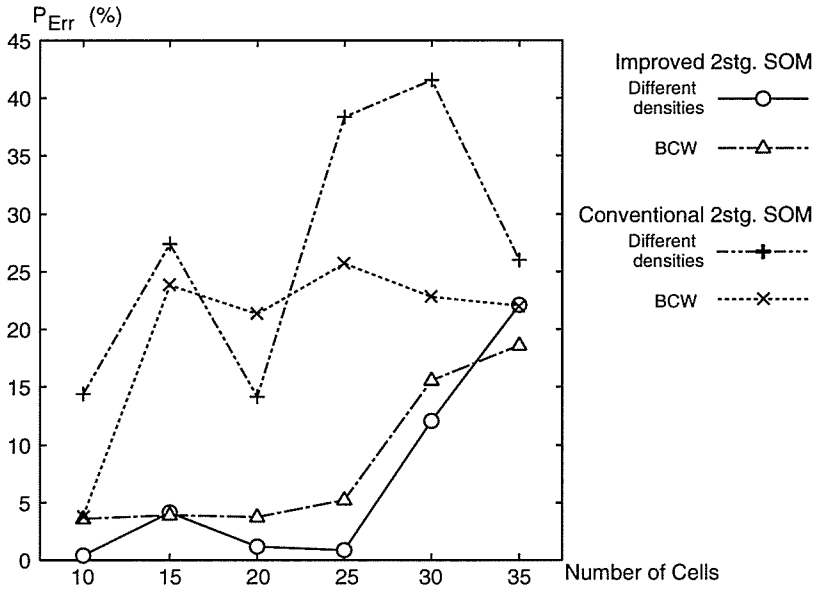


図 5.6 BCW データおよび密度の異なる人工データに対するセル数と誤分類率の関係

なければならない．そのため，セル数の変動に対するクラスタリング結果の安定性は重要な要素である．

図5.6は，クラスタ内のデータの密度が大きく異なる，BCW データと図4.1(c)に示す人工データをクラスタリングしたときの，従来型2段階SOMと拡張2段階SOM双方における，競合層のセル数と誤分類率の関係を示している．従来型2段階SOMは，セル数の増加にともなって，クラスタリング性能が急激に悪化する傾向が見られる．これに対して，拡張2段階SOMでは，ある程度のセル数の変動に対しても，誤分類率が低く抑えられている．このことから，拡張2段階SOMにおけるクラスタリングの安定性が確認できたと考えられる．

5.5 本章のまとめ

本章では，前章において示された2段階SOMの問題点について検討し，その原因について定性的な分析を行った．その結果，2段階SOMのTHSOM過程において，競合層セ

ルのコードベクトル間距離に基づいた，各セルにおける活性度算出法の問題点を指摘した．この問題点を解決するために，セル単位のコードベクトル間距離ではなく，隣接セル間のコードベクトル間距離の変化量に基づいた，各セルの不活性度を定義し，これをもとに拡張2段階SOMの提案を行った．

拡張2段階SOMに対して，人工データおよびUCI MLデータベースから引用した実データを用いた性能評価実験を行ない，拡張2段階SOMでは，従来型2段階SOMでは困難であった，密度の異なるデータの正確なクラスタリング実現できることを示した．さらに，UCI MLデータベースのIris，Breast-cancer wisconsin，Vowel，Thyroid glandデータセットによるクラスタリング実験を行ない，階層型クラスタリング手法におけるクラスタリングの正確さが大きく異なるようなデータセットに対しても，拡張2段階SOMでは，いずれも良好なクラスタリング結果が得られることを示し，拡張2段階SOMの有効性を確認した．

第6章

結言

6.1 本論文のまとめ

本論文では，自己組織化マップ（SOM）のクラスタリング問題への適用に関して，そのクラスタリング性能の向上を目的とした学習アルゴリズムの改善を行った．各章における成果を以下にまとめる．

