論　文

最小全域木による大規模パターンの分布解析

後藤　雅典[†a)]　石田　良介[††]　蔡　文杰[†††]　内田　誠一[††††]

Distribution Analysis of a Large-Scale Pattern Set Using Minimum Spanning Tree

Masanori GOTO[†a)], Ryosuke ISHIDA[††], Wenjie CAI[†††], and Seiichi UCHIDA[††††]

あらまし　本研究の究極の目標は,「パターンの真の分布」を解明することである．その際．単一のクラスの分布だけでなく，複数クラス間の関係も解明の対象とする．パターン空間に存在しうる全てのパターンを収集するのは実現不可能なため，できる限り多くのクラスラベル付きパターンを収集した上で，その分布構造を解析することで，この目標に挑む．真の分布の解明を目指す以上，解析手法として，何らかのモデルによる近似や低次元化など，パターン間の近傍関係に誤差が入り得るものは適切でない．そこで本論文では，パターンの相対位置関係を誤差なく保存しうるネットワーク解析手法により大規模パターンの分布構造の解析を行う．具体的には，各パターンを 1 ノードとし，その近傍関係によりエッジを付与してネットワークを構成し，その構造を解析する．本論文では，ネットワークの作成手法として最小全域木を適用し，分布解析の対象として約 50 万個の活字数字画像と約 80 万個の手書き数字画像を用いた実施例を示し，パターン数の増加によるパターン分布の変化を明らかにする．

キーワード　文字認識，大規模パターン認識，ネットワーク解析，最小全域木，分布解析

1. まえがき

本研究の究極の目標は,「パターンの真の分布」を解明することである．言い換えれば，存在しうる全てのパターンの成す分布を明らかにすることである．これが達成できれば，アウトライアの分布，複数クラス分布間のオーバーラップの様子，連結性などの位相構造，分布の正規性や多峰性など，これまで経験的に論じられてきた事項を具体的に吟味することが可能となる．そして，パターン認識にとってより強固な理論的基盤を与えることができる．

本研究では，大量の画像から構成される大規模パターンを用いてパターンの真の分布を解明することを目指す．後述のように，近年の計算機の進歩によって解析が可能となった大規模パターンを用いた研究は多くの成果をあげている [1]〜[4]．しかしながら，これらの研究の多くは大規模パターンによる認識率の改善を主な目標としている．これに対して本研究では，パターンの真の分布を解明するという視点から大規模パターンの分布構造の解析を行う．

パターンの真の分布を解明するためには次の二つの条件を満たすことが重要である．まず，できる限り多くのクラスラベル付きパターンを収集し，解析対象のパターンを大規模化することである．次に，解析による誤差を生じさせない手法を用いて大規模パターンの分布構造を明らかにすることである．何らかのモデルによる近似や低次元化などでは，解析結果に誤差が生じてしまうため，誤差のない解析手法が必要とされる．

第一の条件であるパターンの大規模化については，解析を行うパターン空間の広さと密接に関係する．もし，パターン空間に存在しうる全てのパターンを収集でき，それらのパターンに正しいクラスラベルを付与できれば，パターンの真の分布は明白なものとなるであろう．しかし，画像のパターン空間は極めて広く，存在しうる全てのパターンを収集することは一般的

† グローリー株式会社研究開発センター，姫路市
Research & Development Center, GLORY LTD., 1–3–1 Shimoteno, Himeji-shi, 670–8567 Japan
†† 九州大学大学院システム情報科学府，福岡市
Graduate School of Information Science and Electrical Engineering, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka-shi, 819–0395 Japan
††† (株) オーリッド，別府市
O–RID CO., LTD., 1–2–2 Minami Tateishi, Beppu-shi, 874–0839 Japan
†††† 九州大学大学院システム情報科学研究院，福岡市
Faculty of Information Science and Electrical Engineering, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka-shi, 819–0395 Japan
a) E-mail: gotou.masanori@mail.glory.co.jp

に不可能である．そのため，なるべくコンパクトなパターン空間を前提とすることで，全てのパターンを収集した究極的状況により近い状況を実現する必要がある．本研究では，そのような状況の下でパターンの多寡による分布構造の変化を明らかにし，その漸近的性質を解析することでパターンの真に分布の解明に取り組む．

第二の条件，すなわち解析による誤差がないことについては，以下に述べるような多くの手法では満足できない．例えば，複数クラスの分布間の関係を解析するための最も単純な指標として，認識率が挙げられよう．しかし，認識率からは，クラス分布がどのようにオーバーラップしているかといった，定性的な知見は得られない．認識率による Confusion Matrix を考えれば，正しく認識できないパターンが存在するクラス間の関係を解析することはできるが，それ以外のクラス間ではクラス間の関係性を表現することができない．また，分布構造の可視化手法として，主成分分析による 2 次元や 3 次元といった低次元空間への射影が挙げられる．しかし，主成分分析では射影の際に多くの情報が失われてしまい，原パターン空間での位置関係などが把握できなくなる．

そこで本論文では，大規模パターンの分布構造をネットワーク解析手法により明らかにし，その有効性を実証することを目指す．ここでいうネットワークとは，各パターンを 1 ノードとし，その近傍関係によりエッジを付与して構成される，無向若しくは有向グラフのことを指す．大規模パターンを対象とすれば，それに比例してネットワークも大規模化することになる．いわゆるスモールワールドやスケールフリーなどのネットワークに関する研究の進展もあり，大規模ネットワークの解析手法が様々に提案されている [5]～[7]．

ネットワーク解析手法によるパターンの分布解析の特徴は次の 4 点である．第一に，パターン分布をネットワークで表現することで，個々のパターンの分布内における相対位置関係を誤差なく明示的に扱える．したがって，誤認識パターンをネットワークと組みで考えることで，分布中でどのような位置関係にあるパターンに誤認識が生じているかを解析できる可能性がある．第二に，初期値などのパラメータに依存することなくネットワークを構築することが可能であるので，あるデータセットに対する分布構造の解析結果は一意に定まる．完全に解析対象のデータのみによって定まる解析結果を得られる手法であることは，パラメータの調整が必要な解析手法に比べて，真の分布を解明するうえで有用である．第三に，特定クラスのパターンが集中的に分布している箇所から外れて存在するパターン，いわゆるアウトライア，を検出できる．アウトライアを的確に検出できれば，その個数や分布傾向などから特徴量の有効性が評価できる可能性がある(注1)．第四に，各パターンのクラス情報に基づいてネットワークを粗視化すれば多クラス間の相対位置関係を把握することができる．多クラス間の分布構造が明らかになれば，それに基づいた識別器の検討などの応用が可能である．

