ネットワーク表現による大規模パターン集合の分布構造解析

後藤, 雅典

https://doi.org/10.15017/1807069

ネットワーク表現による
大規模パターン集合の分布構造解析

後藤 雅典

2017年 1月

目 次

第1章

序論

1.1 背景

1.1.1 パターンの分布構造解析

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できる．

この目標のために本研究では，できる限り大量のパターンを収集したうえで，それらの分布をできる限り正確に把握する解析法を提案し，そして実際に解析した結果の吟味を目的とする．後述のように，従来，分布解析に広く用いられている主成分分析 [1] による低次元化や混合分布モデルによる近似 [2] などの手法は，ある意味，巨視的にパターン分布を捉えることを目的としている．したがって，詳細な分布構造の解析という本研究の視点からは不十分である．さらにパターン認識という応用を前提にすれば，むしろ重要なのは，クラス境界付近などの局所的な分布構造の正確な把握である．したがって，従来手法は応用的側面からも，十分とは言えない．

これに対して本研究では，できる限り正確に分布構造を把握する方法として，大規模なパターン集合のネットワーク表現[1] に基づいた分布構造の解析法を提案する．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，パターン間距離を基準とした近傍関係によりエッジを付与して構成される，無向もしくは有向グラフのこと

[1]用語「ネットワーク」の厳密な定義は，一般に定まっていない．本研究での「ネットワーク表現」とは，グラフを用いて集合の要素の関係性を表現することを指す．具体的には，集合の要素を1つのノードとし，その関係性によりエッジを付与したグラフを用いる．

を指す．大規模なパターン集合を対象とすれば，そのサイズ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 [3] やスケールフリー [4] などの複雑ネットワーク [5, 6] 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 [7]．しかしながら，これらの研究ではスケールフリー性やスモールワールド性，クラスター性などの複雑ネットワークが持つ一定の共通性，すなわち大局的な分布構造を解析することに主眼が置かれている．

我々の分布解析では，むしろ局所的な微細構造に注目する．特にパターン集合が複数のクラスから構成されるとき，そのクラス間の境界がどのように構成されるのかについて注目する．例えば “0” から “9” の 10 クラスからなる数字画像集合を扱うとき，特徴空間内におけるクラス “1” と “7” の隣接状況を定量的・定性的に把握する．後に明らかになるように，この局所構造を正確に解明できることこそが，ネットワークによる分布解析の利点である．解析の結果，多クラス分布間の隣接性やオーバーラップの様子などの分布構造を把握できれば，特徴量の妥当性の検討や識別器の効率的な設計が可能となる．

1.1.2　パターンの分布構造解析に必要な二条件

パターンの真の分布構造の解明には，次の二つの条件がある．第一の条件は，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．第二の条件は，利用する分布構造解析手法が，パターン間の位置関係を正確に把握できることである．

第一の条件，すなわちパターン集合の大規模化は，解析を行うパターン空間の次元と密接に関係する．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，パターン空間は次元数のべき乗で拡大するので，高次元になれば存在しうる全てのパターンを収集することは事実上不可能となる．本論文では，パターン集合をできるかぎり大規模化すると同時に，次元数の低いパターンを扱うことで，極力理想の状況に近づけることとする．より高次元なパター

ンについても，提案する分布構造解析手法を用いれば，与えられたパターン集合内のパターンについては正確な分布把握が可能である．しかし，そのパターン集合自体が空間内で粗になるため，一般的な分布推定の問題と同様，局所構造を含めた分布把握という目標達成は困難になる．

第二の条件，すなわちパターン間の位置関係の正確な把握は，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，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．また，複数クラスが混在するような状況においては，例えば何らかの識別法によって得られた認識率や混同行列が，クラス間の分布構造の把握に利用できる．すなわち，特定2クラス間の認識率が低ければ，それら2クラスはオーバーラップして分布している，といった具合である．しかし，それらの手法では，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できるが，それ以外のクラス間ではクラス間の関係性を解析できない．

1.2　本研究の目的

本研究では，大規模なパターン集合（以下，大規模パターン集合）と損失のない分布解析法を用いて，パターンの真の分布の姿の解明に近づくことを目的とする．具体的には，大量の画像から構成される大規模パターン集合の分布構造の解析を課題とする．解析手法に関しては，ネットワーク表現による大規模パターン集合の分布構造の解析手法を提案する．また，分布構造の解析結果の認識応用についても検討する．

1.2.1　ネットワーク表現による分布構造解析

本研究では，前述した二条件を満たすデータセットと解析手法を用いて，大規模パターン集合の分布構造を明らかにする．第一の条件であるパターン集合の大規模

化については，数十万サンプルからなる 2 値数字画像データセットを用ることで対応する．数字画像であれば，比較的小さなサイズの 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可能である．上述のとおり，比較的低次元の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，一般物体画像に比べて，データエントリ企業との連携などにより入手しやすくかつラベル付けも容易であり，クラスごとに十分なパターン数を確保することが容易である．加えて，各パターンに付与されたクラスラベルの曖昧性も極小化できる．

大規模パターン集合の分布構造が明らかとなれば，これまで経験的に論じられてきた多クラス分布間の隣接性やオーバーラップの様子などの事項を具体的に吟味できる．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集合を用いた研究 [8, 9, 10, 11] は，多くの成果を挙げている．ただし，これらの研究の多くは，大規模パターン集合による認識率の改善を主な目標としている．また，大規模パターン集合であっても，クラスごとのパターン数が十分ではないという課題もある．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，クラスごとに十分なパターン数を確保した大規模パターン集合の分布構造を解析する．

第二の条件であるパターン間の位置関係を正確に把握できる手法については，パターンの分布構造をネットワークとして表現することで対応する．その概要を図 1.1 に示す．具体的には，まず大規模パターン集合の分布構造をパターン間距離を考慮したネットワークで表現する．次にその構造を定量的に解析することで，大規模パターン集合が持つ分布構造を明らかにする．各パターンをが 1 ノードに対応しており，パターン間距離を基準とした近傍関係によりエッジが付与されるこのネットワークは，パターン空間中での分布構造を完全に保存している．したがって，そのノードやエッジの情報を解析すれば分布構造を推測できる．

ネットワーク表現による大規模パターン集合の分布構造解析は六つの特長を持つ．

1. パターン間距離を考慮したネットワークでパターン分布を表現することで，個々

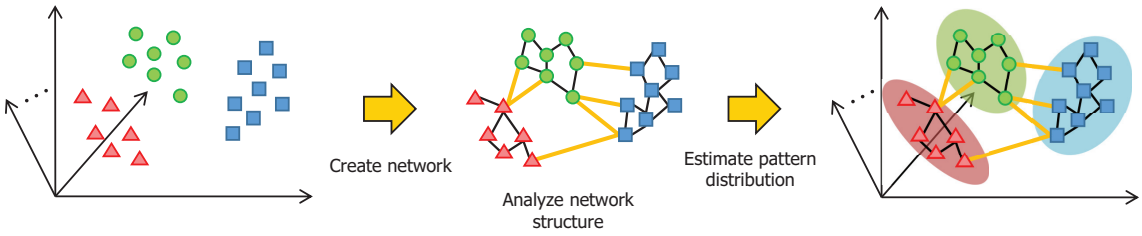


図 1.1: ネットワーク表現による分布構造解析

のパターンの分布内における相対位置関係を正確に扱える．

2. 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できるので，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．

3. 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，他のクラスの分布にオーバーラップして存在するパターン，いわゆるアウトライア[2]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．

4. 何らかの識別器での誤認識パターンをネットワークと組で考えることで，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．

5. ノードやエッジの特徴を解析することで，多クラス間の相対位置関係を把握できる．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．

6. 各パターンに付与されたクラスラベルを利用することで，クラス間の関係性を維持したままネットワークの粗視化が可能となる．この粗視化によって，大規模パターン集合の特徴的な分布構造の可視化が可能となる．

[2]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，本研究で定義する「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．

1.2.2　本研究で用いるネットワーク表現

本研究では，ネットワークとしてパターン間の距離をエッジの重みとした最小全域木 (Minimum Spanning Tree, MST) [12] と相対近傍グラフ (Relative Neighborhood Graph, RNG) [13] を用いる．MST は，木構造という制約により単純なネットワークとなるため，高速に導出できる．また，MST は，全域木という性質によりパターンの分布構造を 1 つのネットワークで表現できる．この性質により，局所的なパターンの近傍関係だけでなく，各クラスごとあるいは大規模パターン集合全体の分布構造も解析できる．RNG は，MST を内包し，かつ閉路を許すグラフ構造であるため，MST に比べてパターン間の近傍関係を余さず表現できる．

大規模パターン集合を対象とした場合，1 パターンを 1 ノードとしたままではネットワークが複雑となるため，何らかの粗視化が必要となる．特に複数クラスが存在し，それらの隣接状況を可視化により把握したい場合には，クラス間境界の状況については元の精度のままで，クラス内の状況についてはあまり重要でないとして一括表現するような粗視化が必要である．そこで本研究では，粗視化したネットワークにより大規模パターン集合の分布構造を可視化する．ここでいう粗視化とは，エッジで結合された同じクラスラベルを持つパターンを統合し，1 つのノードとして表現することである．本論文では，この粗視化処理をネットワークのクラスタ化と呼ぶ．ここで強調すべきは，異クラス間の隣接関係は，この粗視化によって失われない点である．すなわち，粗視化されたネットワークは，パターン認識へ応用するために重要なクラス境界付近などの局所的な分布構造を完全に保持している．パターンの分布構造のネットワーク表現は，パターン数が増加すれば複数のエッジが複雑に交差するため，非常に複雑な構造となる．各パターンのクラスラベルによりネットワークを粗視化すれば，この課題を解決できる．

1.2.3 サポートベクトルマシン (Support Vector Machine, SVM) の学習高速化への応用

分布構造の解析結果は，識別器の学習への応用も可能である．具体的には，クラス境界付近に分布するパターンを解析できるという特長が，Condensing [14, 15] の手法として利用できる．Condensing は，最近傍識別やサポートベクトルマシン (Support Vector Machine, SVM) [16] のようなクラス境界付近の学習パターンを利用してクラスを識別する識別器にとって有用である．クラス境界から離れた箇所に分布する学習パターンを Condensing により削除すれば，最近傍識別の高速化や SVM の学習高速化が可能となる．

そこで本研究では，ネットワーク表現による分布構造解析をパターン認識へ応用した例として，学習パターンの事前選択による SVM の学習高速化の実例を示す．提案手法の解析結果を利用して学習パターンを事前選択すれば，SVM 識別器の認識性能を低下させずに学習を高速化できると期待できる．また，従来提案されている SVM の学習高速化の手法と違い，提案手法は SVM の学習高速化のために新たなパラメータを導入する必要がない．よって，SVM のパラメータ調整に必要な処理時間を含めて考えても，提案手法により十分な学習高速化の効果が得られると期待できる．

1.3 本研究の貢献

本研究の貢献は，以下の四点である．

- MST を用いたネットワーク表現によるパターンの分布構造解析の手法を提案し，そのネットワークの構造的特徴からパターンの分布構造を明らかにする．この手法の特徴は，パターンの分布構造の解析を，MST におけるノードやエッジの特徴などから得られる指標を用いて定量的に行うことである．MST は高速に導出が可能であるので，大規模パターン集合の分布構造の解析に適してい

る．また，約 50 万の活字数字画像と約 80 万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の分布構造を明らかにする．

- RNG とその粗視化を用いたネットワーク表現による分布構造解析の手法を提案し，これによって多クラス間の相互関係が解析できることを示す．この手法の特徴は粗視化，すなわち RNG のエッジで結合された同じクラスラベルを持つパターンを統合し，1 つのノードとして表現することである．木構造という制約を持つ MST に比べて，RNG は閉路を許すグラフ構造であるため，パターン間の近傍関係を余さず表現できる．この RNG をクラス境界付近の局所的な分布構造を完全保存したまま粗視化すれば，パターン認識という応用を前提とした場合に重要となる特徴的な分布構造が容易に解析できる．そして，この粗視化によって，複数のエッジが複雑に交差する大規模なネットワークの可視化が可能となる．また，約 50 万の活字数字画像と約 50 万の手書き数字画像のデータセットおよび複数の特徴量を用いた実験を行い，提案手法によって明らかとなる分布構造の性質と有効性について示す．

- 大規模パターン集合に対する高速かつ省メモリな RNG 構築の手法を提案し，大規模パターン集合の RNG が効率的に構築できることを示す．解析対象のノードをランダムにグルーピングし，各グループに対する各ノードからの最小距離を基準にエッジの付与判定の処理回数を削減すれば，高速かつ省メモリに RNG を構築できる．また，複数のデータセットと特徴量を用いた実験を行い，提案手法の有効性を示す．

- ネットワーク表現による分布構造解析の認識応用として，SVM の学習高速化への応用の実例を示し，特に，データセットサイズの差異が結果に及ぼす影響を複数の実験により検証する．具体的には，分布構造の解析結果により学習パターンを事前選択することで，SVM の学習を高速化する手法を提案する．RNG を用いた事前選択を小規模な学習パターンに適用した先行研究では，SVM の

学習が高速化される一方でSVM識別器の認識率が大きく低下している．これに対して本研究では，約50万の手書き数字画像のデータセットとMNISTの公開データセットおよび複数の特徴量を用いた実験を行い，提案手法によってSVM識別器の認識率を大きく低下させることなく学習を高速化できることを示す．

1.4 論文の構成

本論文は，以下のように構成される．第1章では，本研究の背景および目的について述べた．第2章では，本研究の関連研究について述べる．第3章では，本研究で分布構造を解析する大規模数字画像データセットについて述べる．第4章では，MSTを用いてパターンの分布構造の解析を行い，実験結果について考察する．第5章では，RNGを粗視化したネットワークを用いてパターンの分布構造の解析を行い，実験結果について考察する．第6章では，提案手法のパターン認識への応用例として，学習パターンの事前選択によるSVMの学習高速化の実例を示す．最後に第7章で結論を述べ，今後の課題について言及する．

第2章
関連研究

2.1 大規模パターン認識

近年，インターネット上のデジタルデータを用いた 80 million tiny images [8] や ImageNet [9] など大量のパターンを含む多様なデータセット [10, 11] が整備され，それらを利用した大規模パターン認識の研究が盛んに行われている．これらのデータセットは，Convolutional Neural Network [17] などの機械学習の手法と組み合わさって多くの成果を挙げている [18]．しかしながら，これらの研究事例では，主に認識率の改善について議論されるに留まっており，解析対象としている大規模パターン集合の分布構造に関する考察は行われていない．

ところで，1 クラス当たりのパターン数という点では，依然として大規模なデータセットは存在しない．例えば，ImageNet はクラスごとのパターン数の偏りが大きい．また，80 million tiny images のサブセットである CIFAR-10 [19] は 10 クラスに対して各々6,000 パターンが含まれるのみである．これは，文字認識の分野で従来から広く利用されている MNIST [20] のデータセットと同程度のパターン数である．1.1.2 で述べたように，パターンの真の分布構造を明らかにするためには，クラスラベルが付与された大規模なデータセットにより，クラスごとに十分なパターン数を確保する必要がある．

2.2 集合の要素の関係性のネットワーク表現

集合の要素，すなわち人やものごとなどの様々な事象の関係性は，ネットワークとして表現できる．具体的には，各事象を 1 つのノードとし，その関係性をエッジで表現したグラフは，各事象の関係性をネットワークとして表現する．このネット

ワークの構造を解析すれば，各事象の分布構造が明らかとなる．また，ネットワークは，各事象が持つ複雑な分布構造を可視化する手法としても有用である．

2.2.1　ネットワークの解析

現実世界に存在する巨大で複雑なネットワークを解析する試みが多数なされている [7]．例えば，人と人の交流によって構築されるソーシャルネットワーク [21, 22] や Web ページのリンク構造 [23, 24]，論文の共著者の関係性 [25] などの解析が行われ，各ネットワークがスケールフリー性 [4, 26] やスモールワールド性 [3, 27]，クラスター性 [3] などの性質を持つことが発見されている．これらの研究では，複雑ネットワークが持つ一定の共通性，すなわち大局的な分布構造を解析することに主眼が置かれている．

2.2.2　ネットワークによる関係性の可視化

様々な情報を含んだ大規模パターン集合の解析が，目的に応じた可視化手法を用いて行われており，様々な大規模パターン集合が，可視化の手段とともに実際に提供されている．例えば，地球環境情報統融合プログラムのデータ統合・解析システム DIAS(Data Integration and Analysis System) [28] では，地球規模での様々な観測結果の 2 次元，3 次元のグラフやヒートマップによる可視化結果が利用可能である．DDBJ(DNA Data Bank of Japan) [29] では，塩基配列情報の系統樹による可視化結果が利用できる．特に分布構造（例えば，各パターン間の複雑な関係性や相互作用）を解析する場合には，ネットワークを用いた分布構造の可視化による解析が行われる．

大規模パターン集合の分布構造を可視化する場合，パターンの持つ全ての関係性をネットワーク化すれば，可視化結果は非常に複雑なものとなってしまう．この複雑なネットワークを粗視化せずに直接可視化する場合，描画に必要な計算量が増大するため，高速に描画する手段が必要となる．例えば，マルチレベル法 [30] や代数

的マルチグリッド法 [31] などの手法が研究されている [32]．しかしながら，複雑なネットワークを直接可視化しても，その結果から得られる情報は少ない．例えば，Boitmanis ら [33] のインターネットトポロジーの可視化例は，非常に複雑なネットワークとなっており，その可視化結果からネットワークの性質を直感的に理解できない．

パターンの持つ関係性に何らかの仮説を適用，すなわち解析する関係性を限定してネットワークを粗視化すれば，大規模パターン集合の特徴的な分布構造をより明確に可視化できる．例えば，ゲノム解析では遺伝子情報と癌の因果関係を仮定し，その共起構造を共起ネットワークによって可視化 [34] している．購買行動分析では購入された物品の関連性を仮定し，アソシエーションルールをネットワークによって可視化 [35] している．

本研究では，多クラス分布間の隣接性やオーバーラップの様子を解析対象として，パターンの分布構造をネットワーク表現により可視化する．具体的には，まず，全てのパターンからネットワークを構成する．次に，同じクラスラベルかつ隣接するノード群をを統合し，1 つのノードで表現することでネットワークを粗視化する．このようにして粗視化されたネットワークは，主要なクラス分布間の隣接性だけでなく，クラス境界付近などの局所的な分布構造も完全に保持している．

2.2.3　近接性グラフ

ノード間にエッジを付与する条件は様々に定義できるため，特性の異なる近接性グラフ (Proximity Graph) [36] が多数提案されている．例えば，前述した MST と RNG 以外にも最近傍グラフ (Nearest Neighbor Graph) [12] や Gabriel グラフ [37]，Voronoi 図の双対グラフである Delaunay グラフ [37] などが知られている．

近接性グラフによるネットワーク表現の特長は，そのグラフ構築にパラメータを必要としないことである．パターンの真の分布を明らかにするうえでは，解析に用いるパラメータによって解析結果が異なる手法は適切ではない．後述のように，次元削減や確率密度分布の推定による分布解析法では，何らかのモデルによる近似や

低次元化を行うため，解析に用いるパラメータによって解析結果が変化する．これに対して，ネットワーク表現を用いた分布構造解析法では，解析結果がパターンの分布構造のみによって一意に定まるため，分布構造を正確に把握できる．

しかしながら，ほとんどの近接性グラフは大規模パターン集合の分布構造を表現するには適さない．例えば，パターン間距離をエッジの重みとした最近傍グラフを用いてパターンの分布構造を表現すれば，局所的なパターンの近傍関係を表現する小規模な多数のグラフとなり，各クラスごとあるいは大規模パターン集合全体の分布構造が解析できない．また，Gabriel グラフや Delaunay グラフは，そのグラフを構築するために必要な計算量が課題となる．

2.2.4 最小全域木 (Minimum Spanning Tree，MST) によるネットワーク表現

MST は，エッジに重みを持つ全域木，すなわち与えられたノード集合を重み付きのエッジにより単連結にする木構造のうちで，エッジの重みの総和が最小となるグラフである．図 2.1(a) に MST の例を示す．MST は，木構造という制約を持つため単純なネットワークとなる．MST は，前述した最近傍グラフを内包しており [38]，局所的な分布構造を表現できる．具体的には，パターン間の距離をエッジの重みとして MST を構築すれば，最近傍ノード，すなわちあるパターンから距離が最も近いパターンに対してエッジが必ず付与される．また，全域木である MST は，パターンの分布構造を 1 つのネットワークで表現できるため，局所的な分布構造だけでなく各クラスごとあるいは大規模パターン集合全体の分布構造も解析できる．そして，Prim のアルゴリズム [39] など MST を高速に導出できるアルゴリズムが複数存在している [40]．以上のように，MST は，構造を記述するために必要なメモリ空間量が少なく，高速な構築アルゴリズムが存在するため，大規模パターン集合の分布構造の解析に適している．

上述した特長により，MST を利用した多くの研究事例が報告されている．例えば，

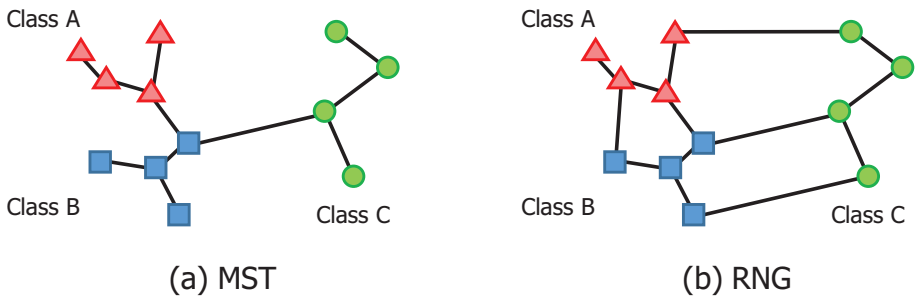


図 2.1: グラフ構造の例

画像のセグメンテーションへの応用 [41, 42, 43] や階層型クラスタリング [44]，雑音除去フィルタ [45] への応用事例が報告されている．また，パターン認識に関連する研究だけでなく，微分方程式の数値解法のためのグリッド生成 [46] やデータマイニング [47]，医学分野での疾患の感染地域 [48] や遺伝子発現データ [49] のクラスタリングへの応用例もある．

2.2.5 相対近傍グラフ (Relative Neighborhood Graph，RNG) によるネットワーク表現

RNG は，あるノード対が相対的に近傍である場合に，そのノード間にエッジを付与することで得られる無向グラフである．ここでいう相対的に近傍とは，図 2.2 に示すように, あるノード対に対して各ノードを中心とし，ノード間の距離 r を半径とする 2 つの超球が重なる空間（同図の斜線部）に他のノードが存在しないことである．図 2.1(b) に RNG の例を示す．RNG のグラフ構造の定義については，5.2.1 で詳説する．

RNG は，前述した MST を内包しており [38]，ノード間の近傍関係をより明示的に表現できる．特に，RNG は閉路を許すグラフ構造であるため，木構造という制約を持つ MST に比べて，パターン間の近傍関係を余さず表現できる．

上述した特長により，RNG を利用した多くの研究事例が報告されている [38]．例

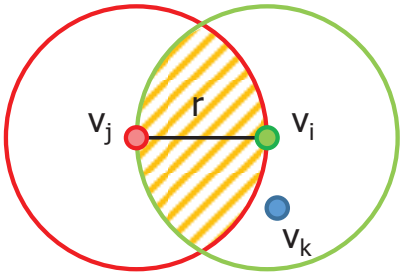


図 2.2: 相対的に近傍なノード対

えば，画像分類 [50, 51] や画像分割 [52]，クラスタリング [53, 54, 55]，特徴選択 [56] への応用事例が報告されている．また，都市構造 [57] や細菌の摂食行動 [58] のモデル化への応用例もある．

従来提案されている RNG 構築のアルゴリズム [59] の多くは，2 次元あるいは 3 次元程度の低次元空間での RNG 構築を前提としているため，高次元空間では効率的に RNG を構築できない．例えば，RNG のスーパーグラフである Delaunay グラフからエッジを削除する手法 [13] では，高次元空間中での Delaunay グラフの構築処理が課題となる．具体的には，Delaunay グラフの構築処理に必要な Voronoi 分割の計算量は d 次元の N 個のパターンに対して $d \leq 4$ の場合，$O(N^{\lfloor d/2 \rfloor+1})$ となる [60, 61]．ここで $\lfloor d/2 \rfloor$ は，$d/2$ 以下の最大の整数である．また，領域を分割する手法 [62, 63] やノード対間の角度からエッジの付与判定の回数を削減する手法 [64] でも，高次元空間に拡張した場合に領域分割や角度計算が煩雑になるという課題がある．2 次元や 3 次元の空間中で効率的に RNG を構築する手法 [65] も提案されているが，本研究が対象とする高次元空間には適用できない．そのためパターン数が少ない場合には総当たり法 [13] を用いて RNG を構築している研究事例もある [66]．しかしながら，総当たり法により N 個のパターンに対して RNG を構築する場合の計算量は $O(N^3)$ である．また，高速化のために，全ての距離をメモリ上に蓄積しようとすれば，$O(N^2)$ のメモリ空間量が必要となる．本研究では大規模パターン集合の分布構造を RNG を用いて解析するため，RNG 構築に必要な計算量とメモリ空間量の課題を解決する必要がある．