第2章では，従来の代表的なクラスタリング手法について概説し，非階層的手法であるk-means法は代表ベクトルの初期状態によってクラスタリング結果にばらつきがあることや，個々のクラスタを1つの代表点だけで近似するために任意形状のクラスタの抽出が困難であることを指摘した．また，最短距離法，最長距離法，ウォード法などに代表される階層的手法は，手法によってクラスタリング結果が大きく異なる場合があり，対象データに適した手法を選択するには経験に基づく試行錯誤が必要であることや，対象データの大規模化への対応が困難であることを指摘した．以上をふまえた上で，SOMおよびSOMを用いたクラスタリング手法の概要とその特長を述べ，クラスタリング問題にSOMを適用することで，従来のクラスタリング手法の問題点を解決できる可能性を示した．

第3章では，SOMの基本学習アルゴリズムにおいて，学習後に生じる不活性セルが，学習後のクラスタ抽出の際にクラスタ境界の正確な推定を阻害することを示し，SOMをクラスタリング問題に適用する上での問題点を明らかにした．クラスタリング問題への適用に

おけるSOMの学習アルゴリズムの問題点を解決するために，従来の基本学習アルゴリズム（BSOM）と，BSOMにおける近傍関数にしきい値作用を取り入れた，しきい値SOM（THSOM）とを段階的に適用する2段階SOMを提案した．その際，THSOMの学習法を再検討し，学習中のセルの勝利回数と学習後のコードベクトル間距離を組合わせることによって，競合層の各セルに対する活性度を定義し，不活性セル判定の正確化を図った．この2段階SOMの学習アルゴリズムに対して基礎的な学習実験を行った結果，2段階SOMでは，SOMにおける位相保持写像の性質を維持しつつ，不活性セルの発生を抑制することが可能であることを確認した．

第4章では，2段階SOMのクラスタリング性能を定量的に評価するために，人工的に作成したデータセットを用いてクラスタリング実験を行った．等方的な分散を持つ正規分布型データセットのクラスタ抽出において，k-means法ではクラスタ数の増加によってクラスタ抽出に失敗するケースが増えるのに対し，2段階SOMでは正確なクラスタ抽出を安定的に行えることを示した．また，クラスタごとのデータ密度が異なっていたり，任意形状のクラスタで構成されるような非正規分布型データセットのクラスタ抽出においては，階層的クラスタリング手法（最短距離法，最長距離法，ウォード法）では，抽出対象のクラスタの特徴によって適用すべき手法が異なること，2段階SOMではその傾向が緩和されることを確認した．一方で，クラスタごとのデータ密度が異なるようなデータセットに対しては，k-means法と同様に，クラスタの境界を正しく推定できない場合が多く発生することを明らかにした．

第5章では，2段階SOMのTHSOM過程における，各セルの活性度の算出方法に改良を加えた，拡張2段階SOMの提案を行った．従来型2段階SOMでは，セルの活性度をコードベクトル間距離に基づいて算出していたため，クラスタ毎にデータ密度が異なるような場合に，各セルの活性度のグラフから不活性セルを正確に検出できないという問題があった．これに対して拡張2段階SOMでは，隣接セル間のコードベクトル間距離の変化量に注目してセルの活性度を定義することにより，人工的な評価用データセットを用いた学習実

験において，データ密度の異なるクラスタで構成されるデータセットに対して，不活性セルの検出が良好に行えることを示した．さらに，UCI ML データベースの実データを用いたクラスタリング実験を行ない，従来のクラスタリング手法および従来型2段階SOMを用いたクラスタリング手法に対して，誤分類率および競合層セル数の変動に対する安定性の面から，拡張2段階SOMの有効性を確認した．

これら一連の研究成果から，本研究の目的である，自己組織化マップを用いたクラスタリングの高性能化については，十分に達成できたと考えられる．

6.2 今後の課題

本論文では，主にクラスタリング時の誤分類率によって，提案手法を含め，種々の従来手法におけるクラスタリング性能を評価している．しかしながら，大規模なデータへの適用を考えた場合，誤分類率が低いことはもちろん重要であるが，それと同時に，クラスタリングに要する計算時間も，クラスタリング性能を示す重要な指標となる．したがって，時間的コストの観点から，従来の階層的クラスタリング手法と，2段階SOMを用いた手法の場合の計算時間について，理論的および定量的な評価を行なわなければならない．