本手法によるパターンの分布解析は，新たな視点での分布解析の手法であり，対象とするパターンの属性（例えば，文字，顔，一般物体など）は限定されない．しかしながら，手法の妥当性を検証する点においては，タグ付けされたクラスの曖昧性を極小化する必要がある．そこで本論文では，最も単純な解析対象として数字画像を対象とした実施例を提示し，本手法の有効性を検証する．

本論文では，パターンの真の分布を解明するために必要な，パターンの分布構造を誤差なく解析し複数クラスの位置関係を明確にすることを目標とした分布解析を行う．具体的には，ネットワークとして最小全域木（Minimum Spanning Tree，以下 MST）を利用し，パターンの分布解析の対象として約 50 万個活字数字画像と約 80 万個の手書き数字画像を用いた実施例を示す．パターン間の距離をエッジの重みとして構築した MST は，高次元空間におけるパターンの分布構造をノード間の近傍関係として表現するので，高次元空間中でのパターンの分布構造の情報を失うことなく様々な形で可視化することができる．MST は木構造という制約をもつため単純なネットワークとなるが，少なくとも一つの最近傍のノード間は必ずエッジが付与されかつ高速に導出が可能であるので，大規模なパターンの分布解析として有効な手法である．また，文字であれば，画像サイズを制限しても十分に認識を行うことが可能であるので，全てのパターンを収集した状況にできる限り近い状況を作り出すことができる．本論文では，本手法の詳細を述べるとともに，大規模な活字，手書き数字の分布構造の解析結果からパターンの多寡によるパターン分布の変化を明らかにするこ

(注1)：用語「アウトライア」の厳密な定義は一般に定まっていない．したがって本論文で定義する「アウトライア」もその一例でしかない．後の実験結果によりその妥当性を検証する．

とで，本手法による分布構造解析の妥当性を検証する．

2. 関連研究

大規模パターンを用いたパターン認識に関する研究は数多く報告されている．例えば，Torralba ら [1] は，インターネットから収集した 8000 万個の画像を用いれば，最近傍決定則のような単純な識別手法であっても高い認識率が達成できることを示している．また，インターネット上のデジタルデータを用いた大規模パターン認識の例としては，他にも文献 [2]～[4] 等で用いられたものがある．しかしながら，これらの研究事例は主に認識率の改善について議論されるにとどまっており，解析対象としている大規模パターンの分布構造に関する考察は行われていない．

パターン認識に限らず，多変量解析を行う分野ではデータの分布解析に関する手法が数多く提案されている．その代表的な例として，主成分分析や独立成分分析，多次元尺度構成法，Isometric feature mapping などの高次元空間での本質的な特徴を保ったまま次元削減を行う手法 [8]～[14] や最尤法やベイズ推定，混合分布モデルなどにより特徴空間での確率密度分布を推定する手法 [12], [13] がある．しかしながら，次元削減を行う手法では前述のようにパターン間の近傍関係に関する情報が失われてしまうという課題があり，確率密度分布を推定する手法では未知の分布に対して適切な初期値や確率モデルなどのパラメータを選択することが困難であるという課題がある．このため，パターンの真の分布を解明するためには新たな解析手法の導入が必要である．

ネットワーク表現が分布構造の解析に有効である実例として，クラスタリングへの応用が挙げられる．例えば，MST による手法では画像のセグメンテーション [15] のようなパターン認識に関連する研究だけでなく，医学分野での疾患の感染地域や遺伝子発現データのクラスタリングへの応用例もある [16], [17]．また，スペクトラルクラスタリング [18], [19] では本研究と同じように一つのパターンを 1 ノードとしたパターンのネットワークを構成し，そのスペクトル（固有値）を計算することでパターン間の類似度行列を算出してパターンの分布構造に基づくクラスタリングを実現している．

以上のような先行研究に対し，本研究では大規模パターンの相対位置関係，すなわち分布構造を低次元化やモデルによる近似などの誤差なくネットワークにより表現し，各パターンのクラスラベルを利用してパターン及びクラスの分布構造の解析を行うことでパターンの真の分布の解明を目指す．本論文では，文献 [20]～[22] での分布構造の解析を発展させ，手書き数字画像に加えて活字数字画像の分布構造の解析を行い，パターン数の増加によるパターン分布の変化を明らかにするだけでなく本手法による分布構造解析の妥当性についても考察する．

3. MST を用いた分布解析

3.1 解析対象

本論文では，16 × 16 画素の 2 値数字画像の分布解析を行う．実際に用いたデータセットの詳細については，**4.** で後述する．数字画像であれば 16 × 16 画素程度の小さなサイズの 2 値画像でも，パターンのクラスを正しく識別することが十分に可能である．前述のとおり，比較的小さなパターン空間中で大量のパターンを用いた解析を行えば，全てのパターンを収集した究極的状況により近い状況でのパターン分布が明らかとなる．また，数字画像は入手性が高いので，顔や風景などの一般的な物体画像に比べてクラスごとに十分なデータ数を確保することが容易である．加えて，数字であれば各パターンに付与されたカテゴリー情報の曖昧性も極小化できるのでクラス分布も含めた厳密な議論が容易である．なお，解析対象である各パターンは，256 次元の 2 値ベクトルとしてパターン空間中に分布するので，各パターンは，256 次元のハイパーキューブの頂点にのみ存在する(注2)．