上述したRNG構築アルゴリズムとは別のアプローチとして，RNGを逐次的に更新することでRNGを構築する手法が考えられる．実際に，逐次更新によるRNGの構築手法がいくつか提案されている [51, 67, 68]．逐次更新によってRNGを構築すれば，パターン数の多寡によるRNGの構造の変化を効率的に解析できる．しかしながら，逐次更新によるRNG構築法で構築されるグラフはRNGの近似グラフであり，パラメータによって構築されるRNGが異なる場合があるという課題がある．よって，逐次更新によるRNG構築法は，本研究が対象とする大規模パターン集合の分布解析には適さない．そこで，本研究では，大規模パターン集合に対して，高速かつ省メモリにRNGを構築する手法についても提案する．

2.3 分布解析とパターン認識

従来から提案されているパターン認識の手法の多くは，分布解析の手法を応用している．パターン認識に応用されている分布解析の手法は，大きく二つに分類できる．第一は，高次元特徴空間でのパターンの主たる分布構造を保ったまま各パターンを識別可能かつできるだけ低次元な特徴空間へ射影し，次元数を削減する手法である．第二は，クラスごとに確率密度分布を推定し，パターンの分布構造をモデル化する手法である．

2.3.1 次元削減による分布解析

一般に，パターン認識を行う場合，識別対象となるパターンを何らかの特徴抽出手法を用いて特徴空間に射影し，その特徴空間中でクラスを識別する評価関数を定義する．パターンのクラスを精度良く識別するためには，自クラスの分布が密となり異クラス間の分布が粗となる特徴空間に識別対象となるパターンを射影する必要がある．その際，識別対象となるパターンを射影する特徴空間が高次元化すれば特徴抽出や識別関数の計算量が増加するため，次元削減の手法が用いられる．

パターン認識に応用されている次元削減の手法は様々である．例えば，主成分分析 [1] や独立成分分析 [69]，多次元尺度構成法 [70]，Isometric feature mapping (Isomap) [71] などの手法が知られている．これらの手法は，高次元特徴空間での本質的な分布構造の特徴を保ったまま次元削減を行えるため，パターン認識に広く応用されている．また，次元削減後の特徴空間の分布構造は，高次元特徴空間での本質的な分布構造の特徴を保つため，次元削減の手法は分布解析にも応用できる．

スペクトラルクラスタリング [72, 73] では，グラフ構造を利用した次元削減の手法が応用されている．具体的には，まず本研究と同じように1つのパターンを1ノードとしてグラフ構造を定義し，グラフ構造を表現する Laplacian matrix を生成する．そして，その行列のスペクトル（固有値）を利用して次元削減を行い，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．

次元削減の手法では，分布構造に対して寄与度の低い情報は解析できない．パターンを低次元の特徴空間に射影することで次元削減する場合，高次元特徴空間でのパターンの分布構造を解析し，目的に応じた低次元の特徴空間を定める必要がある．よって，次元削減によって解析できる情報は，分布構造の主たる基底に対する情報のみである．

次元削減による分布解析の手法は，パターンの分布構造の可視化にも応用できる．具体的には，次元削減の手法を用いてパターンの分布構造を2次元あるいは3次元にまで次元削減すれば，パターンの分布構造を可視化し，直感的に把握することも可能となる．例えば，Najafi ら [74] の Isomap による研究事例では，顔画像の分布構造や手書き数字画像の特定のクラス対の大局的な分布構造が可視化されている．しかしながら，分布構造への寄与度が低い情報は次元削減により失われてしまっている．

2.3.2 確率密度分布の推定による分布解析

確率密度分布の推定法を用いて分布解析を行うこともできる．最尤法やベイズ推定，混合分布モデルなど [2, 75] により確率密度分布を推定すれば，大局的な分布構造が解析できる．これらの確率密度分布の推定法は，大きく三つに分類できる [2, 76].

第一は，比較的少数のパラメータをもつパラメトリックモデルを用いる手法である．この手法では，パラメトリックモデルを用いて確率密度分布を表現し，そのモデルがデータに対して尤もよく合うようにパラメータを調整し，分布を推定する．この手法は，最もよく利用される手法である．しかしながら，大規模パターン集合のように真の分布が複雑な場合にはモデルの推定が困難となる．

第二は，特定の関数型を仮定せず，データから分布の形を決めるノンパラメトリックモデルを用いる手法である．この手法も，大規模パターン集合の分布構造解析には適用しにくい．例えば，最も簡単なノンパラメトリック手法のひとつであるヒストグラムによって推定された密度関数は，分布の滑らかさを表現できない．また，カーネル関数に基づく手法や k-近傍法では，パラメータによって推定される確率密度分布が変化する．

第三は，パラメトリックモデルよりも一般的な関数型を表現するセミパラメトリックモデルを用いる手法である．この手法は，上述の２つの手法の中間的な手法で，複雑な分布を表現するためにパラメータの数を系統的に増やすことができる．

有限のパターンから確率密度分布を推定することは，容易ではない．特に，高次元空間でのパターンの分布を有限のパターンから推定することは困難である [76]．また，分布解析への応用という点では，前述の次元削減を行う手法と同様に，推定結果からクラス境界付近の分布構造などの局所的な特徴を解析できない．

確率密度分布の推定による分布解析の手法でも，次元削減と同様に分布構造を可視化できる．例えば，岩田ら [77] は，パラメトリック埋め込み法を用いて分類済み Web ページや手書き数字画像の分布構造を可視化している．同様の手法としては，Hinton ら [78] や山田ら [79] の研究事例も報告されている．また，岩田らは同手法を拡張してクラス分類器の特徴を可視化し，直感的に理解することも提案している [80]．この可視化による分類器の解析は視覚的判別分析 [81] と呼ばれ，Fisher 線形判別法 [82] を応用した可視化なども提案されている．しかしながら，これらの可視化結果からはクラスごとの分布構造が定性的に把握できるだけであり，複数の結果を定量的に比較できない．

2.4 SVMの学習高速化

大規模パターン集合をSVMの学習パターンとして用いる場合，SVMの学習時間が実用上の課題となる．効率的なSVMの学習法であるchunking [83]やSequential Minimal Optimization (SMO) [84]を適用しても，SVMの学習の計算量は学習パターン数のべき乗で増加する．例えば，SMOではN個のd次元特徴ベクトルを学習パターンとした場合，SVMの学習の計算量は$\Omega(N^2d)$となる [85].

SVMの学習において，1回の学習当たりの計算量を減らすことの有効性は非常に大きい．これは，SVMはカーネルパラメータやコストパラメータなどいくつかのパラメータを最適化する必要があり，パラメータを変えながらSVMの学習を多数回繰り返す必要があるためである．例えば，パラメータの最適化として一般的なk-分割交差検証を用いたグリッドサーチ [86]では，まさに複数のパラメータの組み合わせそれぞれに対してSVMの学習を繰り返すため，SVMの学習高速化が重要となる．

SVMは，サポートベクトルと呼ばれるクラス境界付近のパターンのみを用いて識別器を構成する．よって，学習パターンからサポートベクトルの候補となるパターンを適切に選択できれば，SVM識別器の認識率を低下させることなくSVMの学習を高速化できると期待される．

学習パターンの事前選択によるSVMの高速化手法は大きく二つに分類できる．第一は，事前選択の有無によって選択されるサポートベクトルが変化しない手法 (safe method) である．第二は，事前選択の有無によって選択されるサポートベクトルが変化する手法 (non-safe method) である．

2.4.1 Safe methodによるSVMの学習高速化

学習パターンの事前選択を行うsafe methodとして，Ogawaら [87]の手法が知られている．Ogawaらは，pathwise SVMにおいてサポートベクトルとして選択されないパターンを検出し，あらかじめ学習パターンから削除することでSVMの学習を高速化している．しかしながら，サポートベクトルとして選択されるパターンは

学習に用いるパラメータによって変化する．よって，この手法にはパラメータごとに学習パターンの事前選択を行う必要があるという課題がある．

Safe method による SVM の学習高速化には，特徴量の次元削減というアプローチも考えられる．例えば，Ghaoui ら [88] は，テキスト分類の課題においてクラス識別に寄与しない情報を削除することで特徴量の次元数を削減し，SVM の学習を高速化している．しかしながら，次元削減を行う手法は，削減する特徴量の次元の探索と SVM 識別器によるクラス識別を同じ特徴空間中で行う必要があるため，カーネル型の SVM に適用できないという課題がある．

2.4.2 Non-safe method による SVM の学習高速化

分布解析による学習パターンの事前選択

学習パターンの事前選択にも分布解析の手法が応用されている．分布解析による事前選択の基本的な考え方は，クラス境界付近のパターンの選択である．これは前述のように，SVM がクラス境界付近のパターンのみをサポートベクトルとして識別器を構成するためである．この事前選択は，逆にクラス境界付近ではないパターンを削除する手法と見てもよい．具体的な手法として，クラスタ分析を用いる手法 [89, 90] や異クラスのパターンまでの距離を基準とする手法 [91] をはじめ，多くの事前選択法 [92, 93] が提案されている．

ところで，事前選択法自身に何かしらのパラメータが必要になるのは得策ではない．すなわち，そのパラメータの最適設定に別途計算が発生し，結果的に事前選択による計算量軽減の効果が薄れてしまう．実際，上述した事前選択法は，それぞれ固有のパラメータを持っている．よって，そのパラメータの最適化に関する処理時間を含めて事前選択の効果を議論すべきと思われるが，文献中にそうした記述はなされていない．

ネットワークによる学習パターンの事前選択

近接性グラフを用いた学習パターンの事前選択法も提案されている．近接性グラフによる事前選択の特徴は，事前選択のためにパラメータを必要としないことである．例えば，Gabriel グラフを用いる手法 [94, 95] や RNG を用いる手法 [92, 93, 95, 96] などが提案されている．いずれも，グラフにより表現されたパターンの分布構造に基づいて，学習パターンを事前選択する．

このグラフ化にはパラメータが不要なため，結果的にパラメータフリーな事前選択法を構築できる．よって，パラメータ調整のための新たな計算は発生しない．しかしながら，SVM の学習高速化を目的とする場合，事前選択に必要な処理を含めた処理時間で高速化の効果を評価し，事前選択の有無による認識率の変化を評価して総合的に手法の有効性を議論すべきと思われる．この点では，Gabriel グラフによる事前選択は，その密なグラフ構造のために事前選択による学習パターンの削減効果が低く，高次元空間でのグラフ構築の処理時間やグラフ構造を保持するメモリ空間量も課題となる．また，上述した RNG を用いた事前選択の先行研究では，学習パターン数が 100〜4,500 程度の小規模パターンを用いた実験結果が示されているに留まっており，学習パターン数が大規模化した場合の事前選択による SVM 識別器の認識率の変化についての考察は行われていない．

第3章

実験試料

3.1　解析対象例としての数字画像

1.2で述べたとおり，パターンの真の分布を解明するためには，解析対象のパターンを大規模化する必要がある．この条件を満たすために，本研究では16 × 16画素の2値数字画像を対象とした分布解析を行う．実際に用いたデータセットについては，次節で詳説する．

2値数字画像を解析対象とする利点は次の三点である．

1. 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．1.2.1でも述べたように，比較的低次元の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．

2. 数字画像は，一般物体画像に比べてデータエントリ企業との連携などにより入手しやすく，クラスごとに十分なパターン数を確保することが容易である．クラスごとに十分なパターン数を確保すれば，パターン数の多寡による分布構造の変化が解析可能となる．

3. 数字であれば，各パターンに付与されたカテゴリ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，256次元のハイパーキューブの頂点にのみ存在する．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．

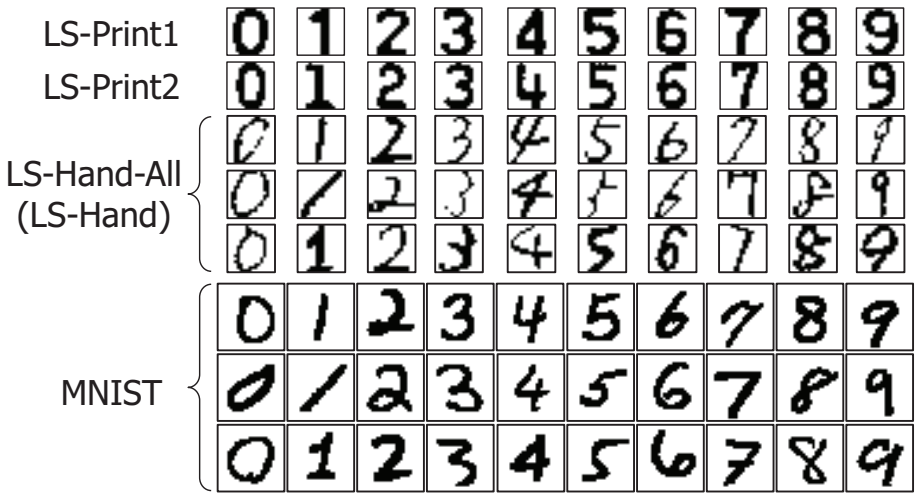


図 3.1: 数字画像のパターン例

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．

本研究で使用する大規模な数字画像データセットは独自に収集したものであるため，他のデータセットでも実験を行い，提案手法の有効性についての信頼性を高める必要がある．そこで，第 5 章および第 6 章では，公開データセットである MNIST [20] を用いて追加の実験を行い，提案手法の有効性を検証する．

3.2 三つの数字画像データセット

本研究では，分布の異なる三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，独自に収集した 535,494 サンプルの活字数字画像データセット（以下，LS-Print）と 822,714 サンプルの手書き数字画像データセット（以下，LS-Hand-All），および公開データセットである MNIST の手書き数字画像データセット（以下，MNIST）を使用して実験を行う．図 3.1 に本研究で使用する数字画像の一例を示す．

LS-Print と LS-Hand-All に含まれる画像は，16 × 16 画素の 2 値画像である．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” までのクラスラベルを付与している．また，MNIST は，公開されている 28 × 28 画素の 256 階調画像を実験に使用する．

LS-Print 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，2 つのサブセット (LS-Print1, LS-Print2）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，各サブセットのサンプル数は，LS-Print1 が 259,153 であり，LS-Print2 が 276,341 である．

LS-Hand-All 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，その他のクラスが約 6 万ずつである．第 5 章および第 6 章での実験では，異クラス間の関係性をエッジ数の多寡や SVM 識別器の誤認識数などにより，クラスごとに定量的に解析する場合がある．その際，特定のクラスにパターン数の偏りがあれば，クラスごとの定量的な解析結果を単純に比較できない．そこで，第 5 章および第 6 章での実験では，LS-Hand-All から各クラスのパターン数が等しくなるようにサンプリングしたサブセット（以下，LS-Hand）を用いて実験を行う．LS-Hand のサンプル数は 518,850 である．

3.3 画像特徴

本研究で使用する特徴量は，各画像の輝度値による輝度特徴と文献 [97, 98] の線素方向特徴（以下，方向特徴）を用いる．輝度特徴は，各画像の輝度値をひとつの特徴ベクトルとする．よって，輝度特徴の特徴ベクトルは，LS-Print と LS-Hand-All の 16 × 16 画素の画像では 256 次元，MNIST の 28 × 28 画素の画像では 784 次元となる．方向特徴は，入力画像を 64 × 64 画素に拡大して輪郭を検出し，垂直，右下，

水平，左下のそれぞれの方向に対応した特徴画像を作成し，各特徴画像に対してサンプリングを行って得られた 1024 次元の特徴量をひとつの特徴ベクトルとする．

輝度特徴は非常に単純な特徴量である．このため輝度特量によるパターンの分布は，画像中での文字の位置ずれなどの影響により，クラスごとに広範な分布構造を持つと予測される．一方，方向特徴は文献 [97, 98] で文字認識を目的として使用されている特徴量である．方向特徴では，画像中での文字の位置ずれなどの影響が特徴抽出時に吸収されるため，パターンの分布が輝度特徴に比べてクラスごとに密な分布構造を持つと予測される．本研究では，このように傾向が大きく異なると予測される 2 つの特徴量による大規模パターン集合の分布構造を解析し，比較することで提案手法の基本的な特性と有効性を検証する．

3.4 パターン間距離

第 4 章での実験では，パターン間の距離尺度としてハミング距離を用いる．本研究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．

第 5 章および第 6 章での実験では，パターン間の距離尺度としてユークリッド距離を用いる．本研究で提案するネットワーク表現による大規模パターン集合の分布構造解析は，ハミング距離に限らず，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．その実例として，第 5 章および第 6 章ではパターン間の距離尺度としてユークリッド距離を適用し，2 値数字画像の分布解析だけでなく 256 階調の数字画像である MNIST の手書き数字画像データセットの分布解析を行う．

第4章
MSTによるパターン分布構造の解析

4.1　まえがき

本章では，MST [12] を用いたネットワーク表現による大規模パターン集合の分布構造解析法を提案する．各パターンを1ノードとし，パターン間の距離をエッジの重みとして構築したMSTは，パターン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．提案手法は，パターン空間中での分布構造を表現したMSTを用いてパターンの分布構造を解析する．MSTは，木構造という制約により単純なネットワークとなるため，高速に導出が可能である．また，MSTは，全域木という性質によりパターンの分布構造を1つのネットワークで表現できるため，局所的なパターンの近傍関係だけでなく各クラスごとあるいは大規模パターン集合全体の分布構造も解析できる．本章では，提案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とで，提案手法による分布解析の妥当性を検証する．

4.2　MSTによるパターン分布のネットワーク表現

本章では，ネットワークとして2.2.4で述べたMSTを用いて大規模パターン集合の分布構造を表現する．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．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4.1に示す．

2.2.4でも述べたように，パターン間の距離を基準にMSTを作成することで，パターン空間における近傍関係を保存できる．具体的には，まずMSTは最近傍グラフを内包しており，局所的なパターンの近傍関係を表現できる．パターン間の距離

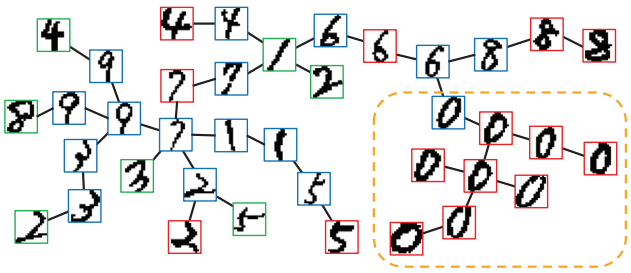


図 4.1: MST の例

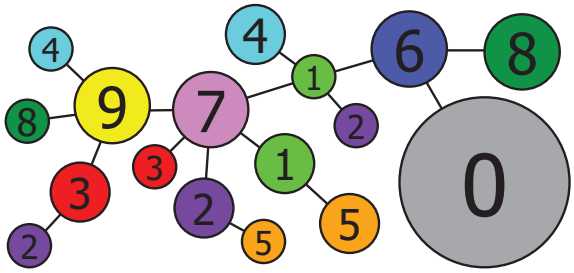


図 4.2: クラスタ木の例

をエッジの重みとして MST を構築すれば，最近傍ノード，すなわちあるパターンから距離が最も近いパターンに対してエッジが必ず付与される．次に，全域木である MST は，パターンの分布構造を 1 つのネットワークで表現できるため，局所的な分布構造だけでなく各クラスごとあるいは大規模パターン集合全体の分布構造も解析できる．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できることは，パターンの分布構造の解析手法として有効である．

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 4.1 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．

加えて本章では，MST を 1.2.2 で述べたクラスタ化，すなわち図 4.1 に例示した互いにエッジで接続される同じクラスラベルを持つノード群を統合したネットワー

ク表現による解析も行う．本論文では，このクラスタ化されたMSTをクラスタ木と呼ぶ．図4.1のMSTから生成したクラスタ木を図4.2に示す．ここで，各ノードの大きさはクラスタに含まれるパターン数に比例している．クラスタ木においてクラスタ間のエッジはMSTのエッジでもあるため，エッジで結合されるクラスタの分布はパターン空間中で近傍関係にある．すなわち，クラスタ木によるネットワーク表現はクラスタ同士の近傍関係を保持する．以上の特長により，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とらえられると期待できる．

MSTによる分布構造表現には，次の四つの特長がある．

1. 最近傍ノードに対してエッジが必ず付与されるため，局所的なパターンの近傍関係を表現できる．例えば，パターン変形の核となるようなパターンは，その周囲に多数の類似パターンが分布することになるため，そのパターンに対応するノードは多数のエッジを持つ．

2. 全域木であるため，大局的に見てパターン空間全体の構造を表現できる．すなわち，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表現できる．

3. 局所的に見ると類似画像がエッジで結合されるため，MSTをクラスタ化してクラスタ木を構成すれば，パターンの特徴的な分布構造が表現できる．

4. 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[40]，ネットワークの構造を記述するために必要なメモリ空間量が少ない．

なお，4.4および4.5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ランダムに選択する．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．

4.3　実験試料

本章では，分布の異なる二つの数字画像データセットの分布構造を解析し，その結果を比較することで提案手法の妥当性を考察する．具体的には，大規模活字数字画像データセットとしてパターン数が 535,494 である LS-Print，大規模手書き数字画像データセットとしてパターン数が 822,714 である LS-Hand-All の分布構造を解析する．各データセットの画像例は，すでに図 3.1 に示したとおりである．また，特徴量は輝度特徴と方向特徴を使用し，パターン間の距離定義はハミング距離を用いる．

4.4　分布解析結果：分布構造の定量化

まず，MST のノード次数とエッジ重みに着目し，LS-Print と LS-Hand-All 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持つエッジの数は，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．

4.4.1　ノード次数の頻度分布

図 4.3 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，LS-Print，LS-Hand-All 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．前述の通り，パターン数の少ない実験では，パターンをランダムに選択し，使用するパターンの選択と MST 生成を複数回繰り返して結果を確認したが，個々の実験結果での頻度分布の差異は少なく，パターン数の増加により次数の大きなノードが出現しやすくなる傾向は共通であった．

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の増加は，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