また，本論文で提案した2段階SOMによるクラスタリング手法は，従来型2段階SOM，拡張2段階SOMともに，最終的にはSOMのコードベクトル間の距離に基づいてクラスタの抽出を行う．そのため，望ましいクラスタ分割を得るための，しきい値設定が非常に重要であり，最適なしきい値を自動設定するための工夫，あるいは，データ間の距離に基づかないクラスタ抽出法の提案が必要と思われる．これについては，k-means 法において情報量規準を組合わせた手法が提案されており [33]，本研究においても，学習後のSOMに対して，赤池情報量規準 AIC[34] あるいはベイズ型情報量規準 BIC[35] を用いたクラスタ抽出法について，すでに検討を始めている [36][37]．今後，このアプローチに対して詳細な研究を進め，SOMを用いたクラスタリングの更なる性能向上を行うことが課題である．

謝辞

本研究は，鳥取大学工学部 伊藤 良生 教授のご指導のもとに行われました．本稿を終えるにあたり，終始，熱心なご指導を賜った同教授に心より感謝申し上げます．また，不甲斐ない私を常に暖かく見守って頂き，励ましてくださったこと，厚く御礼申し上げます．

本研究をまとめるにあたり，ご指導頂いた，鳥取大学工学部 李 仕剛 教授，近藤克哉 教授に深く感謝申し上げます．

本研究を行うにあたり，様々な面でご援助，ご指導を頂いた鳥取大学工学部 笹岡直人 助教，立木純夫 技術専門職員，ならびに鳥取大学大学院工学研究科博士後期課程情報生産工学専攻 教員各位に心より感謝致します．

社会人学生として本研究を遂行するにあたり，松江工業高等専門学校 前校長 宮本 武明 博士，現校長 荒木 光彦 博士には多大なご配慮を賜りました．厚く御礼申し上げます．さらに，終始暖かいご支援とご配慮を賜りました松江工業高等専門学校 情報工学科 教員の皆様をはじめとする，教職員各位に厚く御礼申し上げます．また，本研究を進めるにあたって，多大なご協力を頂きました， 松江工業高等専門学校 堀内 匡 准教授，原 元司 准教授に深く感謝致します．

最後に私事ながら，博士課程修了まで様々な援助や励ましを頂いた，両親，弟，及び友人一同に心より感謝致します．

2009 年 1 月

参考文献

[1] Williams, W.T. and Lance, G.N.: "Hierarchical classificatory methods", in *Statistical Methods for Digital Computers*, Enslein, K., Ralson, A. and Wilf, H.S. (ed.), Chap.11, John Wiley & Sons, (1977)

[2] A.K. Jain, M.N. Murty and P.J. Flynn: "Data Clustering: A Review", ACM Computer Surveys, Vol.31, No.3, (1999)

[3] 鳥脇純一郎: 「認識工学 − パターン認識とその応用」, コロナ社, (1993)

[4] Pang-Ning Tan, Michael Steinbach and Vipin Kumar: "Data Mining: Introduction To", Addison-Wesley, (2005)

[5] T. Kamishima: 神嶌敏弘: 「データマイニング分野のクラスタリング手法 (1) −クラスタリングを使ってみよう!−」, 人工知能学会誌, Vol.18, No.1, pp.59–65, (2003–1)

[6] McQueen, J.: "Some moethods for classification and analysis of multivariate observations", in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, (1967)

[7] 麻生英樹, 津田宏治, 村田昇: 「パターン認識と学習の統計学」, 岩波書店, (2003)

[8] 森健一: 「パターン認識」, 電子情報通信学会, (1988)

[9] Sneath, P.H.A. and Sokal, R.R.: *Numerical Taxonomy*, Freeman, London, UK. (1973)

[10] King, B.: "Step-wise clustering procedures", *J. Am. Stat. Assoc.* Vol. 69, pp. 86–101, (1967)

[11] Ward, J.H.Jr.: "Hierarchical grouping to optimize an objective function", *J. Am. Stat. Assoc.*, Vol. 58, pp. 236–244, (1963)

[12] T. Kohonen: "Self-Organizing Maps", Springer-Verlag Verlin Heidelverg, (1995)

[13] 徳高平蔵, 藤村喜久郎, 岸田 悟: 「自己組織化マップの応用 − 多次元情報の2次元可視化」, 海文堂, (1999)