3.2 パターン間距離

本論文では，パターン間の距離尺度として，ハミング距離を用いる．本論文で使用するデータセットは，非常に多くの 2 値画像により構成されるため，距離計算を高速かつ少ないメモリ領域で演算可能なハミング距離は好都合である．また，距離値が白黒が異なる画素数に相当するため，解釈も容易である．例えば，16 × 16 画素の 2 値画像間のハミング距離が 25 であれば，その画像ペアの画素値を比較すると，白黒が異なる画素が 25 画素存在する (これは，全体で 256 画素の画像の約 10%の画素に相当する)．このような理

(注2)：パターンがハイパーキューブの頂点にのみ存在する分布構造は，多少特殊な分布と言える．しかしながら，例えば 2 値画像をぼけ変換して濃淡化した場合，ぼけ変換は線形変換（すなわち線形写像）であるから，このハイパーキューブ上での分布の大局的構造は，ある程度維持されることになる．

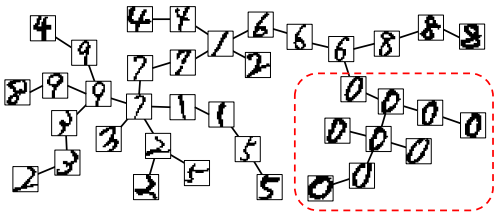

図 1　少数のパターンによる MST の例
Fig. 1　A tiny example of minimum spanning tree.

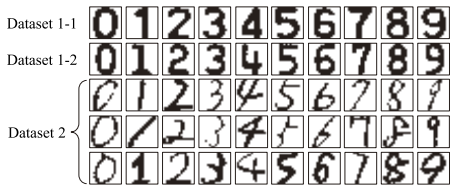

図 2　数字画像のパターン例
Fig. 2　Examples of digit images.

由で本論文ではハミング距離による距離尺度を用いたが，例えばユークリッド距離など他の距離尺度であってもパターン間距離からネットワークを構成して分布解析を行うことが可能である．

3.3　最小全域木

本論文では，ネットワークとして MST を利用する．具体的には，各パターンを 1 ノードとし，パターン間のハミング距離をエッジの重みとして MST を構成する．パターン間の距離を基に MST を作成することで，元のパターン空間における近傍関係を保存することが可能となる．実際に，少数の手書き数字パターンを用いて作成した MST を図 1 に示す．2 値画像において，ハミング距離は画像間で白黒が異なる画素の数となる．したがって，MST の性質により，隣り合うノードは視覚的に近いパターンとなる．例えば，図 1 のような少ないパターン数であっても，右下のクラス “0” のパターンは互いにエッジで接続されている．本論文では，このように互いにエッジで接続される同じクラスラベルをもつノード群をクラスタとして解析を行う．

このようにして作成した MST は，次の四つの特性をもつため大規模パターンの分布解析手法として適している．第一に，前述のとおり，局所的に見ると類似画像が集まってクラスタを構成する．第二に，クラスタ間のエッジも距離が小さいものを選択するため，クラスタ同士の近傍関係を保持する．第三に，大局的に見てパターン空間全体の構造を保持する．更に，大規模なノード数の MST を高速に作成できるアルゴリズムが複数存在し [23]，ネットワークの構造を記述するために必要な空間量が少ない．第四に，全域木であるため同一クラス内だけでなく異クラス間のパターンの位置関係を解析することが可能となる．

なお，以下では，パターンの多寡による影響を調べるために，少数のパターンのみを用いて実験を行う場合がある．その際は，使用するパターンを無作為に選択する．また，選択されるパターンによって生成される MST が異なるため，使用するパターンの選択と MST 生成を複数回繰り返し，分析値の平均をとることとした．

4.　実 験 試 料

本論文では，分布の異なる二つの数字画像データセットの分布解析を行い，その結果を比較することで解析手法の妥当性を考察する．具体的には，535,494 個の活字数字画像（Dataset 1）と 822,714 個の手書き数字画像（Dataset 2）を使用して実験を行った．データセットに含まれる各画像は，1 数字単位に切り分けられており，複数の人間が目視によって “0” から “9” までの正解ラベルを付与している．本論文では，これらの画像をバイキュービック法を用いて 16 × 16 画素へ正規化し，固定しきい値で 2 値化した画像を実験に用いた．図 1 に本論文で使用する数字画像の一例を示す．

Dataset 1 は，紙幣に印字された記番号の活字数字画像である．このデータセットは，二つのサブセット（Dataset 1-1, Dataset 1-2）で構成される．各サブセットは，異なる国で流通している紙幣の記番号の画像であるためフォントが異なっている．記番号は，字輪の組み合わせによって機械的に印字されるためパターンの主な変動要因は，媒体の汚れや画像のボケ，2 値化誤差である．また，各サブセットの画像数は，Dataset 1-1 が 259,153 個，Dataset 1-2 が 276,341 個である．

Dataset 2 は，不特定多数の人間によって筆記された手書き数字画像である．筆記時には特に制約を設けていないため，データセットの画像は多種多様な変形を含んでいる．パターン数の内訳は，クラス “0” が約 18 万個，その他のクラスがおよそ約 6 万個ずつである．

5.　分布解析結果

本章では，MST のノードやエッジの特徴，隣接ノー

ドのクラスに着目した分析解析，及び MST を粗視化した木グラフを用いてクラスの分布構造についての解析を行う．ノードやエッジの特徴からは，ノード間の関係性すなわち局所的なパターン分布の傾向を解析でき，隣接ノードのクラスからは，クラスごとのパターン分布の傾向を解析できる．また，ノードのクラスラベルに基づいて粗視化することでクラスの大局的な分布構造についての解析結果を得ることができる．

5.1 ノードやエッジの特徴

まず，MST のノード次数とエッジ重みに着目し，Dataset 1，2 のパターン分布の解析を行う．あるノードの次数，すなわちそのノードがもつエッジの数は，そのパターンに近いパターンがどの程度存在するかを表す．また，エッジの重みはパターン間の距離を表す．よって，様々なパターン数で MST を作成すれば，そのノード次数とエッジ重みからパターンの多寡によるパターン分布の変化を定量的に解析できる．