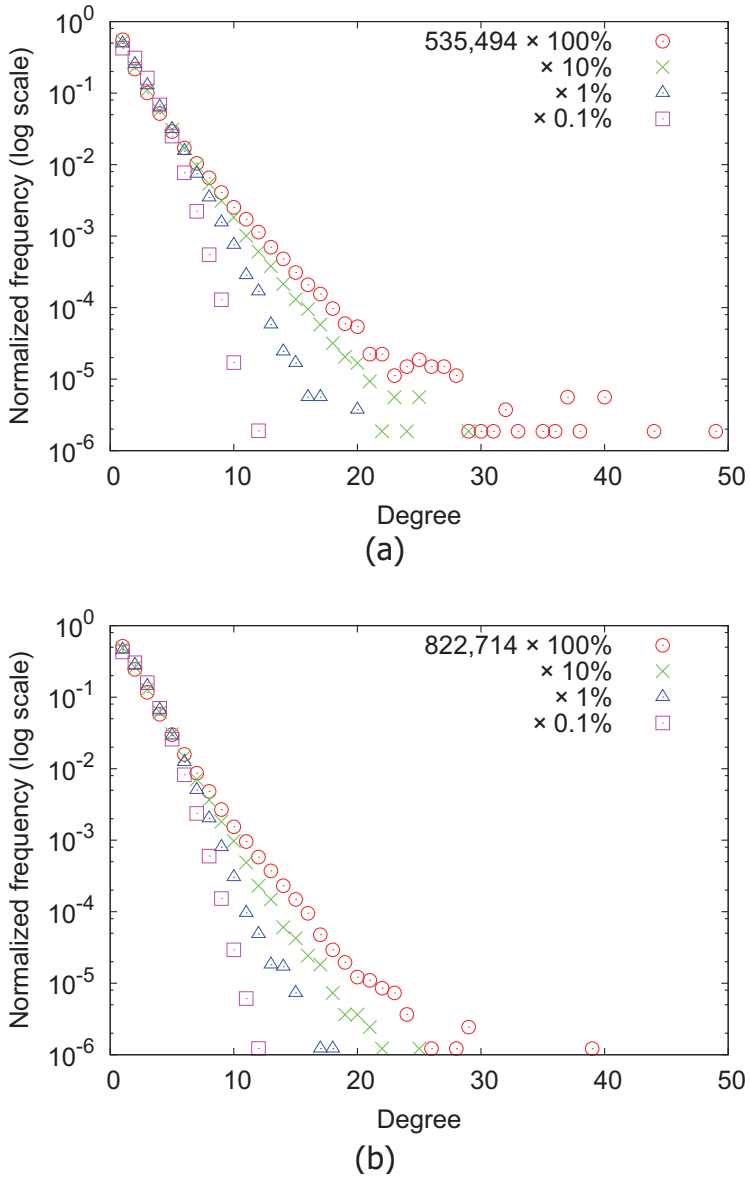


図 4.3: ノードの次数と出現頻度の関係． (a) LS-Print. (b) LS-Hand-All

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解析した．

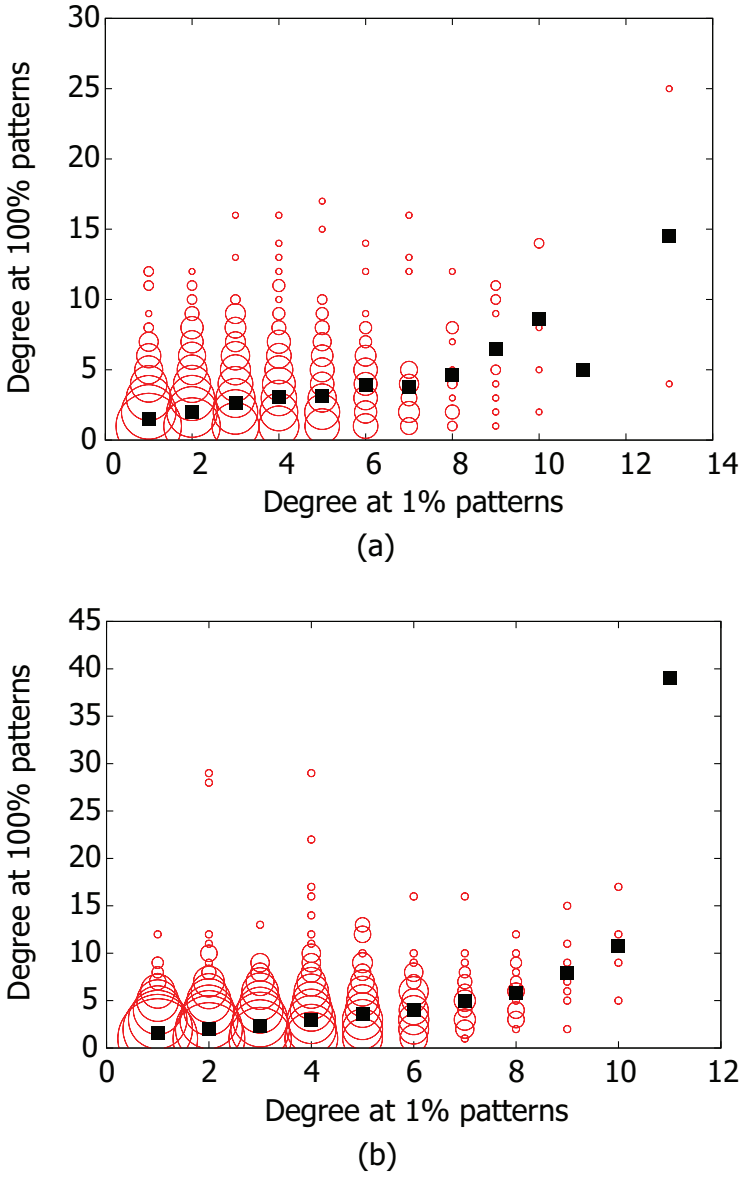


図 4.4: パターン数の増加による次数の変化例 （■：平均）．(a) LS-Print. (b) LS-Hand-All

図 4.4 は，LS-Print と LS-Hand-All のある 1 % 個のパターンを用いた MST において，ある次数を持っていたノードが，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．さらに，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次数変化の平均値を表す．すなわち，1 % 個のパターンを用いた MST での次数ごとに，全パターンを用いた MST での各ノードの次数を平均した値である．

図 4.4 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．例えば，図 4.4(b) から，LS-Hand-All 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 4.3(a)，(b) を比較すると，活字 (LS-Print) よりも手書き数字 (LS-Hand-All) の方が，パターン数の増加により特定のノードの次数が高くなる場合が多い．手書き数字では，異体字のようなパターン変形の核となりうるパターンの存在が予想されるため，このノード次数の傾向は，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，大規模パターン集合を用いて分布構造を解析することの効果である．

4.4.2　エッジ重みの頻度分布

図 4.5 に，パターン数を変えた際の，エッジ重みの頻度分布の変化を解析した結果を示す．同図 (a)，(b) より，LS-Print，LS-Hand-All ともにパターン数が増えることでエッジ重みが小さくなる，すなわちパターン間の距離が近くなる．この実験でもノード次数の解析実験と同様に，選択されるパターンによるエッジ重みの頻度分布の差異は少なく，パターン数の増加によりエッジ重みが小さくなる傾向は共通であった．

図 4.5(a)，(b) を比較すると，エッジ重みの平均と分散は，LS-Print のほうが LS-Hand-All 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

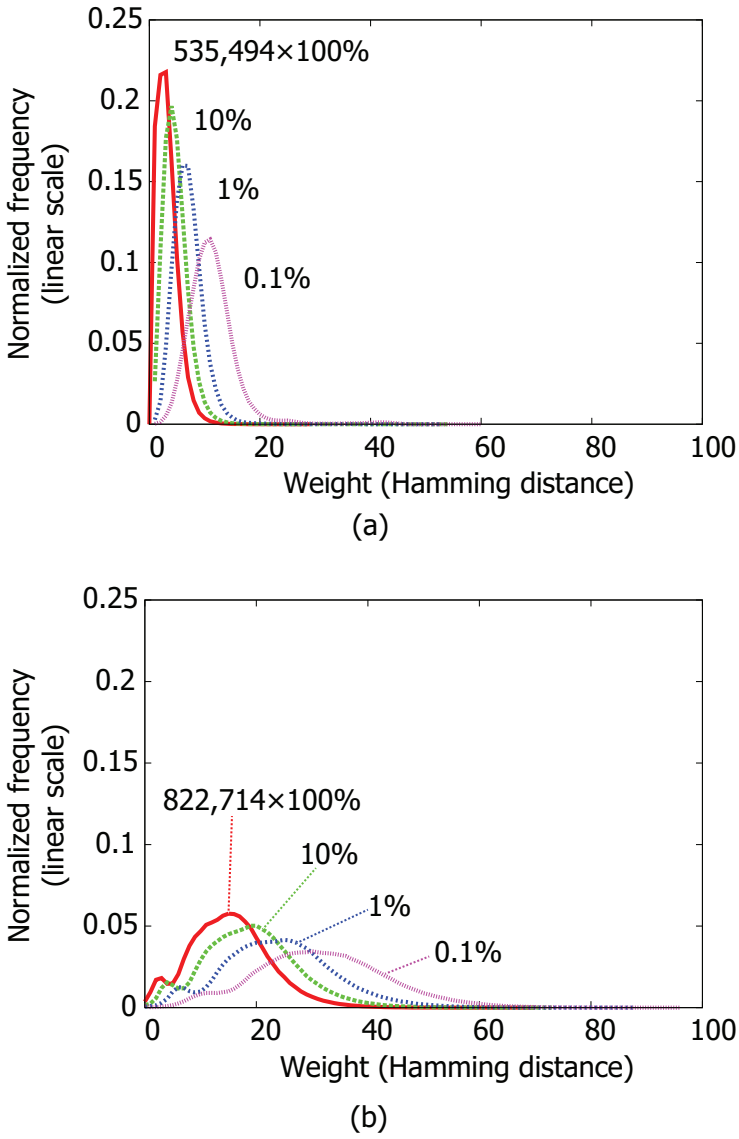


図 4.5: エッジ重みと出現頻度の関係．(a) LS-Print．(b) LS-Hand-All

さくなっている．重みはパターン間の距離であるから，このエッジ重みの傾向は，手書き数字 (LS-Hand-All) よりも活字 (LS-Print) 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，クラスごとに密になっていく様子を示している．なお，図 4.5(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，パターン数の増加とエッジ重みの分布の変化の関係を LS-Print，LS-Hand-All の各々について解析したところ，平均と標準偏差がパターン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．

4.5 分布解析結果：クラスごとの分布構造の定量化

次に，MST の隣接ノードのクラスに着目してパターンの分布構造を解析する．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，すなわち各クラス分布の中心あるいは周縁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，クラスごとの分布構造を解析できる．

4.5.1 隣接ノードのクラスによるノードの分類

図 4.6 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，隣接するノードが持つクラスラベルが，

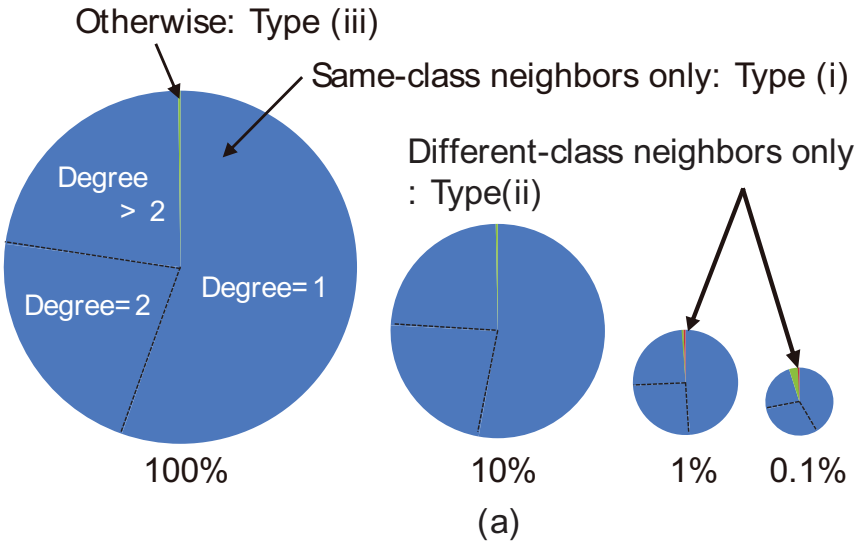
タイプ (i): 全て自ノードと同じクラスである，

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 3 つの場合の割合を示す．

図 4.1 の MST の例では，各パターンの枠線の色が各タイプを表している．具体的には，赤色がタイプ (i)，緑色がタイプ (ii)，青色がタイプ (iii) のパターンである．ここで，タイプ (ii) 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



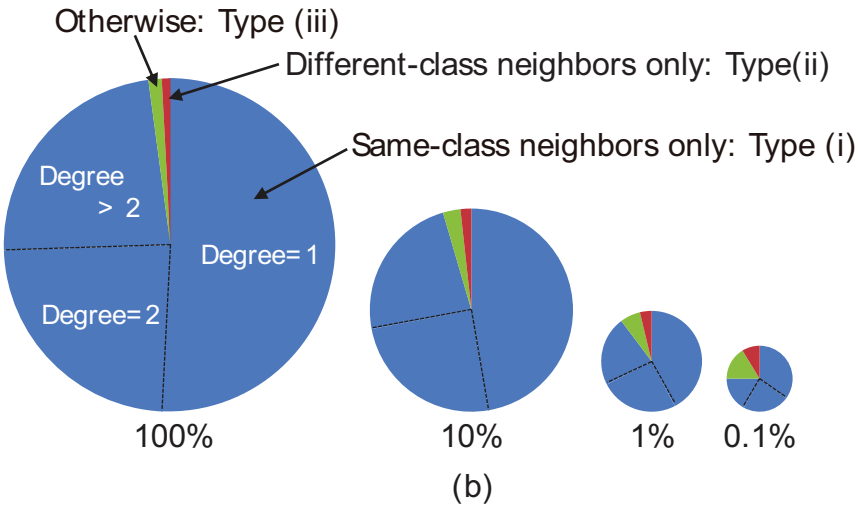


図 4.6: 隣接ノードのクラスラベルによる分類を行ったときの各ノードの割合．(a) LS-Print．(b) LS-Hand-All

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，およびその他（次数が 2）のノードの割合について調べた．

図 4.6 より，ほとんどのノードがタイプ (i) であることが確認できる．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，パターン数の増加にともなって増えていることは，4.4.1 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集中して分布しているような分布構造を持つことを示している．さらに，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，LS-Print 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 4.6 (a)）．

図 4.7 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，手書き数字によらずどちらのタイプも，パターン数の増加にともない，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．LS-Print の実験では，10 % 個以上のパターン数での 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．一方，LS-Hand-All の実験では，パターン数が 10 倍となれば，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，LS-Hand-All の手書き数字について，パターン数が現状の 10 倍の約 800 万になった場合，アウトライアの割合はさら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 %$(= (20+0)/535,494)$，手書き数字については 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

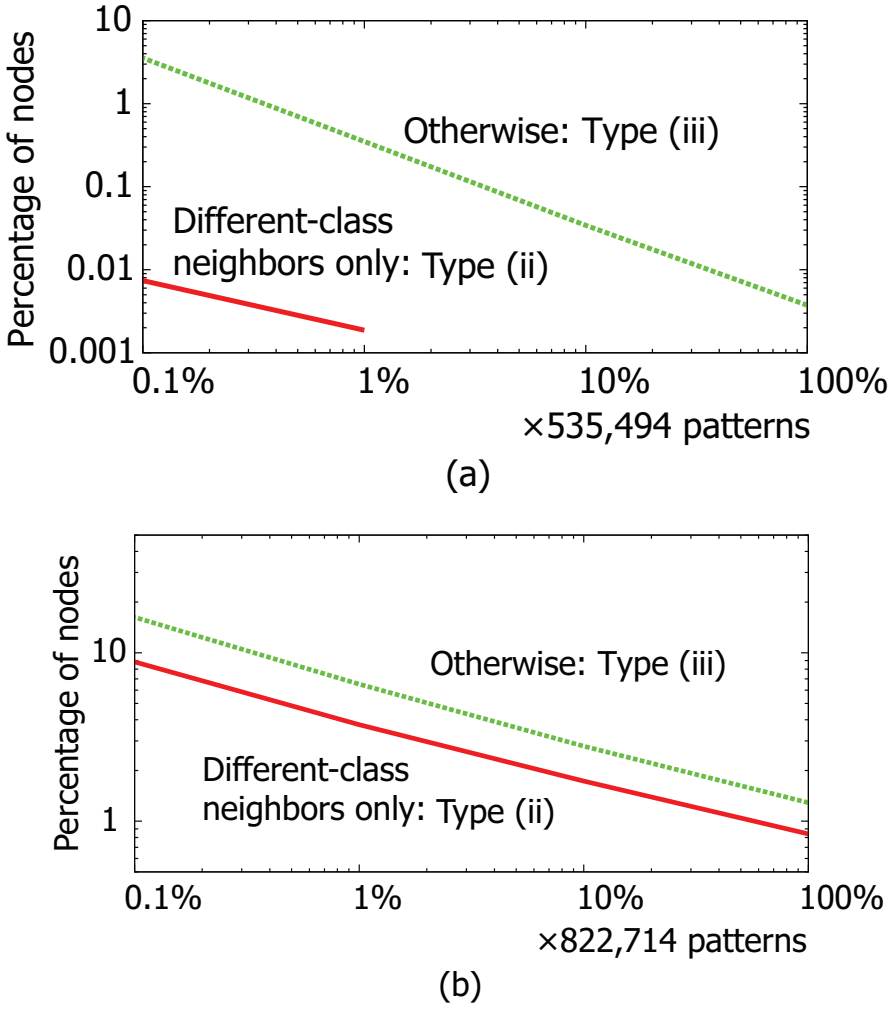


図 4.7: タイプ (ii) とタイプ (iii) の各ノードの割合．(a) LS-Print. (b) LS-Hand-All

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 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．

4.5.2 最近傍決定則による認識率との関係

本項では，前項で考察したLS-Hand-AllのMSTから得られるノードの分類の傾向（図4.7(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，あるノードが持つエッジののうち少なくとも1つは，最近傍パターンに接続されている．よって，異なるクラスと隣接するノードは，最近傍パターンのクラスラベルが自クラスとは異なるパターン，すなわち最近傍認識であれば誤認識されるパターンである可能性が高い．このため，文献 [8] 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，前項で考察したMSTの各タイプ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，LS-Hand-Allの最近傍決定則による認識率の評価実験を行った．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．全パターンを用いずに認識実験を行なう場合には，学習パターンに含まれない画像データから追加のパターンをランダムに選択し，学習パターン数を増加させ，実験を行った．

図4.8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4.7(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ともない，誤認識率は低下する傾向にあり，全パターン(822,714サンプル)を用いて実験を行なった際に，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．この傾向は，前項で述べたMSTのタイプ(ii)とタイプ(iii)のノードの出現傾向(図4.7)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 [8] でも指摘されている．

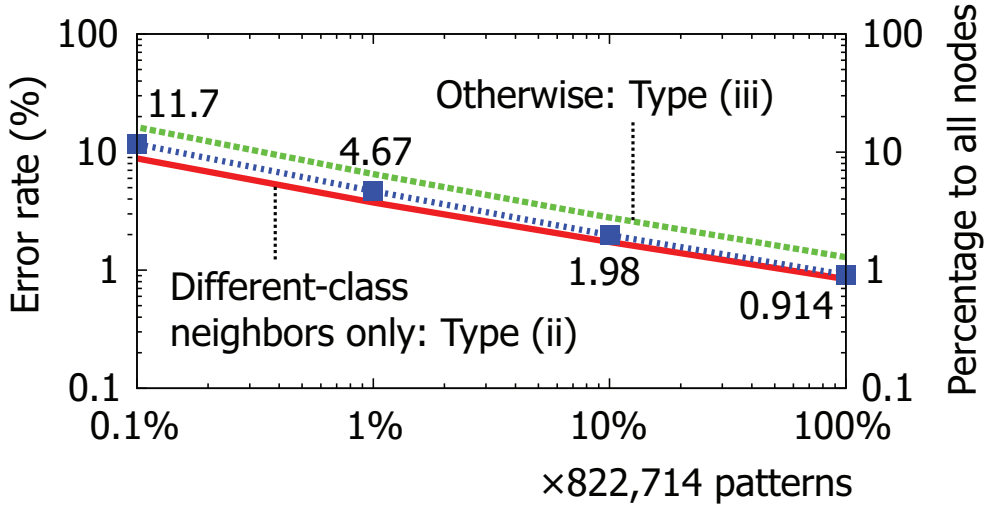


図 4.8: LS-Hand-All におけるパターン数と誤認識率，タイプ (ii) とタイプ (iii) の各ノードの割合の関係

4.5.3 最近傍パターンまでの距離の分布

図 4.9 は，前項で行った手書き数字認識実験での，入力パターンに対する正解クラスおよび不正解クラスそれぞれの最近傍パターンまでの距離を示している．学習パターンが増加するにしたがって，どちらも最近傍パターンまでの距離が減少しているが，正解クラスの最近傍パターンまでの距離の方が全体的に小さい．これは，クラス領域の内側の方が，クラス境界付近よりもパターン分布が密なためである．

特筆すべきは，約 80 万の大規模なパターン分布であるにもかかわらず，不正解クラスのパターンが距離 3（白黒の異なる画素が全体の 1 % あることに相当）を下限として，それ以上近距離に存在しないことである．これは，アウトライアも含めて，各クラスは他クラスと少なくとも一定距離以上離れた状態で存在することを示している．言い換えれば，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存在することが示唆される．

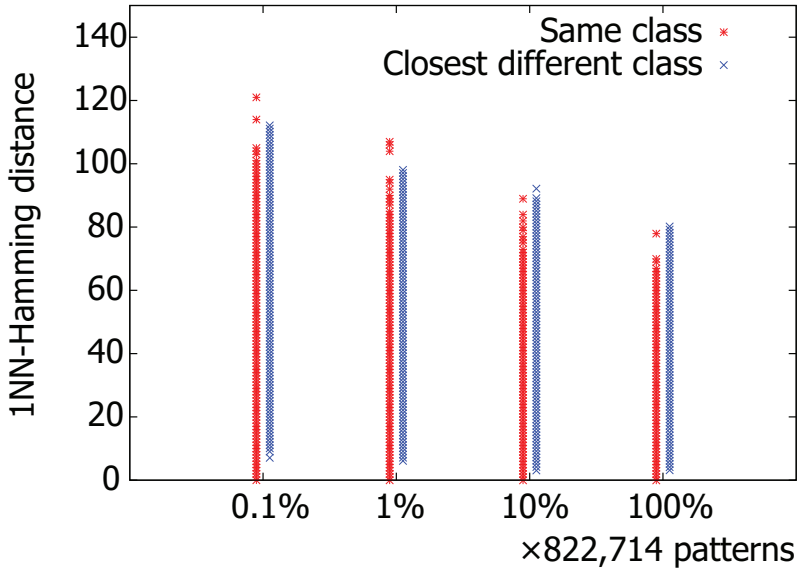


図 4.9: 最近傍パターンまでの距離

4.6 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，4.2 で述べたクラスタ木，すなわち MST においてエッジで結合される同じクラスのノードを統合した新たな木グラフを作成した．例えば，図 4.1 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 1 つのノードに集約表示される．

図 4.10 は，LS-Print，LS-Hand-All 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，一つの円がひと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，LS-Hand-All 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

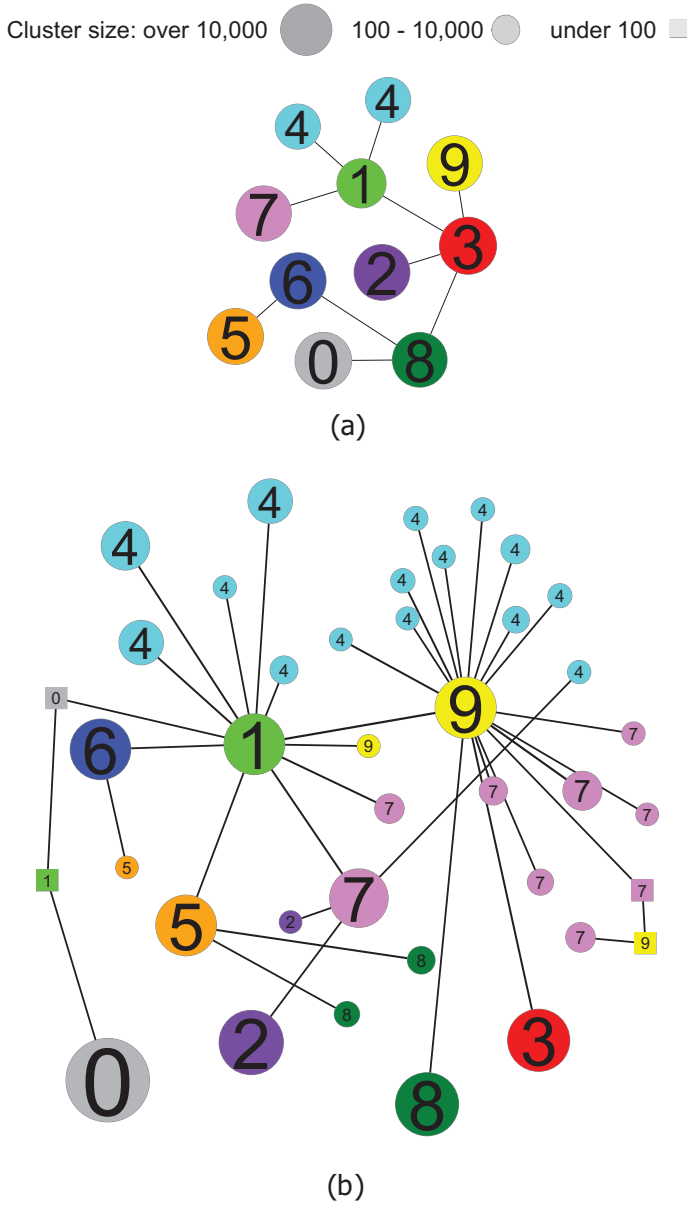


図 4.10: クラスタ木．(a) LS-Print．(b) LS-Hand-All

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，例外として区別するために四角で表記した．

図 4.10(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 3.1 のフォント形状を考えれば，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 4.10(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．この複数クラスタへの分断の程度は，クラスごとに傾向が異なっている．この傾向の差異は，分断化されたクラスが分断化されてないクラスに比べ，パターン分布に粗密が大きいために生じていると予測される．すなわち，複数クラスタへの分断が生じているクラスは，1 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．

図 4.10 に示した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，同図 (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確認できる．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” の形状が，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．数字画像パターンは，全体的に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粗密があることが，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表現によるパターンの分布構造解析の特徴である．

4.7　まとめ

本章では，MST を用いたネットワーク表現による大規模パターン集合の分布構造解析法を提案した．提案手法の特徴は，パターンの分布構造の解析を，MST におけるノードやエッジの特徴などから得られる指標を用いて定量的に行うことである．MST およ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．本章では，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に解析できる．

具体的な解析事例として，約 50 万の活字数字画像と約 80 万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．具体的には，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，大規模パターン集合の分布構造に粗密があることが確認できた．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粗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できる．

解析結果から，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．本章では 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，提案手法を適用可能である．

第5章
クラスタ化された相対近傍グラフ (Clustered-RNG) によるパターン分布構造の解析

5.1 まえがき

本章では第一に，RNG [13] をクラスタ化したネットワークを用いた大規模パターン集合の分布構造解析法を提案する．提案手法は，まず各パターンを 1 ノードとし，パターン間の距離を基準とした近傍関係によりエッジを付与して RNG を構築する．そして，その RNG をクラスタ化したネットワークを用いてパターンの布構造を解析する．ここでいうクラスタ化とは，1.2.2 で述べた粗視化処理である．具体的には，RNG のエッジで結合された同じクラスラベルを持つパターンを統合して 1 つのノードとして表現する．このクラスタ化された RNG (Clustered-RNG) により，単一のクラス分布だけでなく，多クラス分布間の隣接関係やアウトライアの分布などの関係性を同時に解析できる．本章では，提案手法の詳細を述べるとともに，活字および手書き数字画像の輝度特徴および方向特徴を対象とした実施例を示し，提案手法の有効性を検証する．

RNG は第 4 章で用いた MST を内包しており，木構造という制約を持つ MST に比べて RNG は閉路を許すグラフ構造であるため，パターン間の近傍関係を余さず表現できる．後述のように，Clustered-RNG は，RNG の特長を継承するだけでなく新たな特長を持つため，MST では解析できなかった分布構造の特徴を新たに解析できる．提案手法である Clustered-RNG によって新たに解析できる分布構造の特徴は，次の三つである．

1. 識別境界付近のパターン数の多寡，すなわちパターン空間中でのクラス間分布

の粗密が解析できる．

2. クラス間のオーバーラップの様子が解析できる．

3. 最近傍関係にないクラス間の隣接性が容易に解析できる．

本章では第二に，大規模パターン集合に対する効率的なRNG構築法を提案する．大規模パターンの分布構造をClustered-RNGによって解析する場合，Clustered-RNGの構築に必要な計算時間とメモリ空間量が，実用上の課題となる．ここで，Clustered-RNGの構築に必要な計算時間は，ほぼRNGの構築に必要な処理時間となる．よって，高次元空間中の大規模パターン集合に対してRNGを高速かつ省メモリに構築する手法についても提案し，その有効性を検証する．

5.2 RNGとClustered-RNGによるパターン分布のネットワーク表現

本章では，ネットワークとして2.2.5で述べたRNGを利用し，そのRNGをクラスタ化したClustered-RNGによってパターン分布を表現する．具体的には，各パターンを1ノードとし，パターン間のユークリッド距離をエッジの重みとしてRNGを構成する．そして，そのRNGをクラスタ化，すなわちエッジで結合された同じクラスラベルを持つパターンを統合することでClustered-RNGを構築する．

5.2.1 RNG

RNGは，あるノード対が相対的に近傍である場合に，そのノード間にエッジを与えることで得られる無向グラフである．RNGは，最近傍グラフやMSTを内包しており，これらのグラフ表現に比べてノード間の近傍関係をより明示的に表現できる．これらの特長により，RNGはパターン分類や地理分析などの研究にも広く応用されている [38].