[14] 徳高平蔵, 山川 烈, 藤村喜久郎: 「自己組織化マップ応用事例集 − SOMによる可視化情報処理」, 海文堂, (2002)

[15] 寺島幹彦, 白谷文行, 山本公明: 「自己組織化特徴マップ上のデータ密度ヒストグラムを用いた教師なしクラスタ分類法」, 電子情報通信学会論文誌, Vol.J79-D-II, No.7, pp.1280–1290, (1996–7)

[16] 田中雅博, 古河靖之, 谷野哲三: 「自己組織化マップを利用したクラスタリング」, 電子情報通信学会論文誌, Vol.J79-D-II, No.2, pp.301–304, (1996–2)

[17] 青木宏樹, 斉藤利通: 「しきい値作用を有する自己組織化写像の分類機能について」, 電子情報通信学会論文誌, Vol.J83-A, No.9, pp.1122–1124, (2000–9)

[18] E. Uchino, M. Kawamura and K. Nagata: "Dynamic Deletion of Units for Self-Organizing Map by Introducing a New Measure of Unit's Contribution to Learning" *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, Vol.14, No.2, pp.157–164, (2002–4)

[19] Marc M. van Hulle: "Faithful Representations and Topographic Maps: From Distortion to Information-Based Self-Organization" Wiley-Interscience (2000)

[20] K. Koike, S. Kato and T. Horiuchi: "A Two-stage Self-Organizing Map with Threshold Operation for Data Classification", *Proceedings of the 2002 SICE Annual Conference*, pp.2900-2902, (2002)

[21] 加藤 聡, 小池 健太, 堀内 匡: 「2 段階 SOM の提案とそのクラスタリング問題への適用」, 電気学会論文誌 C(電子・情報・システム部門誌), Vol.125, No.1, pp.14–20, (2005–1)

[22] 齋藤堯幸, 宿久 洋: 「関連性データの解析法 − 多次元尺度構成法とクラスター分析法」, 共立出版, (2006)

[23] 加藤 聡, 堀内 匡, 伊藤 良生: 「2 段階 SOM を用いた階層的クラスタリングに関する基礎的考察」, 電子情報通信学会スマートインフォメディアシステム研究会 技術研究報告, SIS2005-64, pp.27–31, (2006–3)

[24] 加藤 聡, 堀内 匡, 伊藤 良生: 「2 段階 SOM によるクラスタリングの性能評価と改良手法に関する研究」, 第 8 回自己組織化マップ研究会 2007 講演論文集, pp.9–12, (2007–3)

[25] 加藤 聡, 堀内 匡, 伊藤 良生: 「改良型 2 段階 SOM によるクラスタリングの性能評価」, 第 23 回ファジィシステムシンポジウム講演論文集, pp.665–670, (2007–9)

[26] 加藤 聡, 堀内 匡, 伊藤 良生: 「2 段階 SOM の拡張とそのクラスタリング性能の評価」, *RISP Journal of Signal Processing*, Vol.13, No.1, pp.77-85, (2009–1)

[27] UCI Machine Learning Repository:
http://www.ics.uci.edu/ mlearn/MLRepository.html

[28] Kurita, T.: "An efficient agglomerative clustering algorithm using a heap", *Pattern Recognition*, Vol. 24, No. 3, pp. 205–209, (1991)

[29] Zahn, C.T.: "Graph-theoretical methods for detecting and describing gestalt clusters", *IEEE Trans. Comput. C-20(Apr.)*, pp. 68–86, (1971)

[30] Gower, J.C. and Ross, G.J.S.: "Minimum spanning trees and single-linkage cluster analysis", *Appl. Stat.*, Vol. 18, pp. 54–64, (1969)

[31] Backer, F.B. and Hubert, L.J.: "A graph theoretic approach to goodness-of-fit in complete-link hierarchical clustering", *J. Am. Stat. Assoc.*, Vol. 71, pp.870–878, (1976)

[32] Fisher, R.A.: "The use of multiple measurements in taxonomic problems", *Annual Eugenics*, Vol. 7, Part II, pp. 179–188, (1936)

[33] Pelleg, D. and Moore, A.："X-means: Extending K-means with Efficient Estimation of the Number of Clusters", *Proc. of the 17th International Conference on Machine Learning (ICML-2000)*, pp.727–734, (2000)

[34] 赤池弘次，甘利俊一，北川源四郎，樺島祥介，下平英寿：「赤池情報量規準 AIC」，共立出版，2007.