5.1.1 ノード次数の頻度分布

図 3 に，パターン数を変えた際のノード次数の頻度分布の変化を解析した結果を示す．同図 (a)，(b) より，Dataset 1，2 ともにパターン数が増えることで次数の大きなノードが出現しやすくなることがわかる．

パターン数の増加に従って次数の大きなノードの出現頻度が増加することは，パターンの分布が一様ではなく，局所的に集中してパターンが分布する領域があることを示唆している．なぜなら，MST のエッジは最近傍か，最近傍に準ずるパターンとの間にのみ存在するため，ノードの次数が大きいということは，その周囲にパターンが密に存在していることを示すと考えられるからである．このことを確認するためにパターンの多寡による各ノードの次数変化を調査した．

図 4 は，Dataset 1，2 のある 1%個のパターンを用いた MST において，ある次数をもっていたノードが，全パターンを用いた場合に次数をどのように変化させるかを例示している．更に，各点の円の大きさは，その次数変化をしたノードの個数を反映している．また，■でプロットされた点は横軸に対する平均値である．

同図より，ノードの次数は，パターンの増加により均等に変化するのではなく一部のノードの次数が大きく増加していることが分かる．例えば，図 4 (b) から，Dataset 2 では 1%個のパターンを用いたときに次数が 4 であったノードは，全パターンを用いた場合にその次数は 1 から 29 の間に分布するが，次数の平均はおよそ 3 でありノードの次数が 10 以上に増加するノードはごく一部であることが分かる．この傾向も，パターンの分布の局所的な集中を示している．特に，次数が大きく増加している一部のノードは，その周囲に多数のノードが分布していることを表しているので，これらのノードは，パターン変形の核のようなパターンであると理解できる．

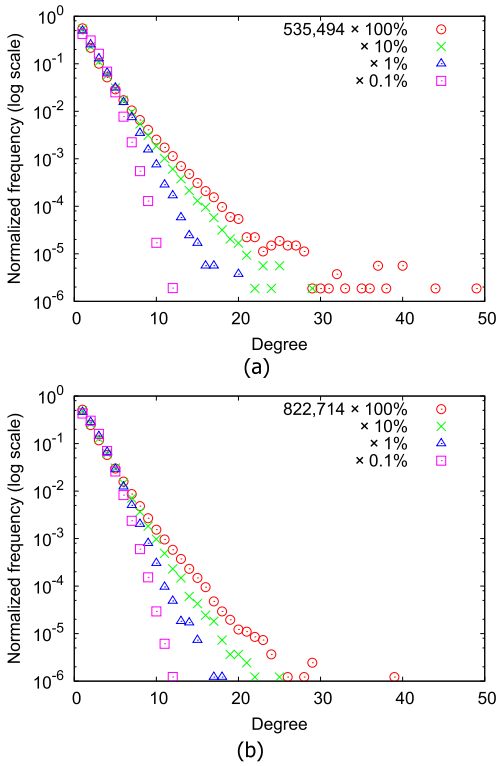

図 3 ノードの次数と出現頻度の関係．(a) Dataset 1 (活字)，(b) Dataset 2 (手書き文字)

Fig. 3 Distribution of node degree. (a) Dataset 1 (machine-printed patterns), (b) Dataset 2 (handwritten patterns).

図 3 (a)，(b) を比較すると，パターン数を増加させたときに活字よりも手書き文字の方が特定のノードの次数が高くなる場合が多いことが分かる．これは，手書き文字は異体字のようなパターン変形の核となりうるパターンが存在しているということから直感的に予想される傾向と合致する結果である．このような，その周囲にパターンが集中しているパターンの存在を明示的に検出できることは，大規模パターンを用いて分布解析を行うことの効果である．

5.1.2 エッジ重みの頻度分布

図 5 に，パターン数を変えた際の，エッジ重みの頻

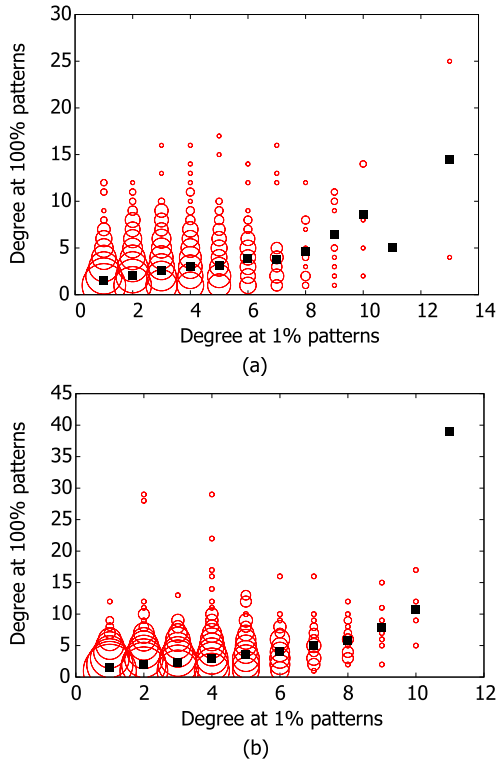

図 4　パターン数の増加による次数の変化例（■：平均）．(a) Dataset 1 (活字)，(b) Dataset 2 (手書き文字)

Fig. 4　Growth of Node degree with increase of the number of patterns. (■: average) (a) Dataset 1 (machine-printed patterns), (b) Dataset 2 (handwritten patterns).

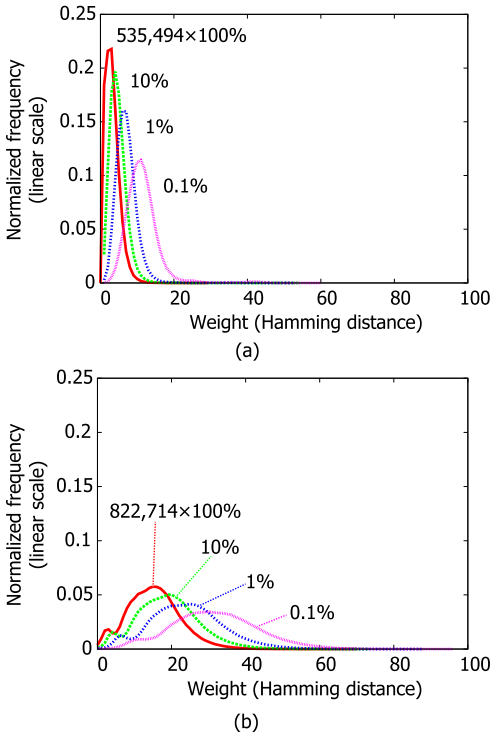

図 5　エッジ重みと出現頻度の関係．(a) Dataset 1 (活字)，(b) Dataset 2 (手書き文字)

Fig. 5　Distribution of edge weight. (a) Dataset 1 (machine-printed patterns), (b) Dataset 2 (handwritten patterns).