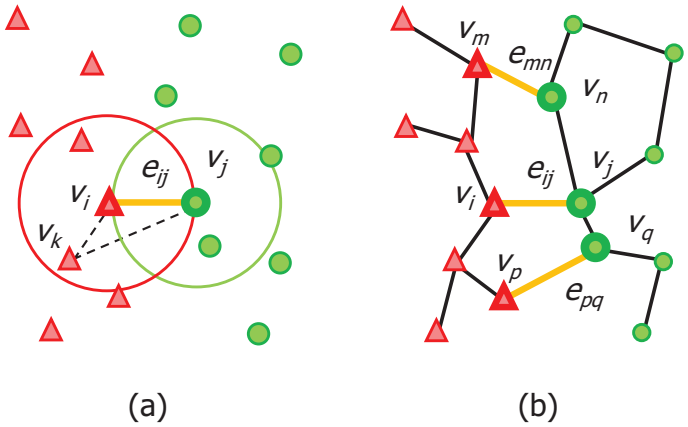


図 5.1: RNG の構成条件．(a) ノード対にエッジを付与する条件．(b) RNG の構成例

RNG のグラフ構造は，以下のように定義される．RNG を構成する各ノードを $\boldsymbol{V} = \{v_1, v_2, \ldots, v_n\}$，各エッジを $\boldsymbol{E} = \{e_{ij}\}$ とおく．各エッジ e_{ij} は，$d(v_i, v_j) \leq \max\{d(v_i, v_k), d(v_j, v_k)\}$，$(k = 1, 2, \ldots, n, k \neq i, k \neq j)$ を満たす．ここで $d(v_i, v_j)$ は，v_i と v_j との距離である．すなわち，図 5.1(a) に示したようなある 2 ノード v_i と v_j について，それぞれを中心とする 2 つの超球が重なる空間に他のノード v_k が存在していない場合に限り，v_i と v_j の間にエッジ e_{ij} が付与される．図 5.1(a) のノード分布に対する RNG を同図 (b) に示す．本章では，各ノードは数字画像の d 次元特徴ベクトルに対応し，ノード間の距離定義はユークリッド距離を用いる．図 5.2 に少数の手書き数字画像で作成した RNG の一例を示す．

RNG は，次の三つの性質により，クラスラベルが付与された大規模パターン集合の分布解析に適している．

1. 局所的に見ると，類似画像の近傍性が保持される．類似画像に対応するノードは，短いエッジで密に結合されるため，同じクラスのノードは互いに結合され，クラスタを構成する．

2. 異クラス間についても近傍関係をエッジとして保存するため，各クラス間の隣接関係を保持する．

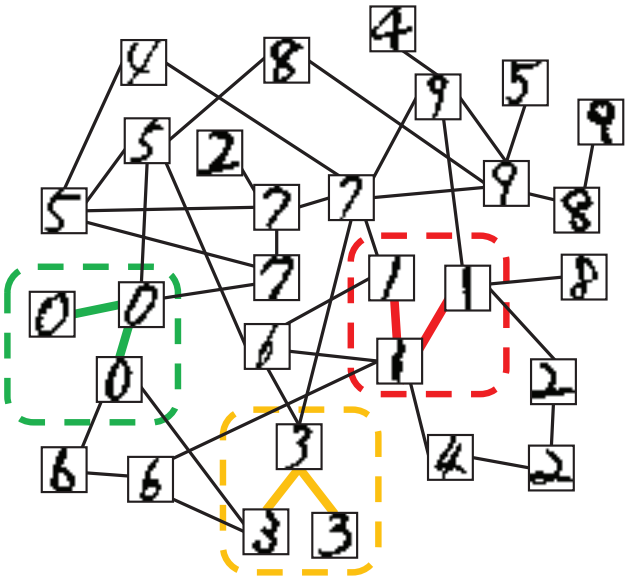


図 5.2: 少数サンプルによる RNG の例

3. 相対的な近傍関係にあるノード間にのみエッジを持つため，大量のノードに対して粗なネットワークとなり，記述に必要なメモリ空間量が少ない．

特に，上述した第二の性質によって，RNG は第 4 章で用いた MST に比べ大規模パターン集合の分布構造を表現する手段として適している．この性質は，RNG のエッジが距離の遠いノード間にも付与されうるという特長によるものである．RNG では，相対的な近傍関係を基準にエッジを付与するため，必ずしも近傍にあるノード間のみにエッジが形成されるわけではない．例えば図 5.3 に示すように，距離の遠いノードであってもその間の領域に他のパターンがなければエッジが付与される．すなわち，相対的に近傍関係にあるパターン間の隣接性を完全に保存できる．この特長により，RNG による分布解析では，分布の離れた異クラスの間の隣接関係やクラス境界付近のパターン数などの隣接性もそのネットワークによって保存される．この点については，5.2.3 で詳説する．

RNG による解析では，上述した相対的に近傍関係にあるパターン間の隣接性を保存できるという特長により，クラス間の隣接関係が明らかとなる．クラス間の隣接

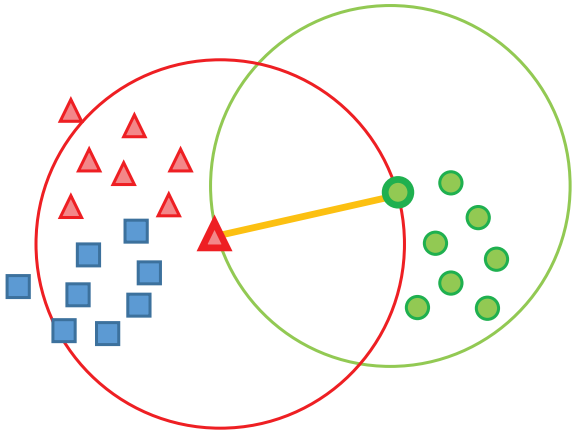


図 5.3: 距離が遠いノード間に付与される RNG のエッジ

関係を解析する簡便な方法として，最近傍法による混同行列を用いることが考えられる．しかしながら，混同行列では誤認識が発生しないクラス間（例えば図 5.3 に△と○で示した近傍性の低いクラス間）の隣接性を解析できない．

RNG による分布構造解析の特長は，大規模パターン集合の隣接関係を完全に保存していることである．よって，大規模パターン集合の分布構造を RNG を用いて可視化すれば，大局的な分布構造と局所的な分布構造を同時に解析できる．実際に，図 5.2 では，RNG の各ノードをパターンに対応させることで，パターンの分布構造を可視化できている．しかしながら，パターン数が増加すれば複数のエッジが複雑に交差するため，RNG による可視化結果は非常に複雑なものとなる．

5.2.2 Clustered-RNG

提案手法である Clustered-RNG は，RNG のエッジで結合された同じクラスラベルを持つパターンを統合して 1 つのノードとして表現することで，分布構造のネットワークをクラスタ化したものである．図 5.4(a) は，図 5.2 の RNG をクラスタ化した Clustered-RNG である．図 5.2 の RNG において，クラス “0”，“1”，“3” などのパターンを統合することで，図 5.4(a) に示した Clustered-RNG が得られる．ここで，

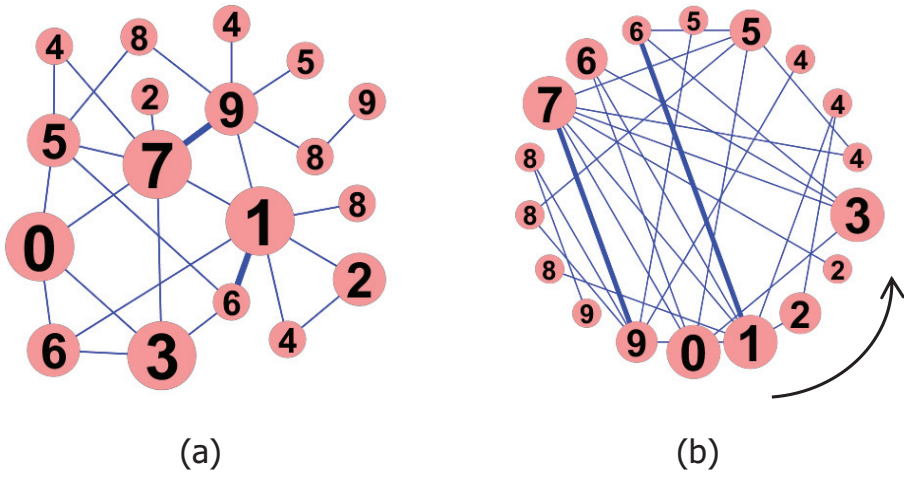


図 5.4: Clustered-RNG によるネットワーク表現（ノードのサイズ：クラスタに含まれるパターン数，エッジの太さ：クラスタ間のエッジ数）．(a) Clustered-RNG. (b) Clustered-RNG の円環表示

各ノードの大きさはクラスタに含まれるパターン数に比例しており，各エッジの太さはクラスタ間のエッジ数に比例している．本研究では，Clustered-RNG を図 5.4(b) のように各ノードをクラスラベルの順に反時計回りに円環表示し，パターンの分布構造を可視化する．また，ノードの大きさとエッジの太さは，各 Clustered-RNG におけるクラスタに含まれるパターン数とクラスタ間のエッジ数の相対的な多寡が可視化できるよう，個別にスケーリングを行う．

Clustered-RNG は，前述の RNG の特長を継承するだけでなく新たな特長も持っている．Clustered-RNG による分布構造のネットワーク表現の特長は，以下の二点である．

第一に，Clustered-RNG は多クラス間の隣接性をエッジの太さで明示している．これにより，クラスタ間エッジの多寡による識別境界付近のパターン数の多寡の解析が可能となる．これは 5.1 で，提案手法によって解析できる分布構造の特徴として第一に述べたものである．具体的には，Clustered-RNG のエッジの太さは，クラス間の境界付近に分布するパターン数に比例する．よって，エッジの太さで表現さ

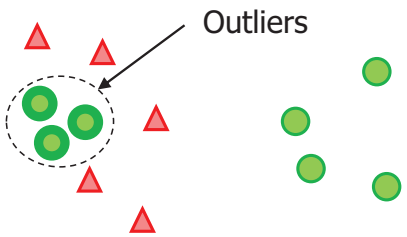


図 5.5: Clustered-RNG によって解析できるアウトライアの例

れる多クラス間の隣接性から識別境界付近のパターン数の多寡を容易に解析できる．すなわち，パターン空間中でのクラス間分布の粗密が解析できる．

このパターン空間中でのクラス間分布の粗密は，Clustered-RNG による可視化結果から容易に把握できる．RNG のクラスタ化では，異クラス間の隣接関係は失われない．また，このクラスタ化により，1.2.1 で述べた RNG ではパターン数が増加すれば複数のエッジが複雑に交差し，可視化結果が非常に複雑となるという課題も解決できる．

第二に，Clustered-RNG は他のクラスにオーバーラップして分布するパターン（以下，アウトライア）を，1 つのクラスタとして検出できる．これにより，クラス間のオーバーラップの様子の解析が可能となる．これは 5.1 で，提案手法よって解析できる分布構造の特徴として第二に述べたものである．

特定クラスのパターンが集中して分布している箇所では，同じクラスラベルを持つパターンが互いにエッジで結合され 1 つのクラスタ（以下，主要クラスタ）を構成する．一方，図 5.5 に示したようなアウトライアは，同じクラスラベルを持つ主要クラスタのパターンとはエッジで結合されないため，主要クラスタとは別のクラスタ（以下，アウトライアクラスタ）を構成する．このアウトライアクラスタにより，他のクラスにオーバーラップして分布する単一のパターンだけでなく，小規模なパターン群をも検出できる．よって，各クラスタのエッジを解析すれば，クラス間のオーバーラップの様子が解析できる．

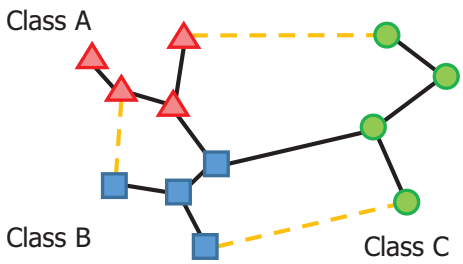


図 5.6: Clustered-RNG により解析できるクラス間の隣接性

5.2.3 Clustered-RNG による分布構造解析の特長

Clustered-RNG による分布構造解析の特長は，次の三点である．第一に，Clustered-RNG は，4.6 で述べた MST によるクラスタ木で解析した多クラス間の隣接関係をより明示的に表現できる．具体的には，5.1 で提案手法よって解析できる分布構造の特徴として第三に述べたように，最近傍関係にないクラス間の隣接性が容易に解析できる．これは，RNG がエッジの重みの総和が最小となるように全てのノードをエッジで結んだ MST を内包しており，MST では表現できない多クラス間の隣接関係を表現できるためである．例えば，図 5.6 のノード分布に対する RNG において，MST では同図に黄色の破線で示したエッジが付与されない．よって，MST ではクラス A，C 間の隣接関係が表現できず，MST によるクラスタ木では多クラス間の隣接関係を十分に解析できない．すなわち，MST を用いた手法で解析できる分布構造は，最近傍のパターン間の関係性にほぼ限定される．一方，Clustered-RNG は前述した RNG が持つ距離の遠いノード間にもエッジが付与されうるという特長を継承しているため，多クラス間の隣接関係を完全に保持できる．

Clustered-RNG が保持する識別境界付近のパターンとその隣接関係は，多クラス識別を行ううえで有用である．例えば，大規模パターン集合を学習データとする一対他識別器の学習において，学習データを RNG によって解析し，自クラスのパターンと RNG のエッジで結合されないパターンを事前に削除して学習すれば，識別器の認識率を大きく低下させることなく学習を高速化できる．この時，事前選択に必

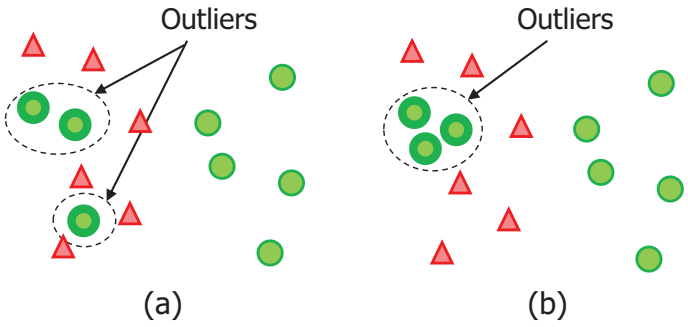


図 5.7: アウトライアの分布．(a) 3-近傍ノードによって検出されるアウトライア．(b) 3-近傍ノードでは検出できないアウトライア

要な情報は Clustered-RNG でも保持できるため，RNG の解析結果を直接用いる場合と比較して，結果を保持するためのメモリ空間量が削減できる．この点については次章で詳説する．

第二に，Clustered-RNG のエッジの太さにより解析できる特徴空間中でのクラス間分布の粗密は，特徴量選択の指標として利用できる．クラス間分布が粗となる特徴量を用いて学習した識別器の認識率は高くなると期待できる．クラス間分布の粗密を解析する手段として，最近傍法などによる識別器を構成し，その混同行列を用ることが考えられる．しかしながら，図 5.6 のような分布については，最近傍法のように単純な手法でも誤認識が発生しないため，混同行列ではクラス間分布の粗密を解析できない．

第三に，Clustered-RNG のアウトライアクラスタのパターンを解析すれば，図 5.7 に示すような他のクラスの分布にオーバーラップして分布するアウトライアを効率的に抽出できる．これは，k-近傍ノードを利用したアウトライア検出よりも優れていると考える．例えば，図 5.7(a) に示したパターン分布では，$k = 3$ とすれば 3-近傍ノードのクラスラベルの最頻値によりアウトライアを検出できる．しかしながら，同図 (b) のような分布の場合には，アウトライアを検出できない．これに対して，Clustered-RNG による解析では，エッジで接続された同じクラスラベルを持つパターンが統合されるため，各クラスタに含まれるパターン数から図 5.7 に示すよ

うなアウトライアを確実に検出できる．

大規模パターン集合を学習データとする場合，その学習データに付与されたクラスラベルの妥当性の確認が課題となる．この課題に対して上述した k-近傍ノードを利用する場合，k の値を変化させてアウトライアの検出を繰り返す必要がある．例えば，文字認識における異体字のように，クラスラベルが同じであるがその形状が異なるパターンが多数存在する場合には，k の値をいくらにすればよいかが自明ではない．これに対して Clustered-RNG による解析では，大規模な学習データから他のクラスの分布中に存在するパターンを効率的に抽出し，解析すれば学習データに付与されたクラスラベルの妥当性が容易に検証できる．

5.3 実験試料

本章では，分布の異なる複数の数字画像データセットの分布構造を解析し，その結果を比較することで提案手法の妥当性を考察する．具体的には，まず，大規模活字数字画像データセットとしてサンプル数が 259,153 である LS-Print1 と，276,341 である LS-Print2 を用いて各々の分布構造を解析する．次に，大規模手書き数字画像データセットとしては，LS-Hand-All から各クラスのサンプル数が等しくなるようにサンプリングしたサブセットである LS-Hand のデータセットを用いた．LS-Hand のサンプル数は，518,850 である．各データセットの画像例は，すでに図 3.1 に示したとおりである．また，特徴量は 256 次元の輝度特徴と 1024 次元の方向特徴を使用し，パターン間の距離定義はユークリッド距離を用いる．

5.4 活字数字画像データセットの分布構造の解析結果

まず，活字数字画像の分布構造を Clustered-RNG を用いて解析する．直感的に理解しやすく，クラスごとに稠密な分布になっていると予測される活字数字画像の分布構造を可視化して比較することで，前述した提案手法の基本的な性質を検証する．

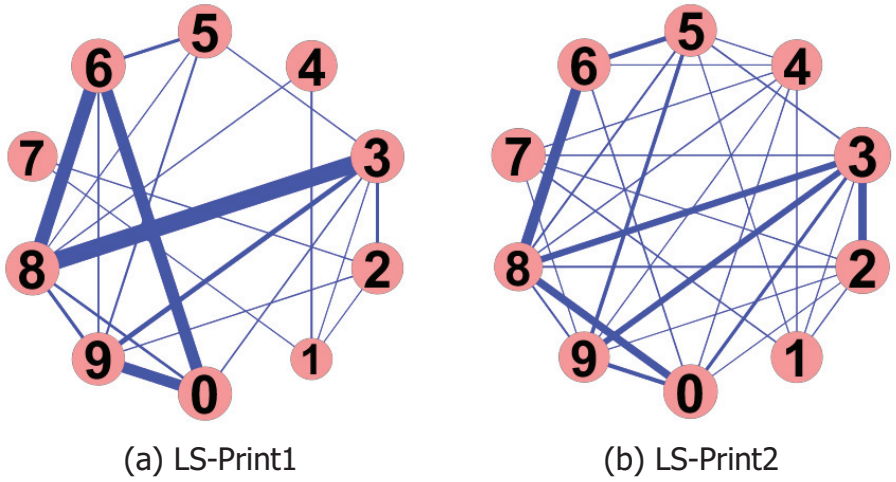


図 5.8: LS-Print の輝度特徴による Clustered-RNG

5.4.1 パターンの差異による分布構造の変化の解析

本項では，輝度特徴による活字数字画像の分布構造を解析し，パターンの差異による分布構造の変化が Clustered-RNG によって解析できることを示す．図 5.8 に，LS-Print1 および 2 の輝度特徴による Clustered-RNG を示す．クラスごとに稠密な活字数字画像の分布構造により，同じクラスのパターンは互いに結合されてクラスタを構成し 1 クラス 1 クラスタとなっている．

一方，異クラス間の関係性に着目すれば，フォントの違いによりクラス対ごとの異クラス間エッジの多寡が異なっている．例えば，LS-Print1 の分布構造である図 5.8(a) では，クラス “3” と “8” が 2,857 本の異クラス間エッジで結ばれているなど特定のクラス対でエッジ数が多く，広範な接面を持つことが推測される．図 3.1 に示した画像例から考えれば，これらのクラス対が広範な接面を持つこと，すなわち類似したパターンが多いことは直感的な理解と相違しない．

クラス対ごとの異クラス間エッジの多寡は，パターンの分布構造の定量的な評価指標となりうる．例えば，クラス “2” と “3” の異クラス間エッジ数は，LS-Print1（図 5.8(a)）で 353 本であったのに対して，LS-Print2（図 5.8(b)）では 1,238 本で

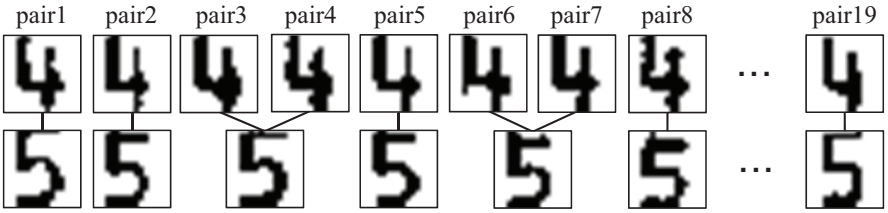


図 5.9: LS-Print2 においてクラス “4” と “5” の異クラス間エッジを持つパターン

あり，多くのエッジを持っている．これは，活字のフォント形状の違いに起因し，LS-Print2 のフォントの形状の方が，前述の 2 値化誤差などの影響を受けて，クラス “2” と “3” に類似したパターンが発生しやすいことを反映している．このような，わずかなフォント形状の差異による分布の変化を的確に解析できる性質は，文字認識用フォントのデザインの定量評価などには有用なものであろう．

図 5.8(b) の LS-Print2 の Clustered-RNG において，クラス “4” と “5” の異クラス間エッジを持つパターンを解析した結果を図 5.9 に示す．同図から，クラス “4” と “5” のクラス境界付近には，本来のフォント形状に対して二値化誤差が加わったパターンが分布していることが確認できる．このように，Clustered-RNG において異クラス間エッジを持つパターンを抽出すれば，各クラス分布の境界付近に存在するパターンの傾向を効率的に解析できる．