[35] 小西貞則，北川源四郎：「情報量規準」，朝倉書店，(2004)

[36] 加藤 聡，堀内 匡，伊藤 良生: 「自己組織化マップによるクラスタリングへの情報量規準の適用に関する実験的検討」，平成 19 年度 電気・情報関連学会中国支部第 58 回連合大会 講演論文集, pp.226–227, (2007–10)

[37] 加藤 聡，堀内 匡，伊藤 良生: 「自己組織化マップと情報量規準を用いたクラスタ抽出法に関する研究」，第 24 回ファジィシステムシンポジウム講演論文集, pp.723–726, (2008–9)

研究業績

1. 学術雑誌発表論文

	著者・論文題目・発表機関	本文
[1]	加藤 聡, 小池 健太, 堀内 匡, “2段階SOMの提案とそのクラスタリング問題への適用, ” 電気学会論文誌C, Vol.125, No.1, pp.14–20, 2005年1月.	第3章
[2]	加藤 聡, 堀内 匡, 伊藤 良生, “2段階SOMの拡張とそのクラスタリング性能の評価, ”Journal of Signal Processing, Vol.13, No.1, pp.77–85, 2009年1月.	第4章 第5章

2. 国際会議発表論文

	著者・論文題目・発表機関	本文
[1]	K. Koike, S. Kato, T. Horiuchi, " A Study on Clustering using Two-stage Self-Organizing Map with Threshold Operation," Proceedings of the International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC2003), Vol.1, pp.581-584, July 2003.	第3章
[2]	S. Kato, T. Horiuchi, Y. Itoh, " A Study on Two-Stage Self-Organizing Map Suitable for Clustering Problems," 2006 IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS'06), pp.677-680, December 2006.	第3章

3. 学会研究会，紀要発表論文

	著者・論文題目・発表機関	本文
[1]	加藤 聡, 堀内 匡, 伊藤 良生, "改良型2段階SOMによるクラスタリングの性能評価," 第23回ファジィシステムシンポジウム講演論文集, pp.665-670, 2007年9月.	第5章
[2]	加藤 聡, 堀内 匡, 伊藤 良生, "2段階SOMによるクラスタリングの性能評価と改良手法に関する研究," 第8回自己組織化マップ研究会2007講演論文集, pp.9-12, 2007年3月.	第5章
[3]	加藤 聡, 堀内 匡, 伊藤 良生, "2段階SOMによる階層的クラスタリングに関する基礎的考察," 信学技報, SIS2005-64, pp.27-31, 2006年3月.	第4章
[4]	加藤 聡, 堀内 匡, 伊藤 良生, " 自己組織化マップを用いたクラスタリングの評価に関する基礎的考察," 計測自動制御学会中国支部 学術講演会論文集, pp.170-171, 2005年11月.	第4章
[5]	加藤 聡, 堀内 匡, 伊藤 良生, "クラスタリング問題に対する2段階SOMの適用に関する評価," 電気学会 電子・情報・システム部門大会講演論文集, pp.1107-1112, 2005年9月.	第4章
[6]	加藤 聡, 堀内 匡, 小池 健太, "2段階SOMを用いたクラスタリングの実験的評価," 第6回自己組織化マップ研究会2005講演論文集, pp.19-23, 2005年3月.	第4章
[8]	加藤 聡, 小池 健太, 堀内 匡, "2段階SOMを用いたクラスタリングに関する基礎的研究," 松江工業高等専門学校研究紀要, 第39号, pp.31-36, 2004年2月.	第3章 第2章

4. その他

平成17年度電気学会 電子・情報・システム部門 「次世代を担う若手技術者・研究者」特集 優秀論文賞

以下の一編に対しての受賞.

加藤 聡, 小池 健太, 堀内 匡, "2段階SOMの提案とそのクラスタリング問題への適用," 電気学会論文誌 C, Vol. 125, No.1, pp.14–20, 2005年1月.

END