度分布の変化を解析結果を示す．同図 (a)，(b) より，Dataset 1，2 ともにパターン数が増えることでエッジ重みが小さく（パターン間の距離が近く）なることがわかる．

図 5 の (a)，(b) を比較すると，エッジ重みの平均と分散は，Dataset 1 のほうが Dataset 2 よりも小さく，同一のデータセット中でもパターンが増加するに従って小さくなることがわかる．重みはパターン間の距離であるから，これは，手書き文字よりも活字の分布構造が密であることと，パターンの増加によってパターン空間が密になっていく様子を示している．なお，図 5 (b) の重み 0 から 10 の範囲でエッジの出現頻度が高いのは，クラス “0” のパターンが他のクラスよりも多く，クラス “0” の分布領域においてパターンが密集しているためである．また，データ数の増加とエッジ重みの分布の変化の関係を Dataset 1，2 の各々について解析したところ，平均と標準偏差がデータ数に対してべき乗則に従って減少していることが確認できた．

5.2　隣接ノードのクラス

次に，MST の隣接ノードのクラスに着目してパターン分布の解析を行う．あるノードの隣接ノードのクラス，すなわちそのノードとエッジで接続されているノードのクラスが同一かどうかは，そのパターンの分布がクラス分布においてどのような位置に存在しているかを表す．よって，MST の隣接ノードの解析を行うことで，クラスの分布構造を解析できる．

5.2.1　隣接ノードのクラスによるノードの分類

図 6 は，MST のパターン数ごとに隣接ノードのクラスによってノードすなわちパターンを分類した結果である．この円グラフでは，隣接するノードがもつクラスラベルが，

タイプ（i）：全て自ノードと同じクラスである，
タイプ（ii）：全て自ノードと異なるクラスである，
タイプ（iii）：その他，

という三つの場合の割合を示す．ここで，タイプ（ii）

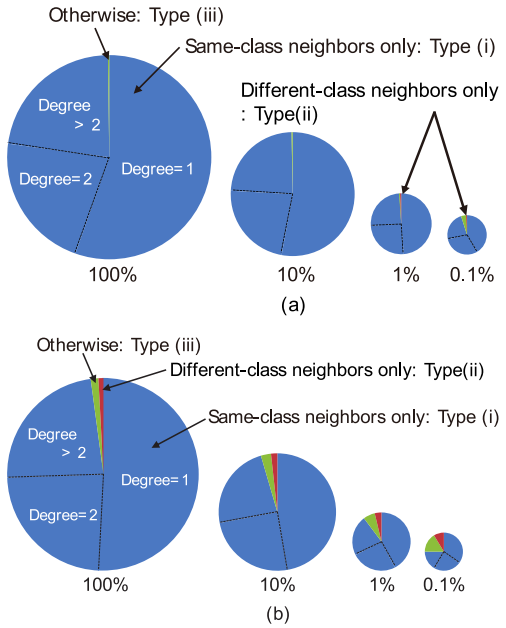

図 6 隣接ノードのクラスラベルによる分類を行ったときの各ノードの割合．(a) Dataset 1 (活字)，(b) Dataset 2 (手書き文字)

Fig. 6 Classification of nodes by their class consistency. (a) Dataset 1 (machine-printed patterns), (b) Dataset 2 (handwritten patterns).

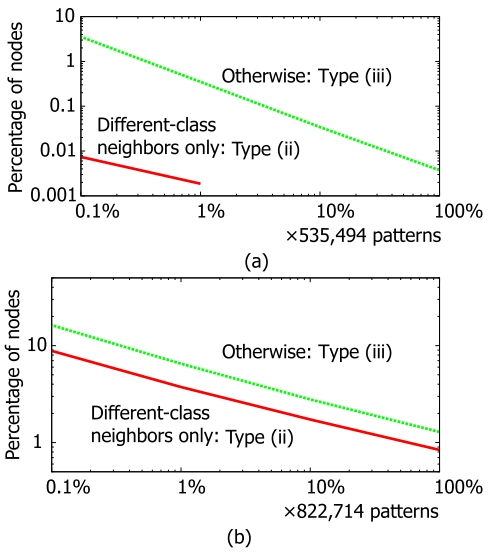

図 7 タイプ（ii）とタイプ（iii）の各ノードの割合．(a) Dataset 1 (活字)，(b) Dataset 2 (手書き文字)

Fig. 7 Percentage of nodes of Types (ii) and (iii). (a) Dataset 1 (machine-printed patterns), (b) Dataset 2 (handwritten patterns).

の場合となったノードは，特定クラスが集中的に分布している箇所から外れて存在するいわゆるアウトライアのパターンに相当する．そして，タイプ（iii）の場合となったノードは，クラス間の橋渡しをしているパターンである．加えて，タイプ（i）の場合となったノードについては，ノードの次数が 1 である葉ノード，次数が 2 より大きい部分木の根ノード，及びその他（次数が 2）のノードの割合について調べた．

図 6 より，ほとんどのノードがタイプ（i）であることがわかる．タイプ（i）のノードのうち次数が 1 であるノードの割合が，MST のパターン数が増加するに伴い増えていることは，**5.1.1** で考察したように，パターン変形の核となりうるパターンの周囲にパターンが分布しているような分布構造をもつことを示している．更に，クラス間をつなぐノードが少ないことから，パターン空間中に同じクラスのパターンが集まった領域があることが理解できる．特に，Dataset 1 ではタイプ（ii）のノードは MST のパターン数が 1%以下の場合にしか出現しておらず，その割合は 0.01%未満であり活字パターンがクラスごとに集中して分布している様子をよく表している（図 6 (a)）．