5.4.2 特徴量の差異による分布構造の変化の解析

本項では，輝度特徴と方向特徴による活字数字画像の分布構造を解析し，特徴量の差異による分布構造の変化が Clustered-RNG によって解析できることを示す．図 5.10 に，LS-Print1 および 2 の方向特徴による Clustered-RNG を示す．輝度特徴の可視化結果である図 5.8 と同様に，クラスごとに稠密な活字数字画像の分布構造により，同じクラスのパターンは互いに結合されてクラスタを構成し，1 クラス 1 クラスタとなっている．

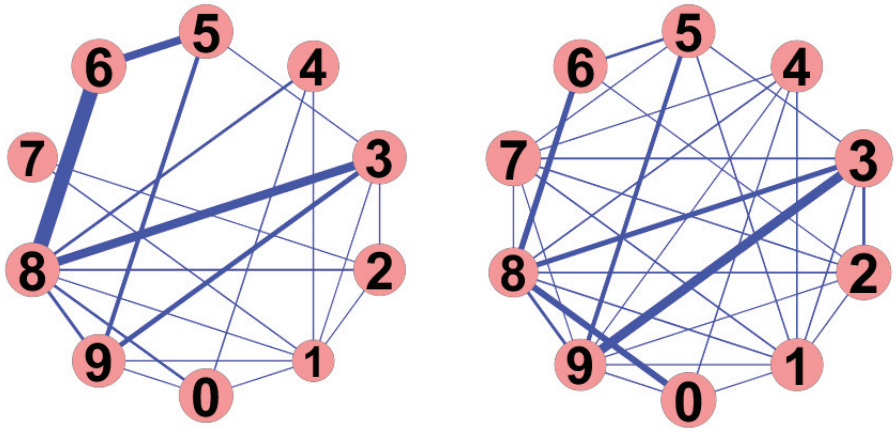


図 5.10: LS-Print の方向特徴による Clustered-RNG

一方，図 5.8 と図 5.10 を比較すると，隣接関係を持つクラス対が異なっている．例えば，図 5.8 で太いエッジを持つクラス “0” と “6” のクラス対は，図 5.10 ではエッジを持たず，方向特徴の特徴空間中では隣接関係がないことが解析できる．実験に用いた各特徴量のパターン間の距離定義は，輝度特徴では各画素値の一致度に相当し，方向特徴では画像中での小領域に含まれる線素の方向の一致度に相当する．このことと図 3.1 に示したフォントの形状から考えれば，クラス “0” と “6” の隣接関係の特徴量による差異は，直観的な理解と相違しない．このように，Clustered-RNG の可視化結果から，特徴量ごとのパターンの分布構造が解析できる．

同様に，図 5.8 と図 5.10 では，クラス間のエッジの多寡も異なっていた．例えば，図 5.10 においてエッジ数は，最大でも LS-Print1 のクラス “6” と “8” の 164 本であった．これは，前項で述べた図 5.8 のエッジ数に比べて大幅に少ない．すなわち，LS-Print1 および 2 の方向特徴による分布構造は，輝度特徴による分布構造に比べてクラス間の分布が粗となっている．このように，Clustered-RNG の可視化結果から特徴量ごとのパターン分布の粗密が解析できる．

Clustered-RNG により解析できるクラス間の隣接性は，混同行列により解析できる隣接性とは異なるものある．5.2.3 でも述べたように，最近傍法による混同行列を

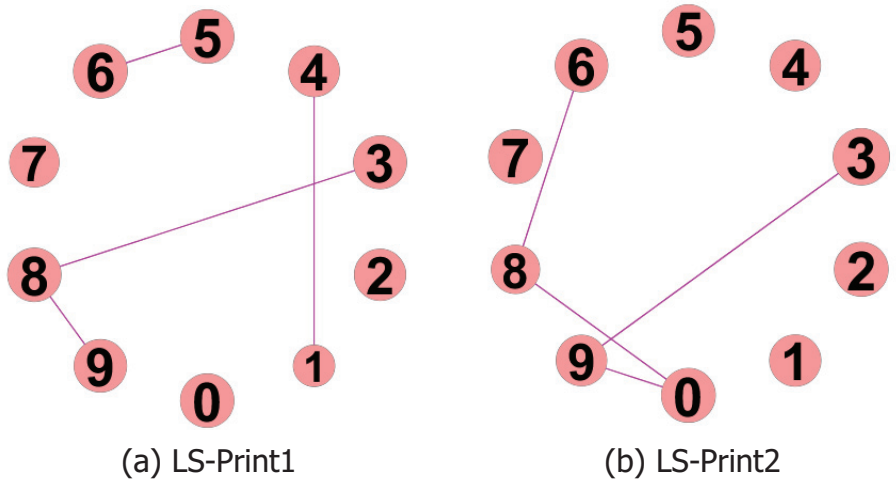


図 5.11: LS-Print の方向特徴による混同行列

用いれば，誤認識が発生するクラス対についてクラス間の隣接性が解析できる．LS-Print1 および 2 の方向特徴による分布構造を混同行列により解析した結果を，図 5.11 に示す．混同行列で分布解析する場合は，同図に示したクラス対間の高々数個の誤認識パターンだけを頼りにする他なく，このような分布構造では，混同行列を用いてクラス間の隣接性を解析できない．一方，Clustered-RNG では異クラス間の隣接性が完全に保存されるため，各クラスの分布構造に基づくクラス間の隣接性を解析できる．

5.5　手書き数字画像データセットの分布構造の解析結果

次に，手書き数字画像のデータセットである LS-Hand の分布構造を Clustered-RNG を用いて可視化する．活字に比べて変形の大きい手書き数字の分布構造を Clustered-RNG を用いて解析し，活字数字画像データセットで行った考察に加えてクラス間のオーバーラップの解析結果の実例を示し，提案手法の妥当性について検証する．

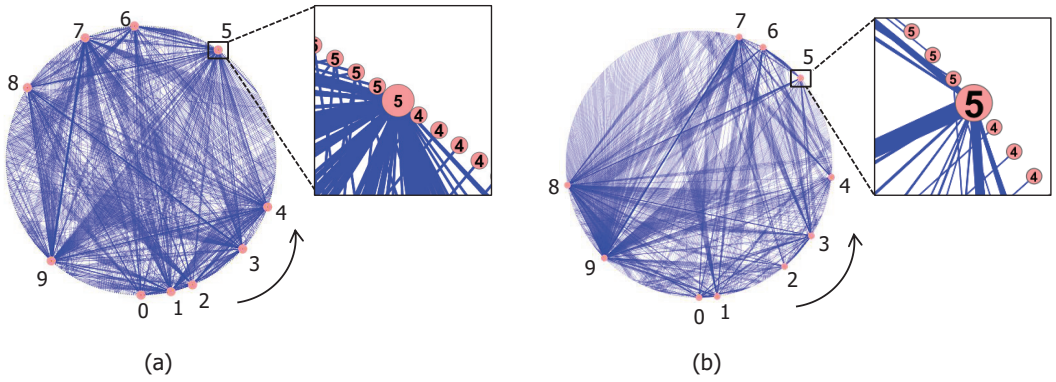


図 5.12: LS-Hand の Clustered-RNG (a) 輝度特徴. (b) 方向特徴

5.5.1 特徴量の差異による分布構造の変化の解析

本項では，輝度特徴と方向特徴による手書き数字画像の分布構造を解析し，特徴量の差異による分布構造の変化が Clustered-RNG によって解析できることを示す．LS-Hand の輝度特徴による Clustered-RNG を図 5.12(a) に，方向特徴による Clustered-RNG を図 5.12(b) に示す．どちらの結果も図 5.10 に示した LS-Print1 および 2 の解析結果と違い，多くのパターンを含む主要クラスタがクラスごとに 1 つと，含まれるパターン数が少ないアウトライアクラスタが多数検出されている．アウトライアクラスタ数は，全体で輝度特徴では 716，方向特徴では 724 であった．また，Clustered-RNG の構築時間はどちらも約 6 時間であり，大規模パターン集合の高次元空間中での分布構造（パターン数は約 50 万，次元数は 1024）を現実的な処理時間で解析できることが確認できた．実験に用いた PC のスペックは Intel Core i7 3.4GHz，メモリ 16GB であり，OS は Windows 7 である．Clustered-RNG の効率的な構築方法については 5.6.1 で詳説する．

まず，図 5.12(a)，(b) を比較すれば，特徴量によって主要クラスタ間のエッジ数が大きく異なっていることが確認できる．この傾向がより明示的になるよう，各 Clustered-RNG から主要クラスタのネットワークを抽出して可視化した結果を図 5.13 に示す．主要クラスタ間の最大エッジ数は，輝度特徴ではクラス “7” と “9” の 30,593 本であっ

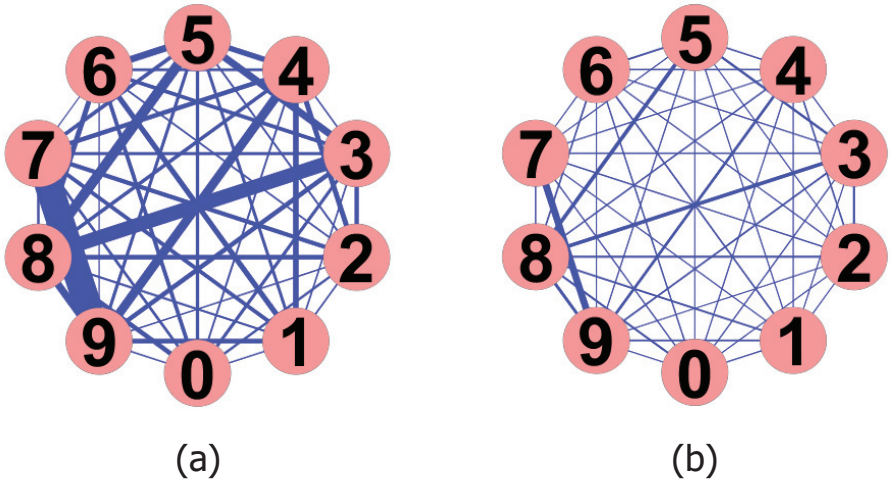


図 5.13: LS-Hand の Clustered-RNG に含まれる主要クラスタのネットワーク．(a) 輝度特徴．(b) 方向特徴

たのに対して，方向特徴ではクラス “7” と “9” の 6,624 本となっている．方向特徴は，画像中の小領域における各方向成分の多寡を特徴ベクトルとしているため，輝度特徴に比べ各画像中での数字の位置ずれを吸収できる．このことから考えれば，図 5.12(a)，(b) でのクラス間の隣接性の変化は直観的な理解と相違しない．この結果から，大局的な分布構造としては，輝度特徴よりも方向特徴の方がクラスごとに稠密な分布となっていることが確認できる．

次に，図 5.12(a)，(b) を比較すれば，クラスごとのアウトライアクラスタ数の多寡が異なっている．各クラスのアウトライアクラスタに含まれるパターン数の合計値を図 5.14 に示す．輝度特徴ではクラス “4”，“8” などのアウトライアが多いのに対して，方向特徴ではクラス “7” のアウトライアが他のクラスに比べて極端に多くなっている．図 5.12(b) の Clustered-RNG から，方向特徴ではクラス “7” のアウトライアクラスタのほとんどが，クラス “9” の主要クラスタとエッジを持つことが解析できる．すなわち，方向特徴では輝度特徴に比べ，クラス “9” の分布にオーバーラップして分布するクラス “7” のパターンが多いことが解析できる．この結果から，局所的な分布構造としては，方向特徴よりも輝度特徴の方が，異クラスの分布にオー

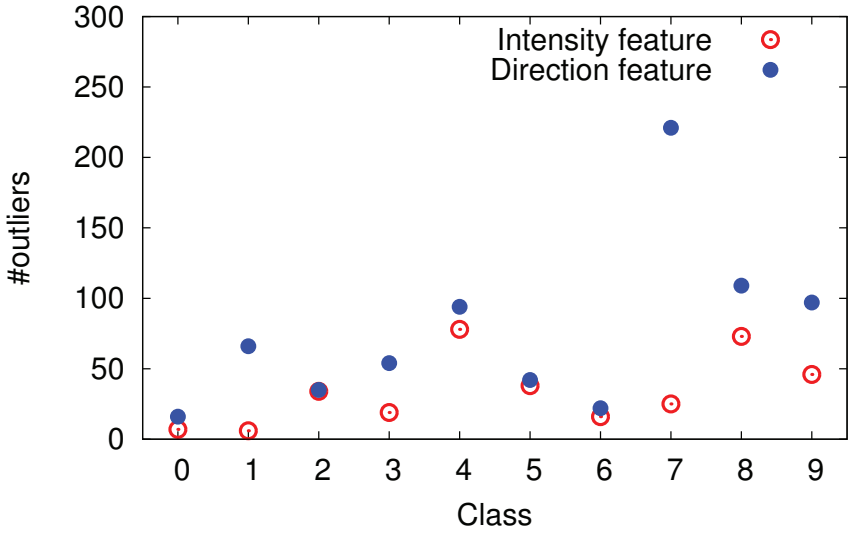


図 5.14: LS-Hand のクラスごとのアウトライア数

バーラップして存在するアウトライア数が少ない分布となっていることが明らかとなった．この Clustered-RNG によって抽出されたアウトライアのパターンについては，次項で詳細に解析する．

Clustered-RNG の異クラス間エッジ数をクラスごとに解析すれば，上述した大局的な分布構造と局所的な分布構造を含んだ大規模パターン集合の分布構造の特性が，明らかとなる．図 5.12(a)，(b) に示した Clustered-RNG について，全てのクラスタが持つ異クラス間エッジ数をクラスごとに解析した結果を図 5.15 に示す．同図から，方向特徴に比べて輝度特徴では異クラス間エッジが多い，すなわちクラス境界付近に多数のパターンが分布していることが解析できる．この異クラス間エッジの解析結果については，5.5.3 で追加の実験を行い妥当性を検証する．

実際にマルチクラスの識別器を構成するうえでは，上述した Clustered-RNG による解析結果を組み合わせて利用することが有効である．特徴空間において自クラスの分布が密で，異クラス間の分布が粗な分布となるように各パターンを射影できる特徴量が，マルチクラスの識別器を構成するための特徴量として適している．しか

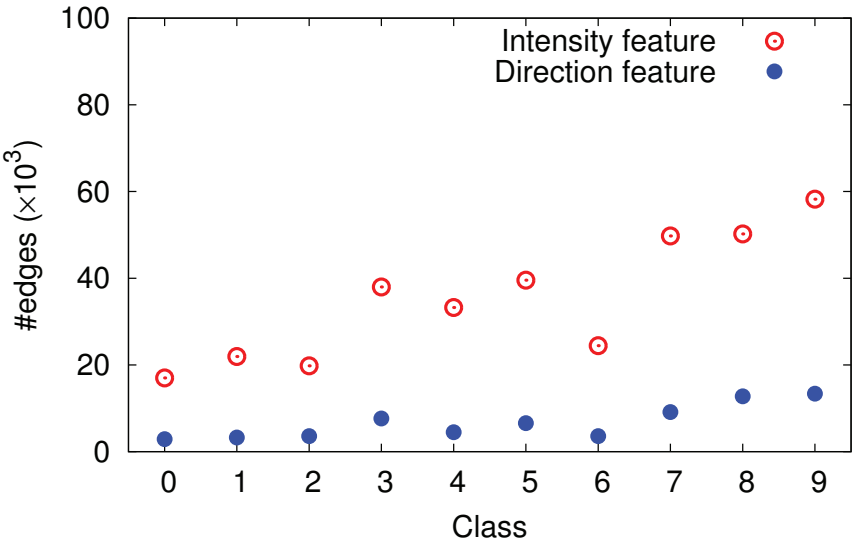


図 5.15: LS-Hand のクラスごとのエッジ数

しながら，全てのクラスに対して，そのような要求を満足する特徴量を見つけることは容易ではない．Clustered-RNG を用いて大局的な分布構造と局所的な分布構造をクラスごとに解析すれば，個々の特徴量の有効性が把握できるだけでなく，複数の特徴量を組み合わせて最終的な認識率を改善することも可能となる．

以上のように，大規模パターン集合の分布構造を Clustered-RNG を用いて可視化すれば，特徴量ごとの分布構造の差異を容易に解析することが可能となる．一般に，マルチクラスの識別器を構成する場合，特徴量によって識別が困難な特定のクラス対が生ずることが多い．この課題に対して，クラス間の隣接性やオーバーラップの差異が容易に解析できれば，識別が困難なクラス対を補完するような複数の特徴量を用いて識別器を構成し，最終的な認識率を改善できると期待できる．

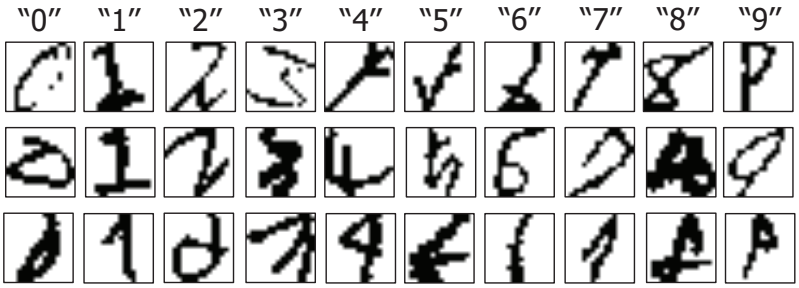


図 5.16: LS-Hand の輝度特徴でのアウトライアクラスタの画像例

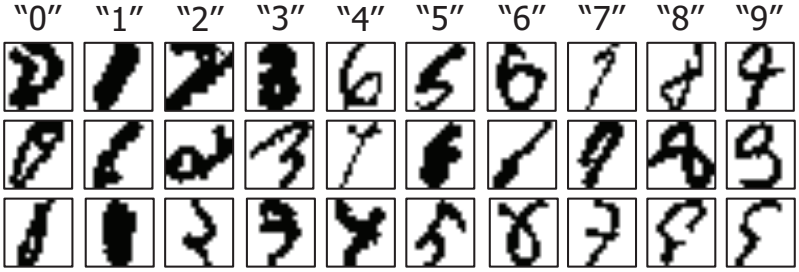


図 5.17: LS-Hand の方向特徴でのアウトライアクラスタの画像例

5.5.2　クラス間のオーバーラップの解析

まず，LS-Hand の輝度特徴による Clustered-RNG（図 5.12(a)）において，アウトライアクラスタとして検出されたパターンを解析した．アウトライアクラスタに含まれていたパターンを図 5.16 に例示する．このように，各クラスの代表的な文字形状のパターンからの変形が大きく，他クラスの代表的な文字形状に近いパターンは，他クラスの分布にオーバーラップして分布するため，アウトライアクラスタを構成する．

次に，LS-Hand の方向特徴による Clustered-RNG（図 5.12(b)）において，アウトライアクラスタとして検出されたパターンの例を図 5.17 に示す．各アウトライアクラスタのエッジから，クラス分布のオーバーラップの状況を解析することも可能である．例えば，図 5.17 で示したクラス “0” のパターンは，最上段がクラス “8”，二段目がクラス “2”，“8”，“9” の主要クラスタとエッジを持つアウトライアクラスタ

のパターンであった．このように，Clustered-RNG の各クラスタに含まれるパターンやエッジを解析することで，どのようなパターンがどのクラスにオーバーラップして分布しているかを解析できる．

図 5.16 と図 5.17 を比較すれば，特徴量によってアウトライアクラスタとして検出されるパターンの傾向が異なっていることが確認できる．具体的には，方向特徴のアウトライアクラスタである図 5.17 の方が，より異クラスの形状に近いパターンを含んでいる．例えば，図 5.16 で例示したクラス “1” のパターンは，輝度特徴による分布構造での結果であるため，画像中での位置ずれによってアウトライアクラスタを構成していると推測される．方向特徴を用いれば，このような位置ずれの影響が吸収できるため，クラス “1” の主要クラスタに含まれることとなる．実際に，図 5.16 で例示したクラス “1” のパターンに類似したパターンは，方向特徴のアウトライアクラスタには含まれていなかった．

このように，異クラスの分布中にオーバーラップして存在するパターンを効率的に解析できるという特長は，大規模パターン認識にとって有用性が高い．5.2.3 でも述べたように，大規模パターン集合を識別器の学習データとする場合，その学習データに付与されたクラスラベルの妥当性の確認が課題となる．これに対して，k-近傍ノードのクラスラベルの最頻値と比較しアウトライアを検出する方法では，k の値をいくらにすればよいかが自明ではないため k の値を変えながら解析を繰り返す必要がある．一方，Clustered-RNG による解析では，アウトライアクラスタとして検出されたパターンのみを確認すれば，クラスラベルの妥当性を効率的に検証できる．実際に我々の実験の初期段階において，最近傍法では検出できていなかった漢数字のパターンを，アウトライアクラスタのパターンを解析することで容易に検出し，スクリーニングできた．

5.5.3 識別境界付近でのパターンの分布構造解析

Clustered-RNG のクラスタ間のエッジからは，識別境界付近でのパターン分布構造が解析できる．例えば，図 5.12(b) のクラス “7” のアウトライアクラスタと，ク

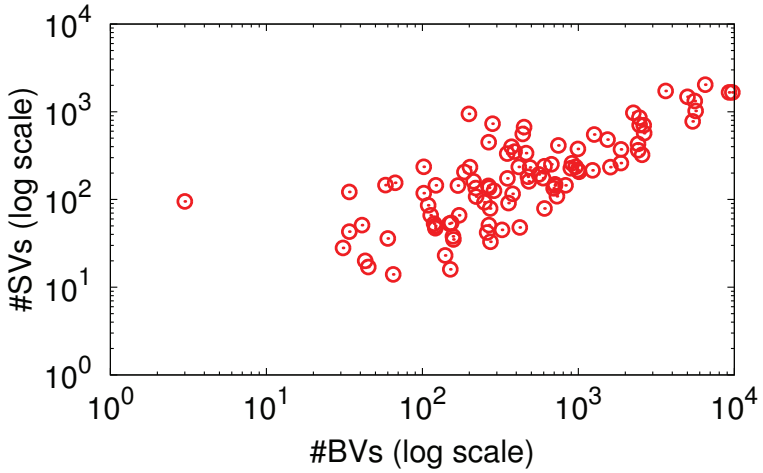


図 5.18: ブリッジベクトル数とサポートベクトル数の関係

ラス “1” や “9” などの主要クラスタ間のエッジから，クラス “7” の一部のパターンがこれらのクラスにオーバーラップして分布していることが解析できる．また，主要クラスタ間が太いエッジで結ばれているクラス “7” と “9” などは，識別境界付近に多くのパターンが分布していることが解析できる．

この解析結果の妥当性を検証する実験を行った．具体的には，まず LS-Hand の方向特徴を用いて，SVM による多クラス識別器のサポートベクトル (Support Vector, SV) の数をクラスごとに解析した．そして，その SV 数をブリッジベクトル (Bridge Vector，BV) 数，すなわち Clustered-RNG において異クラスタ間のエッジを持つパターンの数と，クラスごとに比較した．SVM は，LIBSVM [99] の RBF カーネルを使用した．

図 5.18 に示した実験結果から，各クラスのブリッジベクトル数とブリッジベクトル数には相関性があることが，確認できる．また，SVM がサポートベクトルとして選択したパターンのうち 6 割以上は，ブリッジベクトルでもあった．このことから，Clustered-RNG によって解析できるクラス間の隣接関係の傾向は，妥当性があると判断できる．

Clustered-RNG による解析結果は，パターンの分布構造のみによって一意に定まるため，クラス間分布の粗密を効率的に解析できる．SVM のサポートベクトル数でパターンの分布構造を解析する場合，SVM のパラメータ調整のために学習を繰り返す必要があり，学習パターンが大規模化すれば解析に必要な処理時間が非常に長くなる．一方，提案手法には解析に必要なパラメータが存在しないため，あるデータセットの分布構造の解析に必要な Clustered-RNG の構築処理は 1 回のみであり，効率的な解析が可能となる．

5.6 Clustered-RNG 構築の計算量

提案手法により大規模パターン集合の分布構造を解析する場合，Clustered-RNG 構築の計算時間と計算に必要なメモリ空間量が実用上の課題となる．ここで，Clustered-RNG 構築に必要な計算時間は，ほぼ RNG の構築に必要な計算時間となる．よって，RNG の構築処理を高速化できれば効率的に Clustered-RNG を構築できる．