図 7 に，タイプ（ii）とタイプ（iii）のノードの割合の推移を示す．活字，手書き文字によらずどちらのタイプも，MST のパターン数が増加するに伴いべき乗則に従って全ノードに対する割合が低下していることがわかる．Dataset 1 の実験では，10%個以上のデータ数での MST ではタイプ（iii）のノードが存在しないためノードの割合が減少する度合いを判断することは困難であるが，Dataset 2 の実験では，パターン数が 10 倍になると，どちらのタイプも割合がおよそ 40%へ低下している．このように，タイプ別のノード数の割合の推移からパターン数に対するおおよそのアウトライア数の予測が可能である．例えば，Dataset 2 の手書き文字について，データ数が現状の 10 倍の約 800 万個になった場合，アウトライアの割合は更に少なくなっておよそ 0.386%（約 3,000 個）になると予測される．

タイプ（ii），(iii）の数が少ないということは，半教師付き学習の観点から次のような解釈も可能である．すなわち MST において，タイプ（ii），(iii）のものだけラベル付けされており，タイプ（i）についてはラベルなしだとする．今，leave-one-out 的に，この部分的にラベル付られた MST を構成するノードの一つが入

力だとする．その際，同ノードから MST 上で最近傍のラベル付きノードを参照することで，100%の認識率が達成できることになる．すなわち全体データのうち活字については 0.0037% (= (20 + 0)/535,494)，手書き文字については 2.1% (= (6,878 + 10,553)/822,714) についてのみラベル付するだけで，あとは残りのラベルなしデータと MST の構造を用いて完全な認識が可能になる．現実には，どのパターンがタイプ（ii），(iii）であるかは事前にはわからないので，この手法は使えないが，SVM（Support Vector Machine）と同様，極めて少数のクラス境界パターンの認識に及ぼす重要性が示唆される．

5.2.2 最近傍決定則による認識率との関係

本節では，前節で考察した Dataset 2 の MST から得られるノードの分類の傾向（図 7 (b)）と最近傍決定則による認識率との関係について述べる．MST の性質から，あるノードがもつエッジのうち，少なくとも一つは最近傍パターンに接続されているため，異なるクラスと隣接するノードは，最近傍が異なるクラスのパターンである確率が高い．このため，文献 [1] のように大量の学習パターンを使用した最近傍決定則で認識を行った場合，誤認識されやすい．言い換えれば，前節で考察した MST の各種類のノードの出現傾向と，最近傍決定則による認識を行った場合の誤認識率には相関関係があると予測される．

そこで，最近傍決定則による大量の学習パターンでの手書き数字認識実験を行った．実験には，前節までの実験でも用いた 16 × 16 画素の 2 値画像を使用し，距離尺度についてもハミング距離を用いた．また，認識率の算出には leave-one-out 法を用いた．学習パターン数の変化については，学習パターン数を全体の 0.1%から 100%まで 10 倍刻みで増加させた．全ての画像データを用いずに認識実験を行う場合，学習パターンに含まれない画像データから追加の学習パターンを無作為に選択して学習パターン数を増加させ，実験を行った．

図 8 に誤認識率を対数グラフにプロットした結果を示す．同図には，図 7 (b) に示したタイプ（ii）とタイプ（iii）のノードの割合の推移もプロットしてある．学習パターン数の増加に伴い，誤認識率は低下する傾向にあり，822,714 個全てのパターンを用いて認識を行った際，最も低い誤認識率 0.91%が得られた．

また，この結果から認識に用いる学習パターン数を 10 倍に増加させると，誤認識率がおよそ 40%へ低下していることが分かる．この傾向は，前節で述べた MST のタイプ（ii）とタイプ（iii）のノードの出現傾向 (図 7) と一致しており，両者に相関関係があることが確認できた．なお，大規模データによる認識率の変化に関しては，同様の傾向が文献 [1] でも指摘されている．

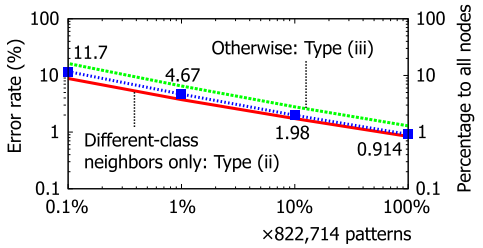

図 8 Dataset 2 (手書き文字) におけるパターン数と誤認識率，タイプ（ii）とタイプ（iii）の各ノードの割合の関係

Fig. 8 Error rates and percentage of nodes of Types (ii) and (iii) under different dataset sizes of Dataset 2 (handwritten patterns).

5.3 クラスタ木による MST の粗視化

次に，MST 上でのクラスの分布に着目してパターン分布の解析を行う．前述のとおり，MST 上では局所的に見ると類似画像が集まってクラスタを構成し，クラスタ同士の近傍関係も保持されている．そこで，MST において隣接する同じクラスのノードを併合した新たな木グラフ (以下クラスタ木) を作成した．例えば，図 1 で言えば，右下のクラス “0” のようなクラスタを構成するノード群はクラスタ木において全て一つのノードに集約表示される．この MST を粗視化したクラスタ木を用いることで，クラスごとのパターンの分布状況やクラス間の関係をより大局的に捉えられると期待できる．

図 9 は，Dataset 1，2 の全パターンを使用した MST から作成したクラスタ木である．クラスタ木では，一つの円が一つのクラスタを表し，数字がクラスを，円の大きさがそのクラスタに属するパターンの数を表している．また，Dataset 2 では全てのクラスタを描くにはクラスタ数が多いため，要素数 100 以下の微小なクラスタは図から省いた．微小なクラスタはあるクラスのパターンが集中して分布する領域から外れたパターンと考えられ，省略してもクラスの分布や隣接関係の大勢に影響はないと判断した．ただし，大きなクラスタ間を結ぶように微小クラスタが存在している場合は，木構造が崩れないようにクラスタを残し，

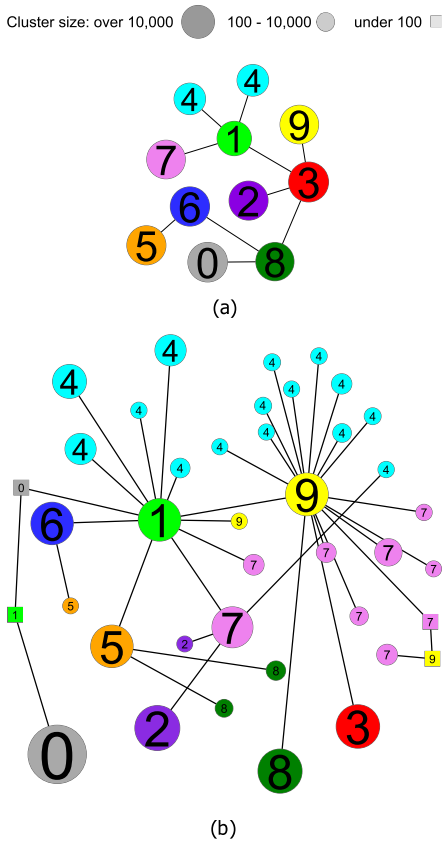

図 9　クラスタ木．(a) Dataset 1 (活字)，(b) Dataset 2 (手書き文字)

Fig. 9　Cluster tree. (a) Dataset 1 (machine-printed patterns), (b) Dataset 2 (handwritten patterns).

例外として区別するために四角で表記した．

図 9 (a) に示すとおり，活字のクラスタ木はクラスごとに巨大なクラスタが存在し，クラス “4” は 2 クラスタに分断化したが，他のクラスは 1 クラス 1 クラスタとなった．クラス “4” が 2 クラスタになった理由としては，図 1 のフォント形状から考えるとデザインの異なるフォントが核となってクラスタを構成したためであろう．クラス “4” 以外のクラスについては，フォントのデザインは異なっているが，自クラス内での分布の広がりよりも他クラスの分布との距離が遠いため 1 クラス 1 クラスタとなっていると予想される．

一方で，手書きのクラスタ木は図 9 (b) のようにクラスごとに複数のクラスタが形成された．例えば，クラス “4” やクラス “7” は複数のクラスタが形成されており，分布領域が分断されている．複数クラスタへの分断の程度がクラスごとに異なるので，分断化されたクラスは分断化されてないクラスに比べ，パターン分布に疎密が大きいと予測される．すなわち，複数クラスタへの分断が生じているクラスは，一つの標準形状とその連続的変形によりクラス全体が生成されているというよりは，異体字のような核となるパターンが複数存在していると考えられる．

クラスタ木のトポロジーは，多クラス間の近傍関係を表現している．例えば，図 9 (b) のクラスタ木からクラス間の近傍関係を見ると，クラス “1” と “9” が多くのクラスタをつなぐハブの役割を果たしていることが分かる．特にクラス “1” は複数の巨大クラスタの架け橋となっている．このクラス “1” の特徴には，数字の “1” が，基本的に縦方向の単純なストロークのみで構成されることが関係している．全体的に，数字画像パターンは横方向よりも縦方向に長い形状をしている．そのため，他クラスのパターンの中でも縦方向に長いパターンがクラス “1” の近くに分布し，どのクラスもクラス “1” と近いという状況を作ったと考えられる．

以上のように，クラスタ木による分布構造の解析では，アウトライアだけでなく異体字に代表されるような同一クラス中での分布の疎密があることが，複数クラスタへの分断という形で可視化できる．また，そのトポロジーからは多クラス間の近傍関係を同時に確認できる．これは，高次元空間中でのパターンの分布構造を保持できるネットワークによるパターン分布解析の特徴である．

6. むすび

本論文では，約 50 万個の活字数字画像と約 80 万個の手書き数字画像のパターン分布を MST を用いてネットワークとして表現し，その構造的特徴からパターン分布の構造を明らかにした．本手法の特徴はパターンの分布解析を，MST におけるノードやエッジの特徴，クラスタ木などのネットワーク解析的な手法により行っていることである．MST 及びそのクラスタ木から得られた解析結果は高次元中での分布構造を保持しているので，パターンの多寡によるパターン分布の変化を明らかにすることができた．本論文では，クラスラベルが正しいことを前提に解析を行ったが，クラスラベルそのものに曖昧性がある場合でも，曖昧なクラスを表すラベル（例えば，クラス “7” または “9” であることを表すクラス “7-9”）を新たに導入すれば同様の解析を行うことができる．

本手法により，MST のノード次数とエッジ重みに着目して解析を行うことで，大規模パターンの分布構造に疎密があることとが確認できた．また，隣接ノードのクラスによるノードの分類を行った場合に，各ノードの割合から大規模パターン認識による誤認識率の変化が予測できることを確認した．加えて，MST において隣接する同じクラスのノードを併合したクラスタ木により MST を粗視化することで，同一クラス内でのパターン分布の疎密を解析できることを確認した．クラスタ木による粗視化によれば，多クラス間の近接関係を特徴的に可視化することができる．本手法を活字数字画像と手書き数字画像に適用し，解析を行った結果から，これらの特性を用いて異なるデータセットの分布構造を定量的に比較することについての妥当性についても検証できた．本論文では 16×16 画素の 2 値数字画像を解析対象としたが，より大きなサイズの画像や多階調の画像であっても例えばユークリッド距離のような距離尺度を用いれば MST を構築できるので本手法を適用することは可能である．

今後の課題として，よりクラス間の関係性に焦点を当てた解析を行う事が挙げられる．更に，様々なパターン空間中での分布解析を行うことで分布構造の漸近的傾向も明らかになると期待できる．また，パターン空間の状況をより適切に表現できるようなグラフ構造の導入を検討する．そして最終的には，これら検討を通して，例えば Support vector やカーネルの意味の再吟味，分布構造に依拠した最近傍認識の高速化，次元解析，正規性仮定の妥当性の検証 [24] といった展開を目指す．更には分布の「不自然さ」（非等方性）から，文字という記号パターンのデザインの概念 [25] についても，何らかの知見が得られるものと期待している．

文　献

[1] A. Torralba, R. Fergus, and W.T. Freeman, “80 million tiny images: A large data set for nonparametric object and scene recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol.30, no.11, pp.1958–1970, Nov. 2008.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” Proc. Computer Vision and Pattern Recognition (CVPR), pp.248–255, Miami, USA, June 2009.