5.6.1 高速かつ省メモリな RNG 構築法

2.2.5 でも述べたように，従来提案されている RNG 構築のアルゴリズム [59] の多くは，2 次元あるいは 3 次元程度の低次元空間での RNG 構築を前提としているため，高次元空間では効率的に RNG を構築できない．最も単純な総当たりによる RNG 構築法である RNG-1 アルゴリズム [13] により，RNG を構築する場合の計算量は，ノード数 N に対して $O(N^3)$ である．また，高速化のために，全ての距離をメモリ上に蓄積しようとすれば，$O(N^2)$ のメモリ空間容量が必要となる．よって，RNG-1 アルゴリズムを用いて RNG を構築した場合，本研究が対象とする大規模パターン集合の分布解析が，実現困難となる．

上述した RNG-1 アルゴリズムは，次の手続きで RNG の構築を行う．

Step 1： 全てのノード対に対して距離 $d(v_h,\ v_k),\ h,k = 1,2,\ldots,N,\ h \neq k$ を計算する．

Step 2 : 各ノード対 $(v_h,\ v_k)$ に対して，その他のノードとの相対距離 $d^x_{\max} = \max\{d(v_h,\ v_x),\ d(v_k,\ v_x)\},\ x = 1, 2, \ldots, N,\ x \neq k,\ x \neq h$ を計算する．

Step 3 : 各ノード対 $(v_h,\ v_k)$ に対して，$d^x_{\max} < d(v_h,\ v_k)$ を満たすノード v_x が存在しなければ，ノード対 $(v_h,\ v_k)$ 間にエッジを付与する．

提案手法は，RNG-1 アルゴリズムを改良して，RNG 構築を高速かつ省メモリに実現する．まず，ノードを複数のグループに分割し，各ノードからの各グループに属するノードまでの最小距離とそのノードの番号のみをメモ化 [100]，すなわちルックアップテーブル (Look Up Table，以下，LUT) としてメモリ上に蓄積することで，必要なメモリ空間量を削減する．そして，この LUT を用いて，あるノード対に対してエッジが付与されないことを検出するために必要な計算量を削減する．具体的には，後述する LUT を利用した二つの判定により，エッジの付与を阻害するノード，すなわち $d^x_{\max} < d(v_h,\ v_k)$ を満たすノード v_x を検出するために必要なノード間距離の計算と比較演算の回数を削減する．このように，ノード対間の角度や領域分割を利用せず，ノード間の距離のみを用いて RNG を構築すれば，ノード間の距離定義を L_2 距離以外への拡張も容易となる．

RNG-1 アルゴリズムに対する提案手法の改良点は，以下の 2 点である．まず，図 5.19 に示すように，RNG-1 アルゴリズムの Step 1 でのノード間の距離計算において各ノードをランダムに P 個のグループに分割し[1]，ノード集合 $\boldsymbol{V}_p,\ p = 1, 2, \ldots, P$ を構成する．そして，ノード v_h から他のノード $v_k,\ k = 1, 2, \ldots, N,\ k \neq h$ に対する距離 $d(v_h,\ v_k)$ を計算し，v_h から各グループのノードに対する最小距離 $d^p_{\min} = \min\{d(v_h,\ v_x)\},\ v_x \in \boldsymbol{V}_p,\ x \neq h$ とそのノードの番号 x を LUT に追加する．図 5.19(a) の例では，v_n に対して，図中に太線で示した 4 つのノードとの距離および各ノードの番号が LUT に追加される．

次に，RNG-1 アルゴリズムのエッジ付与判定，すなわち RNG-1 アルゴリズムの

[1]図 5.19 では，図の単純化のためにクラスごとにグループを分割する例を示しているが，別のグループに比較的距離が近いノードが含まれる方が計算量の削減効果が高くなる．実験では，各クラスのパターンを各グループに均等に分割した．

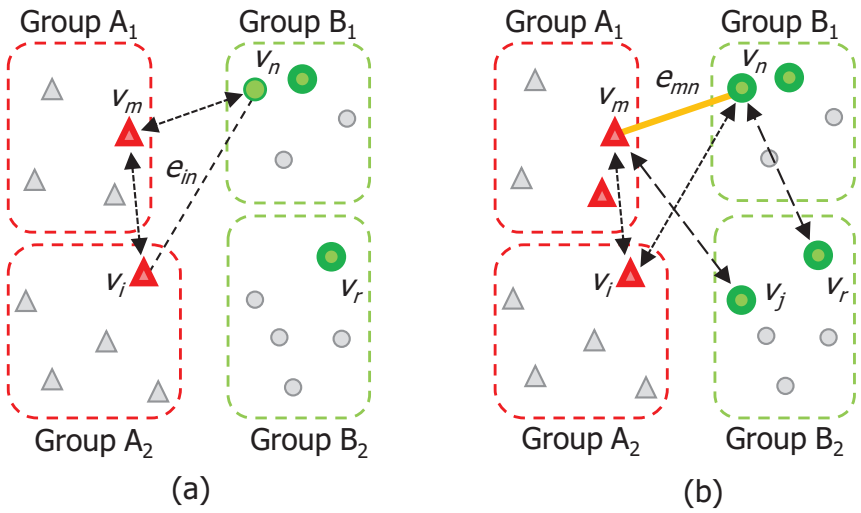


図 5.19: ノードのグルーピングによる RNG の構築．(a) エッジが付与されないノード対の検出．(b) エッジの付与判定の対象となるノードの削減

Step 2 と 3 を行う前に，二つの判定を行う．これらの判定により，エッジ付与判定の対象となるノードが大きく削減でき，提案手法は，全体として $O(N^2)$ の処理時間での RNG 構築を実現する．

第一の判定は，LUT を参照することで，上述したエッジの付与を阻害するノードを早期に，すなわち少ない判定回数で検出する処理である．図 5.19(a) の例では，ノード v_n の LUT に登録されているノード，すなわち図中に太線で示した 4 つのノードのみを対象にノード間距離の計算と比較演算を行えば，ノード対 $(v_i,\ v_n)$ へのエッジの付与を阻害するノードを検出できる．具体的には，例えばノード v_m が $\max\{d(v_n,\ v_m), d(v_i,\ v_m)\} < d(v_i,\ v_n)$ であるため，エッジの付与を阻害するノードとして検出できる．

第二の判定は，LUT を参照することで，エッジ付与判定の対象とする必要のないノード集合を検出する判定である．具体的には，ノード対 $(v_h\ v_k)$ について，ノード v_h，v_k それぞれに対する距離が最小のノード $v_{x'},\ v_{x''} \in \boldsymbol{V}_p$ に対して，$d(v_h,\ v_k) \leq \max\{d(v_h,\ v_{x'}), d(v_h,\ v_{x''})\}$ であれば，ノード集合 $\boldsymbol{V}_p$ に属するノードは，エッジ付与判定の対象から削除することができる．

図 5.19(b) の例では，ノード v_m, v_n の LUT に登録されているノード，すなわち図中に太線で示した 7 つのノードのみを対象にノード間距離の比較演算を行うことで，全てのノード集合に属するノードをエッジ付与判定の対象から削除することができる．例えば，まず，$d(v_m, v_n) \leq \max\{d(v_m, v_i), d(v_n, v_i)\}$ であるため，ノード対 (v_m, v_n) のエッジ付与判定の対象からグループ A_2 に属するノードを削除できる．また，$d(v_m, v_n) \leq \max\{d(v_m, v_j), d(v_n, v_r)\}$ であるため，グループ B_2 に属するノードもエッジ付与判定の対象から削除できる．以下同様に，グループ A_1 およびグループ B_1 に属するノードもエッジ付与判定の対象から削除できる．

また提案手法は，RNG 構築に必要なメモリ空間量の削減に対しても効果がある．これは，エッジ付与判定の回数が少なくなるため，ノード間距離の計算結果をメモリ上に蓄積せず，必要に応じて距離計算を行っても十分高速に処理できるためである．LUT をメモリ上に蓄積するために必要なメモリ空間量は，$O(N)$ でスケールするため，提案手法はメモリ効率の点でも有効な手法である．

5.6.2 Clustered-RNG 構築の計算量の評価実験

本項では，前項で述べた RNG 構築法を適用して Clustered-RNG 構築を行い，その処理時間を評価する．RNG 構築には全てのノード対間での距離計算が必要となるため，Clustered-RNG 構築の処理時間が $O(N^2)$ 以下でスケールしていれば提案手法が効率的なアルゴリズムであると判断できる．

実験諸元

実験では，MNIST と LS-Hand のデータセットを使用し，パターン数の多寡だけでなく特徴量の次元数による処理時間の変化も評価した．特徴量の次元数は，28×28 画素の画像である MNIST では，輝度特徴で 754 次元，方向特徴で 1024 次元であり，16×16 画素の画像である LS-Hand では，輝度特徴で 256 次元，方向特徴で 1024 次元である．

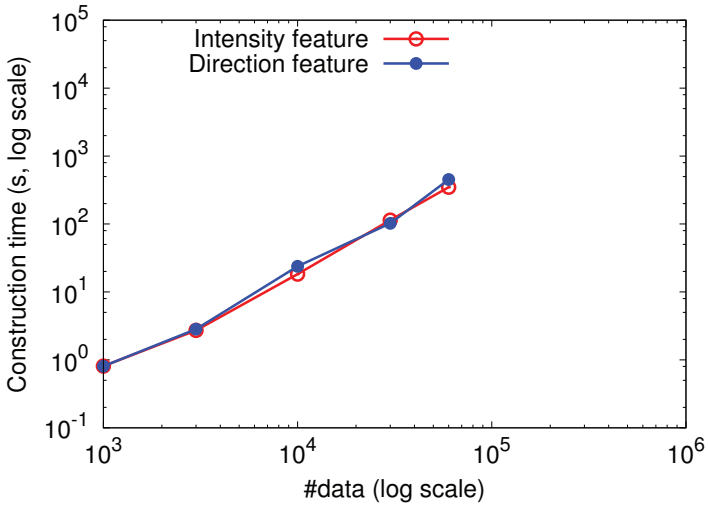


図 5.20: MNIST での Clustered-RNG 構築処理時間

実験では，Clustered-RNG の構築処理を各データセットを用いて 3 回行い，処理時間の平均値を Clustered-RNG 構築の処理時間とした．また，実験に用いた PC スペックは Intel Core i7 3.4GHz，メモリ 16GB であり，OS は Windows 7 である．

実験結果

まず，MNIST のデータセットに対する Clustered-RNG 構築の処理時間を図 5.20 に示す．同図から，Clustered-RNG 構築の処理時間が，特徴量によらず $O(N^2)$ 以下でスケールしていることが確認できる．具体的には，Clustered-RNG 構築の処理時間が約 8 分であった．これに対して，前述した RNG-1 アルゴリズムを用いた場合の処理時間は約 20 時間であり，提案手法により効率的に Clustered-RNG が構築できることが確認できる．

次に，LS-Hand のデータセットに対する Clustered-RNG 構築の処理時間を図 5.21 に示す．同図から図 5.20 の MNIST での実験結果と同様に，約 50 万のデータセットである LS-Hand でも，Clustered-RNG 構築処理時間が $O(N^2)$ 以下でスケールしていることが確認できる．具体的には，方向特徴を用いた場合，Clustered-RNG 構築の処理時間が約 6 時間であった．MNIST の実験結果と違い，特徴量によってわずか

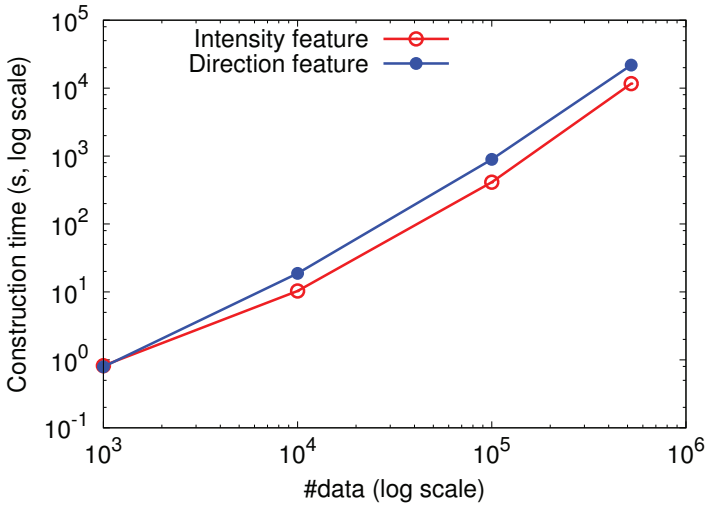


図 5.21: LS-Hand での Clustered-RNG 構築処理時間

ながら処理時間に差異が生じている点は，上述した特徴量の次元数の違いに起因している．

提案手法と RNG-1 アルゴリズムによる Clustered-RNG 構築の処理時間の差は，データセットのサイズに依存するため，MNIST に比べ LS-Hand では処理時間の差がより顕著となる．例えば，サンプル数の少ないサブセットでの処理時間から，LS-Hand の全パターンの方向特徴を用いた場合の RNG-1 アルゴリズムによる処理時間を予測すれば約 9,000 時間となる．実験に用いた RNG-1 アルゴリズムのプログラムでは，高速化のためにノード間距離を全てメモリ上に蓄積している．しかしながら，LS-Hand の約 50 万のノード間距離を全てメモリ上に蓄積するためには 250GB 以上のメモリ空間量が必要となる．そのため，実験に用いた環境ではノード間距離をメモリ上に蓄積する高速化の手法が利用できず，ノード間距離を適宜計算する必要があるため処理時間は 9,000 時間から更に増大する．よって，大規模パターンに対して RNG-1 アルゴリズムを適用した場合，現実的な時間での Clustered-RNG 構築は実現できない．

提案手法では，高速化のためにノードをグルーピングするため，グループ数 p が設定可能なパラメータとなる．本節の実験では，提案手法のノードのグルーピング

は試行ごとにランダムに行いグル―プ数は $p = 64$ とした．

p の値の大きさと計算量の関係の見積りは次のようになる．まず，上述した二つの判定では，小さすぎる p の値を適用すれば，メモ化された最小距離によってエッジ付与を阻害するノードが検出できない可能性が高くなるため，計算量が増加する可能性がある．実際に $p = 8$ として実験した場合，試行ごとの処理時間のばらつきが大きくなる傾向があった．逆に，p の値を大きくした場合，上述の二つの判定で棄却できるグループに含まれるノード数が増加すると期待できる．一方で，探索するメモ化された最小距離の数が増加しかつその距離を保持するメモリ空間量が増加するため，CPU のキャッシュメモリでミスヒットする確率が高くなる．よって全体として処理時間が増加する可能性がある．実際に $p = 200$ とした場合，試行ごとの処理時間のばらつきが大きくなる傾向があった．また，評価に用いたプログラムはグループ単位で処理を並列化するため本節で用いた実験環境の CPU の仕様により p の値が 8 の倍数の場合に最も効率が良い．実際に，p の値を 10 から 100 の間で 8 の倍数から複数設定し，上述した Clustered-RNG 構築の処理時間を評価した結果では，平均処理時間はほぼ一定であった．よって，Clustered-RNG の構築処理の高速化のために p の調整を行う必要はない．

5.7 むすび

本章では第一に，クラスタ化された RNG を用いてパターンの分布構造を解析する手法を提案した．提案手法では，まず各パターンを 1 ノードとし，その近傍関係によりエッジを付与して RNG を構築する．そして，その RNG をクラスタ化，すなわちエッジで結合された同じクラスラベルを持つパターンを統合して 1 つのノードとして表現した Clustered-RNG を用いてパターンの布構造を解析する．提案手法により，多クラス分布間の隣接性やアウトライアの分布状況の解析が可能となる．具体的な解析事例として，約 50 万の活字数字画像と約 50 万の手書き数字画像のデータセットと複数の特徴量を用いた実験を行い，提案手法によって明らかとなる分布

構造の性質と有効性について示した．

提案手法によれば，4.6 で述べた MST によるクラスタ木では解析できなかった多クラス間の隣接関係を解析できる．RNG は MST を内包しており，MST では表現できない多クラス間の隣接関係を表現でき，その隣接関係は RNG をクラスタ化した Clustered-RNG でも保持される．Clustered-RNG の構造とエッジの太さから，特徴空間中でのクラス分布の隣接性とクラス間分布の粗密が解析できる．

提案手法によって解析できる特徴量ごとあるいはクラスごとの分布構造は，マルチクラスの識別器を構成する際に有効である．一般に，マルチクラスの識別器では，特定のクラス対の識別が困難になることが多い．提案手法によって解析できる異クラス間の隣接性やクラス間のオーバーラップの差異を利用すれば，この課題への対策ができる．例えば，異クラス間分布が粗となるように複数の特徴量を組み合わせて識別器を構成すれば，最終的な認識率を改善できると期待できる．

提案手法は，近年研究が進んでいる大規模パターン認識にとっても高い有用性を持つ．まず，学習用の大規模パターン集合に付与されたクラスラベルの妥当性の検証という課題に対して，提案手法により抽出されるアウトライアクラスタのパターンを解析すれば，効率的にクラスラベルの妥当性を検証できる．また，提案手法により抽出できる識別境界付近のパターンとその隣接関係は，大規模パターン集合を学習データとする識別器の学習の高速化に応用が可能である．この点については，提案手法のパターン認識への応用事例として，次章で詳説する．

第二に，高次元空間中の大規模パターン集合に対して，高速かつ省メモリに RNG を構築する手法を提案した．提案手法は，エッジ付与の判定回数を削減することで効率的に RNG を構築する．具体的には，学習パターンをランダムにグルーピングし，各グループに対する各ノードからの最小距離を基準にエッジの付与判定が必要なノードを絞り込むことで，エッジの付与判定の回数を削減する．また，RNG の構築に必要なメモリ空間量が $O(N)$ でスケールするため，提案手法はメモリ効率の点でも有効な手法である．複数のデータセットと特徴量を用いた実験を行い提案手法によって Clustered-RNG が効率的に構築できることを確認した．

第6章
SVMの学習高速化への応用

6.1 まえがき

本章では，Clustered-RNG による分布構造解析法をパターン認識へ応用した例として，学習パターンの事前選択による SVM の学習高速化の実例を示す．提案手法は，Clustered-RNG による分布構造の解析結果を基に，学習パターンから異クラス間エッジ（クラスラベルが異なるパターンを結合するエッジ）を持つパターンを事前選択することで，SVM の学習を高速化する．

すでに 2.4 で述べたように，大規模な学習パターンを用いて SVM 識別器を構成する場合，SVM の学習時間が実用上の課題となる．具体的には，一般に SVM の学習は $O(N^2)$ の計算量が必要であり，SVM はパラメータ調整のために SVM の学習を多数回繰り返す必要がある．よって，大規模パターン集合を学習に用いる場合，SVM の学習高速化は重要な課題となる．SVM はクラス境界付近のパターンを利用してクラスを識別するため，SVM の学習において，各クラスのクラス境界に寄与するパターンのみを事前選択することができれば，SVM の学習時間を短縮できる．

本章では，複数のデータセットと特徴量を用いた実験を行い，ブリッジベクトル，すなわち Clustered-RNG によって抽出される異クラス間エッジを持つ特徴ベクトルを学習パターンとして事前選択した場合の SVM 識別器の認識率と SVM の学習時間を検証する．すでに 5.5.3 で示した実験結果（図 5.18）から，ブリッジベクトル数とサポートベクトル数には相関性があることが確認されている．すなわち，サポートベクトルもブリッジベクトルもクラス境界付近のパターンであるため，ある程度のサポートベクトルはブリッジベクトルである可能性がある．そこで，事前選択の有無による SVM 識別器のサポートベクトルの変化も解析した．

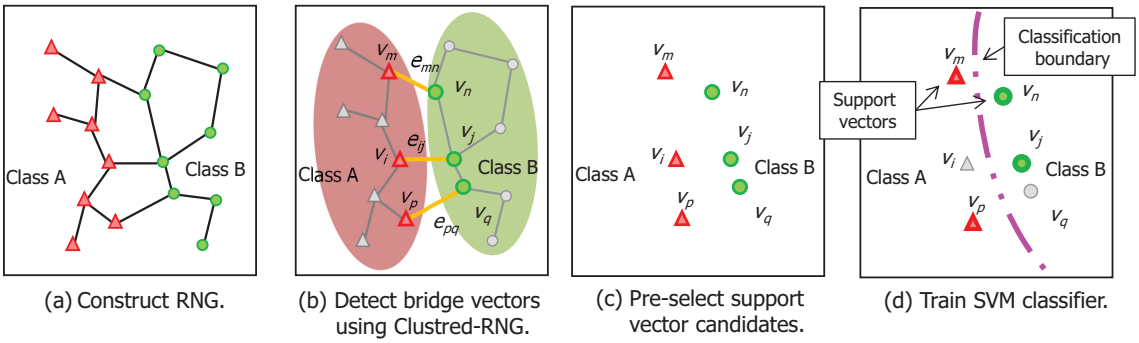


図 6.1: 学習パターンの事前選択による SVM の学習

6.2 事前選択による SVM の学習高速化

6.2.1 SVM の学習パターンの事前選択

Clustered-RNG による分布構造の解析結果から学習パターンを事前選択すれば，SVM の学習を高速化できる．具体的には，ブリッジベクトルを SVM の学習パターンとして事前選択すれば，識別境界付近のパターンのみを用いて効率的に SVM 識別器を構成できる．提案手法による SVM の学習は，次の手順で行う．

1. 学習パターンに対する RNG を構築する（図 6.1(a)）．
2. RNG の粗視化により Clustered-RNG を構築し，ブリッジベクトルを検出する．図 6.1(b) の例では，黄色で示した異クラス間エッジ e_{mn}, e_{ij}, e_{pq} を持つ 6 つのノードがブリッジベクトルである．
3. 検出されたブリッジベクトルを学習パターンとして事前選択する（図 6.1(c)）．
4. 事前選択された学習パターンを用いて SVM の学習を行い，SVM 識別器を構成する．図 6.1(d) の例では 6 つのブリッジベクトルを学習した結果，4 つのサポートベクトルが選択されている．

提案手法は，特に one-vs.-all の 2 クラス SVM を構成する場合に，他クラスの学習パターンを大幅に削減できる．これは，Clustered-RNG の多クラス間の隣接性を同

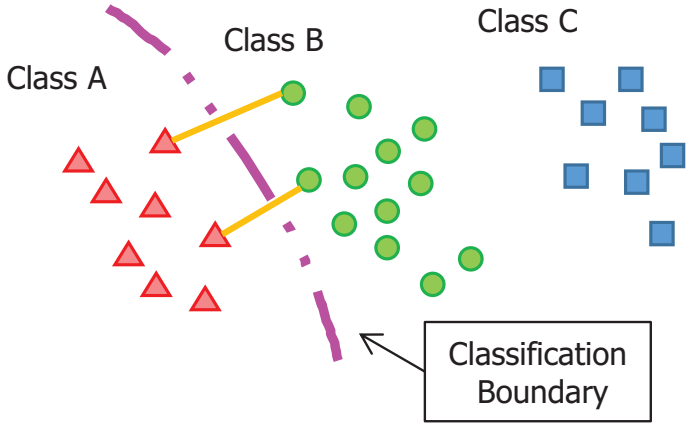


図 6.2: パターン分布と識別平面の例

時に解析できるという特長が利用できるためである．例えば，図 6.2 の分布構造を持つパターンの識別を行う場合，クラス A とそれ以外のクラスを識別するための識別平面を決定するためには，クラス A のパターンおよびクラス A と隣接性を持つクラス B のパターンのみを考慮すればよい．すなわち，クラス C のパターンは学習パターンから削除できる．さらに，同図に黄色で示したクラス A，B 間の異クラス間エッジを持つパターンのみを用いて，全パターンを用いた場合と同等の識別平面を決定できれば，SVM の認識率を低下させずに学習パターンを大幅に削減できる．

実例として，図 5.8(a) に示した LS-Print1 の輝度特徴による Clustered-RNG を考える．例えば，クラス “0”-vs.-all の 2 クラス SVM を構成する場合，クラス “0” は 4 つのクラス（クラス “3”，“6”，“8”，“9”）に対してのみ異クラス間エッジを持つ．よって，クラス “0” と異クラス間エッジを持たないクラス，(クラス “1”，“2” など) の全てのパターンは，事前選択によって学習パターンから削除される．そして，学習パターンとして事前選択されるパターンは，クラス “0” と異クラス間エッジを持つ 4 つのクラスに存在するブリッジベクトルのみであるので，学習パターン数を大幅に削減できる．

ここで強調すべきは，以上の Clustered-RNG による事前選択には，調整を要するパラメータが存在しない点である．2.4 で述べたように，SVM はいくつかのパラメー

タを調整する必要があり，パラメータを変えながらSVMの学習を多数回繰り返す必要がある．ここで，事前選択のために新たなパラメータが導入されれば，SVMの学習の回数がさらに増加することとなり，事前選択により個々のSVMの学習時間が短縮したとしても，SVMの構成に必要な処理全体に対する処理時間の短縮効果は低くなる．一方，提案手法では，事前選択される学習パターンがその分布構造によって一意に定まるため，パラメータ調整に必要なSVMの学習の回数は増加しない．よって，提案手法ではSVM識別器の構成に必要な処理全体に対する処理時間を大幅に短縮できる．