[3] J. Xiao, J. Hays, K.A. Ehinger, A. Oliva, and A. Torralba, “SUN Database: Large-scale Scene Recognition from Abbey to Zoo,” Proc. Computer Vision and Pattern Recognition (CVPR), pp.3485–3492, San Francisco, USA, June 2010.

[4] F. Schoroff, A. Criminisi, and A. Zisserman, “Harvesting Image Databases from the Web,” IEEE Trans. Pattern Anal. Mach. Intell., vol.33, no.4, pp.754–766, Oct. 2011.

[5] L.A.N. Amaral, A. Scala, M. Barthélémy, and H.E. Stanley, “Classes of small-world networks,” Proc. Natl. Acad. Sci. USA, vol.97, no.21, pp.11149–11152, Oct. 2000.

[6] R. Albert and A.-L. Barabási, “Statistical mechanics of complex networks,” Reviews of Modern Physics, vol.74, pp.47–97, June 2002.

[7] A.-L. Barabási, “Scale-Free Networks: A decade and beyond,” Science, vol.325, no.5939, pp.412–413, July 2009.

[8] I. Jolliffe, Principal Component Analysis, Second Edition, Springer-Verlag New York, New York, 2002.

[9] A. Hyvärinen and E. Oja, “Independent component analysis: Algorithms and applications,” Neural Netw., vol.13, no.4, pp.411–430, May 2000.

[10] T.F. Cox and M.A.A. Cox, Multidimensional Scaling, CRC Press LLC, Boca Raton, 2010.

[11] J.B. Tenenbaum, V. De Silva, and J.C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” Science, vol.290, no.5500, pp.2319–2323, Dec. 2000.

[12] R.O. Duda, P.E. Hart, and D.G. Stork, パターン識別，尾上守夫（監訳），新技術コミュニケーションズ，東京，2001.

[13] C.M. Bishop, パターン認識と機械学習（上，下）：ベイズ理論による統計的予測，元田　浩，栗田多喜夫，樋口知之，松本裕治，村田　昇（監訳），シュプリンガー・ジャパン，東京，2007-2008.

[14] 赤穂昭太郎，カーネル多変量解析—非線形データ解析の新しい展開，岩波書店，東京，2008.

[15] P.F. Felzenszwalb and D.P. Huttenlocher, “Efficient graph-based image segmentation,” Int. J. Comput. Vis., vol.59, no.2, pp.167–181, Sept. 2004.

[16] S.C. Wieland, J.S. Brownstein, B. Berger, and K.D. Mandl, “Density-equalizing euclidean minimum spanning trees for the detection of all disease cluster shapes,” Proc. Natl. Acad. Sci. USA, vol.104, no.22, pp.9404–9409, Oct. 2007.

[17] Y. Xu, V. Olman, and D. Xu, “Clustering gene expression data using a graph-theoretic approach: An Application of Minimum Spanning Trees,” Bioinformatics, vol.18, no.4, pp.536–545, April 2002.

[18] J. Shi and J. Malik, “Normalized Cuts and Image Segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol.22, no.8, pp.888–905, Aug. 2000.

[19] A.Y. Ng, M.I. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” Advances in Neural Information Processing Systems, vol.2, pp.849–856, Dec. 2002.

[20] 石田良介，吉田　晃，蔡　文傑，フォンヤオカイ，内田誠一，“大規模数字画像データベースを用いたパターン分布解析，”

信学技報，PRMU2011-117, 2011.
[21] S. Uchida, R. Ishida, A. Yoshida, W. Cai, and Y. Feng, “Character image patterns as big data,” Proc. 2012 International Conference on Frontiers in Handwriting Recognition (ICFHR), pp.477–482, Bari, Italy, Sept. 2012.
[22] 石田良介，吉田　晃，蔡　文傑，フォンヤオカイ，内田誠一，“大規模数字画像データベースを用いたパターン分布解析，” 画像の認識・理解シンポジウム（MIRU2012），IS2-34, Aug. 2012.
[23] R.L. Graham and P. Hell, “On the history of the minimum spanning tree problem,” Annals of the History of Computing, vol.7, pp.43–57, Jan.–March 1985.
[24] 鶴岡信治，村瀬晶彦，木村文隆，横井茂樹，三宅康二，“人間の字種識別基準を用いた自由手書き片仮名文字認識，” 信学論（D），vol. J68-D, no.4, pp.781–788, April 1985.
[25] 小川英光（編著），パターン認識・理解の新たな展開—挑戦すべき課題，電子情報通信学会，東京，1994.

（平成 25 年 8 月 12 日受付，10 月 17 日再受付）

後藤　雅典

2002 京大・工・電気電子卒．2005 グローリー（株）研究開発センター入社．画像パターンの解析・認識に関する研究に従事．

石田　良介

2012 九大・工・電子情報工卒．現在，同大システム情報科学府修士課程．画像パターンの分布解析に関する研究に従事．

蔡　　文杰　（正員）

1990 中国武漢大・理工・電子卒．1996 同大大学院修士課程（情報）了．2007 オーリッド（株）入社．博士（工学）．文字認識に関する研究に従事．

内田　誠一　（正員）

1990 九大・工・電子卒．1992 同大大学院修士課程（情報）了．セコム（株）勤務を経て，現在，同大システム情報科研究院知能システム学部門教授．博士（工学）．画像パターン・時系列パターンの解析・認識に関する研究に従事．2003 本会 PRMU 研究奨励賞，2009 MIRU 長尾賞，2007 IAPR/ICDAR Best Paper Award，2009 本会論文賞，ICFHR2010 Best Paper Award，MIRU2011 優秀論文賞各受賞．IEEE，情報処理学会各会員．