6.2.2 事前選択のSVM識別器への影響

Clustered-RNGによって抽出されるブリッジベクトルのみを学習パターンとしてSVMの学習を行えば，識別境界付近に分布するパターンのみを用いて効率的に学習が行える．ここで，全学習パターンを用いて学習したSVM識別器のサポートベクトルは必ずしもブリッジベクトルではないため，事前選択の有無によってサポートベクトル数や含まれるパターンは変化する．しかしながら，SVM識別器を構成する目的はクラス間の識別平面の表現であり，サポートベクトル集合が異なっていても同等の識別平面が表現できるならば認識率も同等となる．

SVMは，サポートベクトルからの距離の線形和によって識別平面を表現するため，サポートベクトルごとに識別平面への寄与度が異なる．RBF(Radial Basis Function)カーネルKを用いた場合，SVMの識別関数は，$f(v) = \mathrm{sign}(\sum_{k \in \boldsymbol{S}} \alpha_k y_k K(v, v_k) + b)$, $y_k \in \{-1, 1\}$となる．ここで，$\boldsymbol{S}$はサポートベクトル集合，α_kは各サポートベクトルの重み，v_kはサポートベクトルである．よって，重みαが大きいサポートベクトルが，識別平面への寄与度が大きく重要であるといえる．

サポートベクトルに含まれるパターンが完全に一致していなくとも，同等の識別平面を表現するSVMは構成可能である．具体的には，全ての学習パターンを用いて学習したSVM識別器のサポートベクトルのうち，重みαが大きなサポートベクトルが，ブリッジベクトルまたはブリッジベクトルの近傍に存在するパターンであ

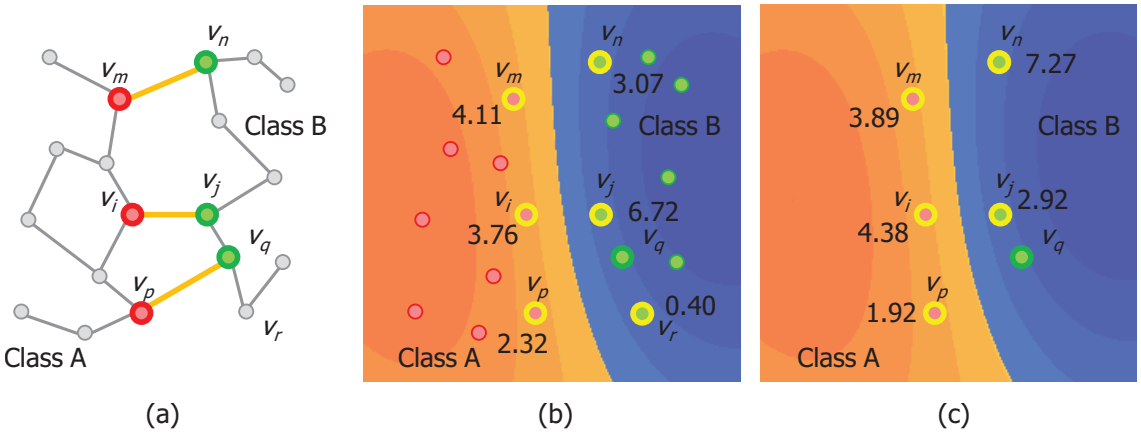


図 6.3: 2 クラス SVM の識別関数の可視化．(a) 学習パターンの RNG. (b) 全てのパターンで学習した場合の識別関数．(c) ブリッジベクトルのみで学習した場合の識別関数

ればよい．例として，図 6.3(a) の分布を持つ 19 個のパターンを学習パターンとする場合を考える．同図に黄色で示したエッジが異クラス間エッジであり，それらのエッジを持つ 6 個のパターンがブリッジベクトルである．次に，図 6.3(b)，(c) は，RBF カーネルによる SVM の識別関数の可視化結果であり，各サポートベクトルを各々の重み α とともに黄色で示している．同図 (b) は全パターンで学習した結果であり，同図 (c) はブリッジベクトルのみで学習した結果である．同図 (b)，(c) を比較すれば，異なるサポートベクトルによって同等の識別平面が表現されていることが確認できる．この事前選択によるサポートベクトルの変化については，6.5 で検証する．

6.3 実験条件

6.3.1 実験試料と実験諸元

本章では，大規模手書き数字画像データセットである LS-Hand と公開手書き数字画像データセットである MNIST を用いて実験を行い，提案手法の妥当性を考察す

る．SVM 識別器の評価のためには学習用と評価用のパターンが必要であるため，学習用として 518,850 パターンの LS-Hand を用い，評価用として LS-Hand とは別の 100,000 パターンを LS-Hand-All から各クラスのパターン数が等しくなるようにサンプリングしたデータセットを用いた．MNIST は，公開されている学習用の 60,000 パターン，評価用の 10,000 パターンを用いた．各データセットの画像例は，すでに図 3.1 に示したとおりである．また，特徴量は輝度特徴と方向特徴を使用し，パターン間の距離定義はユークリッド距離を用いる．輝度特徴は，16×16 画素の画像である LS-Hand では 256 次元，28×28 画素の画像である MNIST では 784 次元となる．方向特徴は，データセットによらず 1024 次元である．

SVM は，LIBSVM [99] の RBF カーネルを使用した．また，コストパラメータ C とカーネルパラメータ γ は，5-分割交差検証を用いたグリッドサーチにより調整した．実験に用いた PC スペックは Intel Core i7 3.4GHz，メモリ 16GB であり，OS は Windows 7 である．

6.3.2 評価項目

事前選択による SVM の学習高速化に関する実験では，事前選択前後のデータセットに加えて，事前選択後のクラスごとのパターン数と同数をランダムに抽出したデータセットを用いる．具体的には，各データセットから (i) 全パターン，(ii) ブリッジベクトル（提案手法），(iii) ブリッジベクトルと同数をランダムサンプリングの 3 通りのデータセットを作成し，クラスごとに one-vs.-all の SVM を構成して学習時間と認識率を比較する．認識率はクラスごとに異なるため，各クラスの SVM 識別器の誤認識率を平均して評価した．

SVM の学習高速化を目的とした手法の有効性を評価するうえでは，処理時間だけではなく，SVM 識別器の認識率も考慮する必要がある．例えば，(iii) では，LIBSVM の実装に用いられている SMO [84] の計算量 $O(N^2)$ に比例して SVM の学習時間が短縮される．しかしながら，クラス境界付近のパターンが学習パターンに含まれないため，認識率が低下すると予想される．一方，(ii) では，学習パターンがクラス境

界付近のパターンで構成されるため，認識率は (i) と同等程度になると期待される．しかしながら，学習パターンがクラス境界付近に偏ることで，(iii) と比べて SVM の学習計算が収束しにくくなり，学習パターン数が同数でも SVM の学習時間が長くなると予想される．そこで，以下の実験では，認識率と SVM の学習時間のトレードオフにより，提案手法の有効性を評価する．

6.4 大規模データセットでの SVM の学習高速化の評価実験

本節では，各学習パターンセットを用いて SVM の学習を行い処理時間と認識率を評価して提案手法の有効性を示す．また，学習パターン数を変化させた場合の実験も行い，パターン数が増加すれば事前選択の有無による認識率の差が小さくなるため，提案手法が大規模パターン集合に対して有効であることを示す．

6.4.1 事前選択の効果の評価

LS-Hand の方向特徴を用いて SVM の学習を行った実験結果を表 6.1 に示す．ブリッジベクトル数は学習するクラスごとに異なるため，(ii)，(iii) については学習パターン数の平均値を記載した．パラメータ調整時間は，グリッドサーチによってパラメータ C，σ を 3×6 通りから選択するための処理時間であり，Clustered-RNG の構築時間は，交差検証用と SVM の学習用の合計 6 種類の Clustered-RNG を構築する処理時間である．

表 6.1 から，事前選択による認識率の大きな低下は生じていないことが，確認できる．一方，事前選択により学習パターン数を 1/50 に削減できている．事前選択されたクラスごとの学習パターン数は，クラス “9”-vs.-all の学習パターンが 20,467 個で最大であった．ランダムサンプリングの学習パターンを用いた学習結果において，サポートベクトル数が少なく認識率が低下していることから考えれば，提案手法は

表 6.1: LS-Hand の方向特徴での実験結果

Training Dataset	Full	Bridge (Proposed)	Random
#Training Data	518,850	10,592	10,592
Clustered-RNG(s)	–	1.16×10^5	–
Parameter Optimization(s)	6.37×10^6	3.72×10^5	2.15×10^4
SVM Training(s)	5.22×10^4	8.50×10^2	1.72×10^2
Total(s)	6.42×10^6	$\mathbf{4.89 \times 10^5}$	2.17×10^4
#SV	3,186	2,848	883
Error Rate(%)	0.061	0.073	0.34

クラス境界付近のパターンを選択的に抽出しているため，少ない学習パターンでも認識率が保持されていると推測できる．

表 6.1 から，事前選択で学習パターンを 1/50 に削減したことが，SVM の学習の大幅な高速化につながったことが確認できる．具体的には，Clustered-RNG の構築時間を含めた学習時間の合計値で比較して，約 13 倍の高速化効果が得られている．特に，全パターンを用いた場合では 70 日以上を要する SVM の学習が，提案手法により 6 日弱に短縮できることは実用上有効である．one-vs.-all の SVM 識別器はクラスごとに独立して学習できるため，提案手法を分散処理の手法と組み合わせれば SVM の学習に必要な実時間をさらに短縮することも可能である．

ランダムサンプリングした学習パターンでの結果は，学習時間が提案手法よりも短縮されるものの，誤認識率が提案手法に比べて約 0.27 % 増加しているため，学習時間と認識率とのトレードオフを考えれば提案手法は有効性があると判断できる．SVM 識別器のパラメータをより適切に調整するために，グリッドサーチにより探索するパラメータの組み合わせが増加すればこの高速化の効果はより大きくなる．よって，提案手法による SVM の学習高速化の効果は，実用上の有効性が高い．

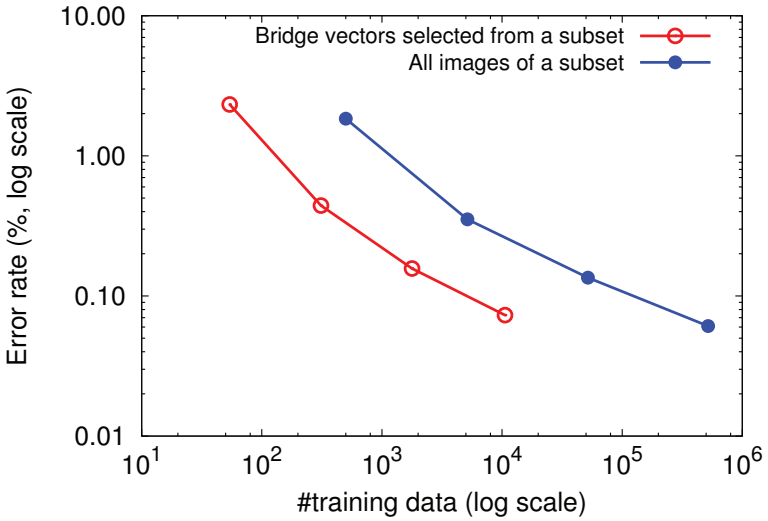


図 6.4: LS-Hand の方向特徴における学習パターン数と認識率の関係

6.4.2　学習パターン数の影響の評価

上述のように，LS-Hand の約 50 万の学習パターンを用いた実験では，事前選択を行ってた学習パターンでも SVM 識別器の認識率は保持される．一方，先行研究 [92, 93, 95, 96] での小規模なデータセットでの実験では，事前選択を行うと SVM の認識率が低下することが確認されている．そこで，学習パターン数と事前選択の有無による SVM の認識率の変化を確認する実験を行った．具体的には，LS-Hand の学習パターンからランダムサンプリングによりパターン数の少ないデータセットを作成し，SVM の学習を行って認識率を確認した．

実験結果を図 6.4 に示す．小規模な学習パターンの場合，事前選択による SVM 識別器の認識率の低下が大きくなることが確認できた．例えば，LS-Hand の 500 個のパターンを学習パターンとした場合，誤認識率は全パターンで 1.84 %，提案手法で 2.33 % であった．また，学習パターン数が増加することで，事前選択の有無による認識率の差が小さくなる傾向も確認できた．提案手法は，適用する学習パターンが大規模化すれば，SVM の学習高速化の効果が高くなるだけでなく，事前選択による認識率の低下が小さくなるため，その有効性が高くなる．

6.4.3　パターン分布構造の影響の評価

6.4.1 で示した表 6.1 の実験結果から，学習パターンが同数でも，その分布構造によってSVM の学習の処理時間が大きく変化することが確認できた．具体的には，クラス境界付近のパターンのみを用いて学習を行った場合，SMO の収束に時間がかかるため，パターン数に対して学習時間が長くなる．この事実から考えれば，クラス境界付近に分布するパターン多い学習パターンセットでは，提案手法を用いて事前選択を行っても高速化の効果が十分ではなくなる可能性がある．そこで，5.5 での実験結果から，方向特徴に比べてクラス境界付近のパターン数が多い分布構造を持つと推測される LS-Hand の輝度特徴を用いて，事前選択の効果に対する特徴量の影響の評価実験を行った．

LS-Hand の輝度特徴を用いたSVM の学習時間を表 6.2 に示す．実験は，LS-Hand のデータセットからランダムサンプリングによりパターン数が少ないサブセットを作成し，そのサブセットを各々学習パターンとしてSVM の学習を行って学習時間と認識率の評価を行った．

表 6.1 と表 6.2 の結果を比較すれば，上述の予測のとおり，提案手法により事前選択を行った場合のSVM の学習時間が，方向特徴では 4.89×10^5 s であったのに対して，輝度特徴では 9.54×10^5 s であり，約 2 倍に増加している．一方，表 6.2 では輝度特徴の場合でも，事前選択によって学習パターン数が約 10 % に削減されることで，SVM の学習時間は 1/5 以下に短縮されており，提案手法の有効性は失われていない．

表 6.2 の学習パターン数と認識率の関係を図 6.5 に示す．同図から，事前選択の影響による認識率の低下はほとんどなく，学習パターン数が増加すればSVM 識別器の認識率が改善していることが確認できる．よって，クラス境界付近に多くのパターンが分布する場合でも，提案手法による事前選択によって，SVM 識別器の認識率を低下させず学習を高速化できる．

また，表 6.1 と表 6.2 の比較から，特徴量の違いにより，サポートベクトル数の多寡が大きく異なっていることが確認できる．具体的には，方向特徴では 3,000 個程

表 6.2: LS-Hand の輝度特徴による実験結果

Training Dataset Size (%)	100	10	1	0.1
Preselection Method	None			
#Training Data	518,850	51,880	5,180	510
Clustered-RNG(s)	–	–	–	–
Parameter Optimization(s)	5.46×10^6	1.62×10^5	3.32×10^3	2.91×10^2
SVM Training(s)	3.73×10^4	1.15×10^3	1.52×10^1	≤ 1
Total(s)	5.50×10^6	1.64×10^5	3.34×10^3	2.92×10^2
#SV	11,741	2,753	768	159
Error Rate(%)	0.085	0.24	0.69	2.2

Training Dataset Size (%)	100	10	1	0.1
Preselection Method	Bridge Vectors (Proposed)			
#Training Data	43,967	6,142	837	107
Clustered-RNG(s)	4.75×10^4	3.55×10^2	5.60×10^0	≤ 1
Parameter Optimization(s)	9.02×10^5	2.14×10^4	9.71×10^2	1.43×10^2
SVM Training(s)	4.34×10^3	7.66×10^1	4.46×10^0	≤ 1
Total(s)	$\mathbf{9.54 \times 10^5}$	$\mathbf{2.18 \times 10^4}$	$\mathbf{9.81 \times 10^2}$	$\mathbf{1.44 \times 10^2}$
#SV	9,102	2,474	562	105
Error Rate(%)	0.081	0.24	0.71	2.2

度であったサポートベクトル数が，輝度特徴では 10,000 個程度に増加している．これは，前述したとおり，方向特徴に比べて輝度特徴ではクラス境界付近のパターン数が多く，識別平面が複雑になっているためだと推測される．また，輝度特徴では全パターンで学習した場合に比べブリッジベクトルを学習パターンとした場合のサ

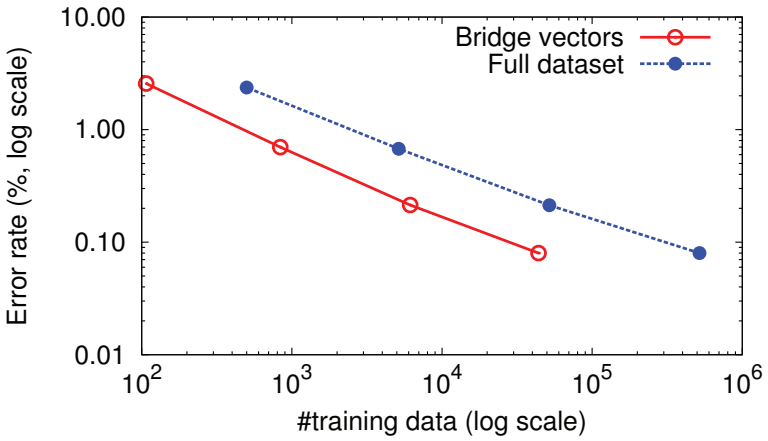


図 6.5: LS-Hand の輝度特徴での認識率

ポートベクトル数が，10 % 以上少なくなっている．すなわち，事前選択を行って学習した SVM 識別器は，全パターンを用いて学習した場合よりも少ないサポートベクトルによって同等の認識率を達成している．

事前選択の有無によって，SVM 識別器にサポートベクトルとして選択されるパターンは変化する．具体的には，図 6.6 に示す包含関係となる．上述したように，事前選択の有無によって，SVM 識別器の認識率はほとんど変化しないため，全パターンを用いて学習した SVM 識別器のサポートベクトルの大半が，事前選択を行った学習パターンを用いて学習した SVM 識別器でも，サポートベクトルとして選択されていると予測される．そこで，事前選択の有無によらずサポートベクトルとなるパターンの割合，すなわち（同図の緑の部分）/（同図の緑＋青の部分）を解析した．

事前選択の有無によらず，サポートベクトルとなるパターンの割合を図 6.7 に示す．同図から，学習パターン数が増加すれば，事前選択の有無によらず，共通のパターンがサポートベクトルとして選ばれる確率が高くなることが確認できる．具体的には，全パターンを学習パターンとした場合には，60 % 以上のパターンが事前選択を行ってもサポートベクトルとして選択されている．

6.2.2 で述べたように，SVM 識別器はサポートベクトルが持つ重み α を適切に設定

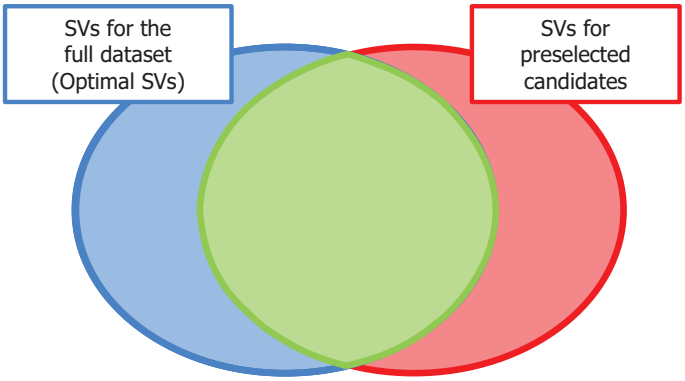


図 6.6: SVM 識別器にサポートベクトルとして選択されるパターンの包含関係

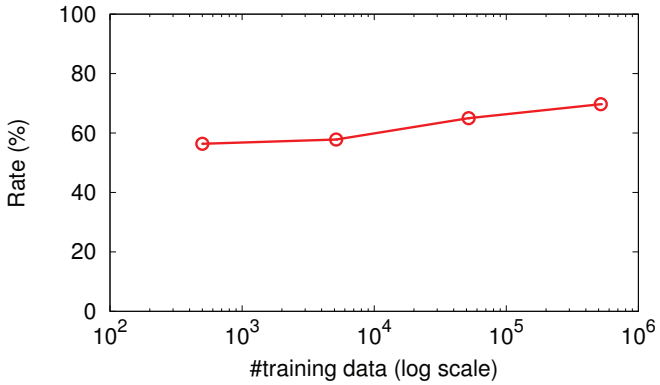


図 6.7: 事前選択の有無によらずサポートベクトルとなるパターンの割合

できれば，異なるサポートベクトル集合を用いて同等の識別平面を表現できる．ここで，識別平面への寄与度が大きいサポートベクトルがブリッジベクトルでなければ，事前選択を行った学習パターンでは，他のパターンを用いて同等の識別平面を構成することとなり，サポートベクトル数は増加すると予測される．このことから考えれば，識別平面への寄与度が大きい，すなわち重み α が大きいサポートベクトルが事前選択の有無によらず共通して選択されているため，少ないサポートベクトル数で同等の認識率が達成できていると推測される．この点について，次節で追加の実験を行って検証する．

6.5 公開データセットMNISTでのSVMの学習高速化の評価実験

前節で示したLS-Handの実験結果から，提案手法による学習パターンの事前選択はSVMの学習高速化に有効であることが検証できた．一方，実験に用いたデータセットが独自の大規模パターン集合であるため，信頼性の点では不十分である．そこで，公開データセットであるMNISTを用いて追加の実験を行い，提案手法の有効性を検証する．加えて，本節ではクラスごとのSVM識別器の解析結果についても示す．

6.5.1 事前選択の効果の評価

公開データセットでの評価実験として，MNISTのデータセットを用いた実験を行った．まず，表6.1と同様に，MNISTの方向特徴を用いた実験結果を表6.3に示す．表6.3から，MNISTでも事前選択の影響による認識率の低下は，ほとんどないことが確認できる．また，SVMの学習高速化についても，Clustered-RNGの構築時間を含めた学習時間の合計値で比較して，約11倍の高速化効果が得られている．この高速化効果は，表6.1に示した方向特徴での結果では，事前選択による高速化効果が約13倍であったことに比べてやや低いものの，同程度の有効性があると判断できる．

一方で，学習パターン数の削減率は，データセットの違いにより顕著な差があらわれている．具体的には，表6.1のLS-Handの結果では事前選択により学習パターン数が1/50に削減できていたのに対して，MNISTでは1/10にしか削減されていない．また，事前選択の有無によってサポートベクトル数が変化せず，ほぼ同数となっている点もLS-Handの実験結果とは異なる傾向である．

図6.4の実験と同様に，MNISTの学習パターンからランダムサンプリングによりパターン数の少ないデータセットを作成し，SVMの学習を行って認識率を評価した結果を図6.8に示す．同図には，比較のために図6.4の結果も付記した．同図から，

表 6.3: MNIST の方向特徴を用いた実験結果

Training Dataset	Full	Bridge (Proposed)	Random
#Training Data	60,000	4,313	4,313
Clustered-RNG(s)	–	1.33×10^3	–
Parameter Optimization(s)	2.61×10^5	2.34×10^4	4.40×10^3
SVM Training(s)	3.02×10^3	1.20×10^2	4.70×10^1
Total(s)	2.64×10^5	$\mathbf{2.48 \times 10^4}$	4.45×10^3
#SV	1,684	1,690	695
Error Rate(%)	0.42	0.47	1.08

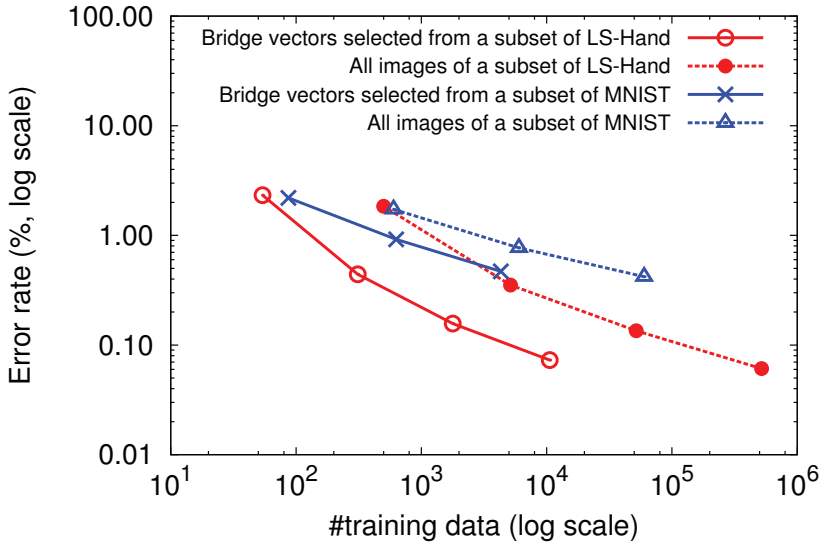


図 6.8: 方向特徴における学習パターン数と認識率の関係

図 6.4 に示した LS-Hand の方向特徴での結果と同様に事前選択の影響による認識率の低下がなく，パターン数が増加するにしたがって認識率が改善していることが確認できる．

表 6.4: MNIST の輝度特徴を用いた実験結果

Training Dataset	Full	Bridge (Proposed)	Random
#Training Data	60,000	5,313	5,313
Clustered-RNG(s)	–	7.61×10^2	–
Parameter Optimization(s)	1.20×10^6	7.57×10^4	2.42×10^4
SVM Training(s)	4.48×10^3	2.06×10^2	6.71×10^1
Total(s)	1.20×10^6	$\mathbf{7.67 \times 10^4}$	2.43×10^4
#SV	3,498	2,712	1,200
Error Rate(%)	0.40	0.40	0.98

次に，表 6.2 と同様に MNIST の輝度特徴を用いた実験結果を表 6.4 に示す．結果から，上述した MNIST の方向特徴での結果と同様に，事前選択の影響による認識率の低下がなく，SVM の学習高速化効果がえられていることが確認できる．また，事前選択による学習パターン数の削減についても，方向特徴での結果と同様に 1/10 にしか削減されていない．

提案手法により事前選択される学習パターンは，パターンの分布構造のみによって一意に定まるため，上述した LS-Hand と MNIST での学習パターン数の削減効果の違いは，データセットの持つ分布構造の違いに起因すると推測される．そこで，MNIST のパターン分布についても，Clustered-RNG による可視化を行った．

MNIST の輝度特徴と方向特徴による Clustered-RNG を，図 6.9 に示す．また，同図から解析したクラスごとのエッジ数を図 6.10 に，アウトライア数を図 6.11 に示す．結果から，MNIST では特徴量の違いによる分布構造の変化が異なることが確認できる．具体的には，LS-Hand では輝度特徴に比べて，方向特徴では異クラス間のエッジ数が少ない，すなわち方向特徴を用いることで各クラスが密な分布構造となっていたのに対し，MNIST では，方向特徴を用いても異クラス間のエッジ数が多く，

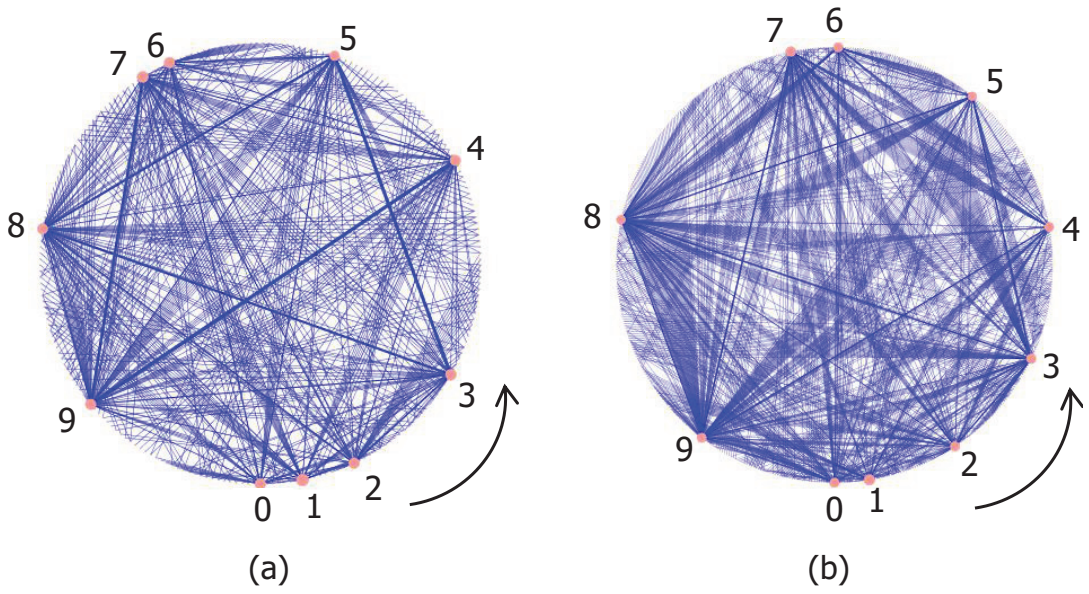


図 6.9: MNIST の Clustered-RNG (a) 輝度特徴. (b) 方向特徴

各クラスの分布が広範な接面を持つことが，図 5.15 と図 6.10 の比較から確認できる．また，パターン数が LS-Hand が約 50 万，MNIST が 6 万であることを考慮すれば，LS-Hand に比べて MNIST では多くのアウトライアが存在することが，図 5.14 と図 6.11 の比較から確認できる．このように MNIST は LS-Hand に比べて複雑な分布構造を持つため，提案手法による学習パターンの削減効果が低く，SVM の学習高速化の効果も低くなったと推測できる．

この分布構造の違いは，SVM 識別器の認識率にも表れている．具体的には，図 6.8 に示したように，6 万程度の学習パターンを全て用いて学習した場合の方向特徴による SVM 識別器の誤認識率が，LS-Hand では 0.13 % であったのに対し，MNIST では 0.42 % であり，MNIST の方が認識率が低い．このように，データセットや特徴量の違いにより識別器の認識率が異なる場合，特徴量や識別器の妥当性の検証が課題となる．この課題に対して，第 5 章で述べた Clustered-RNG による分布構造の解析を行えば，上述したように認識率以外の定量的な指標によって，特徴量や識別器の妥当性が検証できる．

上述した MNIST での特徴量による Clustered-RNG のクラスタ間エッジ数の差異

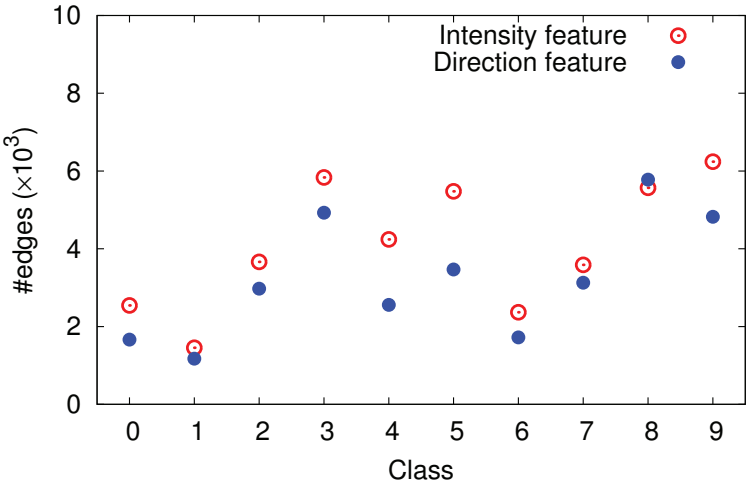


図 6.10: MNIST のクラスごとのエッジ数

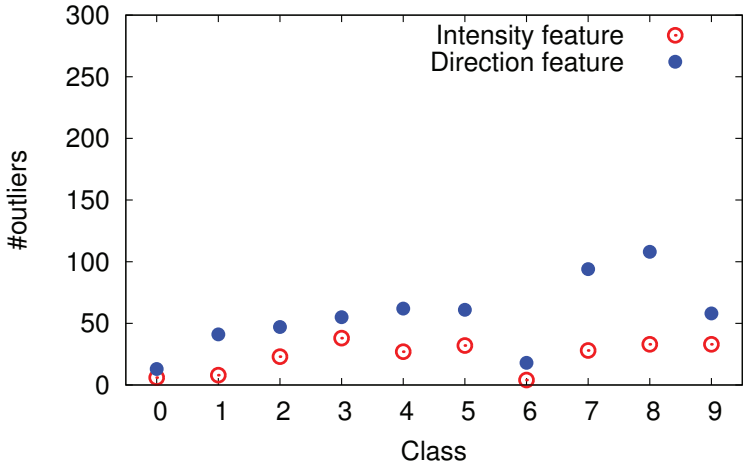


図 6.11: MNIST のクラスごとのアウトライア数

は，特徴量の性能に対する認識の点でも興味深い．具体的には，文献 [97, 98] において文字認識への適用を目的として検討された方向特徴に比べて，非常に単純な輝度特徴の方が，クラス境界付近のパターン数が少ない結果となった．実際に，表 6.3 と表 6.4 の結果からも，輝度特徴を用いた SVM 識別器の方がわずかながら高い認識

率を示している．このように，従来検討されてきた複雑な特徴量比べ，単純な特徴量でも同程度の認識率が示される傾向は，HOG 特徴 [101] での輝度勾配特徴の有効性の再評価やバイナリ特徴 [102, 103] による高速化などの研究事例で示された傾向とも合致している．

6.5.2 事前選択の有無による SVM 識別器の変化の解析

本項では，MNIST の輝度特徴を学習パターンとして事前選択の有無による SVM 識別器の変化を詳細に解析し，提案手法の妥当性を検証する．具体的には，これまでの実験結果では，各クラスの SVM 識別器の平均認識率を基準に評価実験を行ってきたのに対し，本項では，クラスごとの SVM 識別器の認識率を個別に解析する．また，事前選択の有無によるサポートベクトルの変化についても詳細に解析する．

MNIST のデータセットを用いて SVM の学習を行った結果を表 6.5 に示す．まずクラスごとの認識率に着目すれば，事前選択による認識率の変化は図 6.12 となっている．結果から，全てのクラスで事前選択による認識率の低下がなく，ランダムサンプリングした場合に比べ，認識率が高いことが確認できる．また，事前選択によるクラスごとの学習パターン数と認識率の関係は，図 6.13 となっている．結果から，学習パターン数の削減効果はクラスごとに大きな差はないことが確認できる．最も学習パターン数の削減効果高かったのはクラス “1” であり 6 万の学習パターンが 1,977 個に削減できていた．

次に，クラスごとのサポートベクトル数の着目すれば，事前選択によるサポートベクトル数の変化は図 6.14 となっている．同図から，全てのクラスで提案手法による事前選択を行うことでサポートベクトル数が削減できており，識別器としての性能が改善されていることが確認できる．一方で，ランダムサンプリングによって学習パターンを削減した場合は，サポートベクトル数が少なくなるものの，上述したように識別器の認識率も低下する．このように，ランダムサンプリングされた少ない学習パターンでは，各クラスの分布構造が表現されないため，正しい識別平面が学習できない．

表 6.5: MNIST の輝度特徴における実験結果

Class	Full			**Bridge (Proposed)**			Random		
	#Training Data	#SV	Error Rate(%)	#Training Data	#SV	Error Rate(%)	#Training Data	#SV	Error Rate(%)
0	60,000	2,341	0.20	3,681	1,575	0.20	3,681	504	0.68
1		1,422	0.25	1,977	1,023	0.27	1,977	354	0.42
2		3,737	0.33	4,930	2,844	0.34	4,930	1,072	0.89
3		4,711	0.45	7,043	3,779	0.47	7,043	1,566	1.22
4		3,733	0.38	5,768	3,371	0.37	5,768	1,074	1.15
5		4,629	0.44	6,929	3,073	0.43	6,929	1,815	0.73
6		2,225	0.25	3,247	1,632	0.27	3,247	674	0.65
7		3,081	0.43	4,822	2,287	0.44	4,822	1,073	0.97
8		4,701	0.58	6,951	3,791	0.63	6,951	1,985	1.48
9		4,401	0.70	7,778	3,745	0.61	7,778	1,881	1.56
Ave.	60,000	3,497	0.40	5,313	2,712	0.40	5,313	1,200	0.98

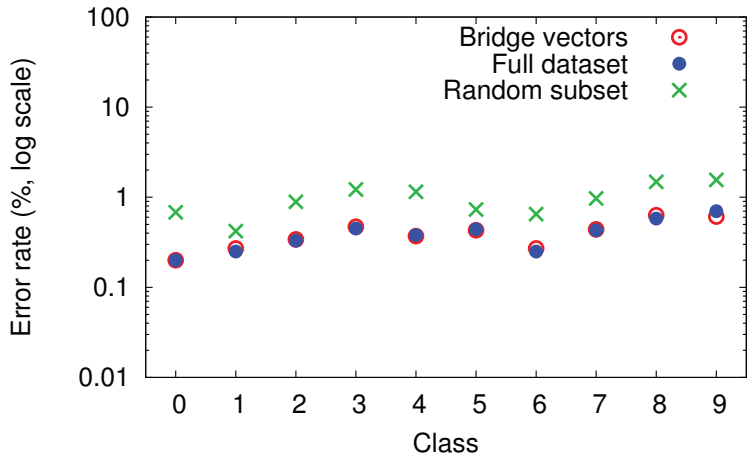


図 6.12: MNIST の輝度特徴におけるクラスごとの認識率

また，クラスごとのサポートベクトルの重み α についても解析を行った．図 6.14 に示した結果から，事前選択を行ってもサポートベクトルとなるパターンは，他のサポートベクトルのパターンに比べ，重み α が大きいことが確認できる．言い換え

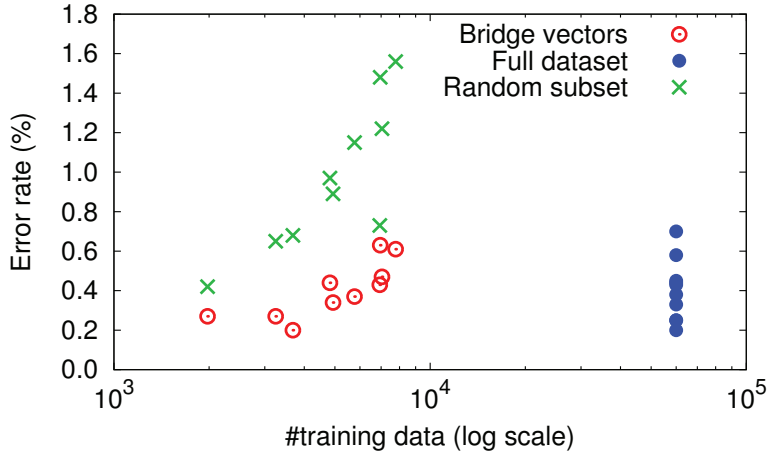


図 6.13: MNIST の輝度特徴におけるクラスごとの学習パターン数と認識率の関係

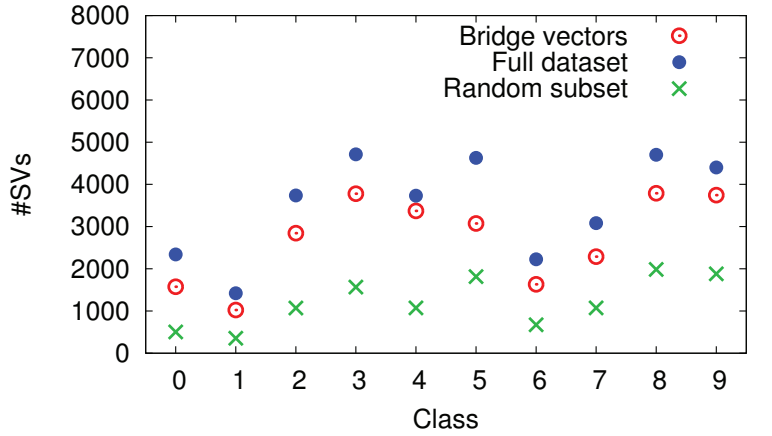


図 6.14: MNIST の輝度特徴におけるクラスごとのサポートベクトル数

れば，クラス間の識別平面を構成するために重要となるパターンは，事前選択の有無によらず SVM 識別器のサポートベクトルとして選択される．事前選択を行って学習した SVM 識別器は，全パターンを用いて学習した SVM 識別器とは一部異なるサポートベクトルを用いて，同等の識別平面を構成する．

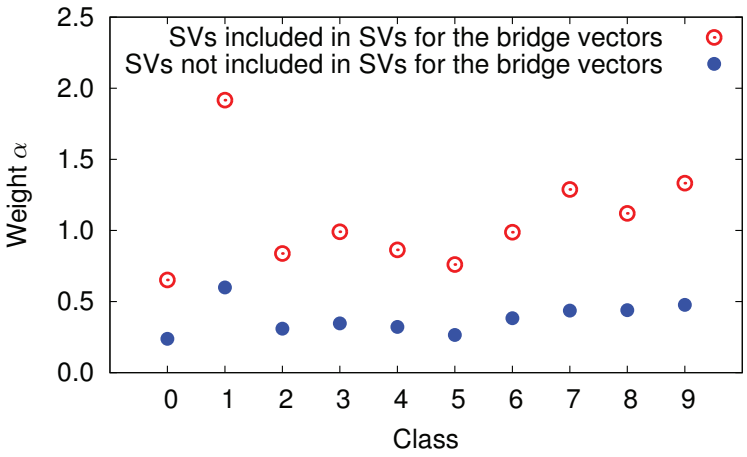


図 6.15: MNIST の輝度特徴におけるクラスごとのサポートベクトルの重み α の平均値

6.6 まとめ

本章では，第 5 章で提案した Clustered-RNG による分布構造解析法の認識応用の例として，SVM の学習高速化のための学習パターン事前選択の手法を提案した．提案手法は，Clustered-RNG による分布構造の解析結果を基に，学習パターンから異クラス間エッジ（クラスラベルが異なるパターンを結合するエッジ）を持つパターンを事前選択することで，SVM の学習を高速化する．

実験結果から，提案手法が大規模な学習パターンでの SVM の学習高速化に有効であることが確認できた．小規模な学習パターンの RNG から学習パターンを事前選択している先行研究では，事前選択により SVM 識別器の認識率が低下している．これに対して，約 50 万の手書き数字画像のデータセットと MNIST の公開データセットを用いた実験を行い，データセットのサイズの差異が結果に及ぼす影響を複数の実験により検証し，提案手法の有効性を示した．提案手法は，SVM 識別器の認識率を大きく低下させることなく，学習時間を大幅に短縮できる．特に，学習パターンが大規模化な場合には，認識率の低下がさらに小さくなることを示した．

本章では，複数のデータセットと特徴量ごとの SVM 識別器の認識率を比較することで，第 5 章で述べた Clustered-RNG による分布構造解析の有用性についても考察した．Clustered-RNG により解析できる異クラス間の隣接性やアウトライア数の定量的な指標により，特徴量や識別器の妥当性が検証できる．この解析結果に基づいてデータセットの分布構造や特徴量の特性を検討すれば，複数の特徴量の組み合わせや，特徴量の改良などが容易となり，識別器の認識率の改善も可能となるはずである．

第7章

結論

7.1 本研究の要約

本研究では，大規模パターン集合の分布構造解析という課題に対し，MST および RNG をクラスタ化した Clustered-RNG を用いたネットワーク表現による分布構造解析の手法を提案した．また，大規模パターン集合に対して，高速かつ省メモリに RNG を構築する手法について提案した．加えて，ネットワーク表現による分布構造解析法の認識応用の例として，学習パターンの事前選択による SVM の学習高速化の実例を示した．

第 1 章では，まず，ネットワーク解析を用いたパターンの分布解析に関する学術的背景について述べ，本研究の目的を明らかにした．本研究の目的は，パターンの分布の真の姿の解明に近づくことである．さらに，大規模パターン集合の分布構造解析を行うために必要となる条件を整理し，それを実現するための大まかなアプローチについて論じた．パターンの真の分布構造を明らかにするためには，二つの条件が必要となる．第一の条件は，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．第二の条件は，利用する分布構造解析手法が，パターン間の位置関係を正確に把握できることである．この二つの条件に対して，本研究では数字画像の大規模パターン集合とネットワーク表現による分布解析法を用いて対応し，大量の画像から構成される大規模パターン集合の分布構造の解析という課題に取り組んだ．加えて，提案手法の解析結果を利用した学習パターンの事前選択が SVM の学習高速化に有効である可能性を指摘した．

第 2 章では，本研究で提案した内容に関連する研究について整理し，本研究の位置づけを行った．次元削減や確率密度分布の推定による分布解析法では，何らかのモデルによる近似や低次元化を行うため，パターン間の位置関係を正確に把握でき

ない．これに対して，パターン間距離を考慮したネットワーク表現を用いた分布構造解析法では，パターンの分布構造がネットワークによって完全に保存され，解析結果がパターンの分布構造のみによって一意に定まるため，パターン間の位置関係を正確に把握できる．また，SVM の学習高速化について提案手法と先行研究の違いを示した．

第 3 章では，本研究で分布構造を解析する大規模数字画像データセットについて述べた．数字画像は，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可能であり，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．そして，数字であれば，各パターンに付与されたカテゴリ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可能となる．

第 4 章では，MST を用いてパターンの分布構造を解析する手法を提案した．提案手法では，まず各パターンを 1 ノードとし，パターン間の距離をエッジの重みとして MST を構築する．そして，その MST の構造からパターンの分布構造を解析する．このようにして構築された MST は，パターン空間における分布構造をノード間の近傍関係として表現する．このネットワークの構造を解析すれば，特徴的な大規模パターン集合の分布構造が明らかとなる．この手法の特徴は，パターンの分布構造の解析を，MST におけるノードやエッジの特徴などから得られる指標を用いて定量的に行うことである．評価実験からは，大規模パターン集合の分布構造にクラスごとおよび同一クラス内での粗密があることが確認できた．また，隣接ノードのクラスによるノードの分類を行えば，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 およびそのクラスタ化，すなわちエッジで結合された同じクラスラベルを持つパターンを統合して 1 つのノードとして表現したクラスタ木による解析結果により，多クラス間の近傍関係を明らかにした．

第 5 章では第一に，クラスタ化された RNG を用いてパターンの分布構造を解析する手法を提案した．提案手法では，まず各パターンを 1 ノードとし，その近傍関係によりエッジを付与して RNG を構築する．そして，その RNG をクラスタ化，す

なわちエッジで結合された同じクラスラベルを持つパターンを統合し，1 つのノードとして表現した Clustered-RNG を用いてパターンの布構造を解析する．RNG は，閉路を許すグラフ構造であるため，パターン間の近傍関係を余さず表現できる．また，RNG をクラスタ化した Clustered-RNG の構造から，単一のクラス分布だけでなく，多クラス分布間の隣接性やアウトライアの分布などの関係性が，同時に解析できる．約 50 万の活字数字画像と約 50 万の手書き数字画像のデータセットを用いた評価実験からは，提案手法により，大規模パターン集合の多クラス分布間の隣接性やアウトライアの分布などの関係性が可視化され，定量的に解析できることが確認できた．そして，複数の特徴量を用いた実験を行い，提案手法によって明らかとなる分布構造の性質と有効性について示した．

同じ第 5 章では第二に，高次元空間中の大規模パターン集合に対して，高速かつ省メモリに RNG を構築する手法を提案した．提案手法では，ノードをランダムにグルーピングし，各グループに対する各ノードからの最小距離を基準にエッジ付与の判定の回数を削減することで，RNG を効率的に構築する．評価実験では，複数のデータセットと特徴量を用いて Clustered-RNG 構築の処理時間を評価し，提案手法によって Clustered-RNG が効率的に構築できることを確認した．

第 6 章では，第 5 章で提案した Clustered-RNG による分布構造解析法の認識応用の例として，SVM の学習高速化のための学習パターン事前選択の手法を提案した．提案手法は，Clustered-RNG による分布構造の解析結果を基に，学習パターンから異クラス間エッジを持つパターンを事前選択する．約 50 万の手書き数字画像のデータセットと MNIST の公開データセットおよび複数の特徴量を用いた評価実験から，提案手法の有効性を確認した．具体的には，Clustered-RNG の解析結果から学習パターンを事前選択すれば，SVM 識別器の認識率を大きく低下させることなく学習時間を大幅に短縮できる．特に学習パターンが大規模化な場合には，認識率の低下がさらに小さくなることを示した．

7.2 今後の課題

今後の課題として第一に，よりクラス間の関係性に焦点を当てた解析を行なうことが挙げられる．高次元空間中での文字のクラス間の関係性がより明確なものとなれば，サポートベクトル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 [104]，などの展開が考えられる．さらには，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 [105] 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．

第二の課題として，一般物体など文字以外のパターンの分布構造の解析が挙げられる．本研究では扱わなかった一般物体のデータセットに提案手法を適用すれば，大規模なデータセットをより効率的に学習することが可能となり，パターン認識に新たな理論的基盤を与えられると考えられる．近年，急速に整備されている非常に多くのクラスのパターンを含んだデータセットの分布構造を解析し，クラスラベルの妥当性やクラス間の隣接性を吟味すれば，より効率的な識別器を構成できるはずである．

第三の課題として，Clustered-RNG による解析結果のさらなる応用の検討が挙げられる．本研究での提案手法を用いれば多クラスの隣接性を同時に解析できる．よって，多クラスの分布構造から識別が容易なクラス対や識別が困難なクラス対を抽出できる．この特長を応用すれば，binary tree 型の SVM [106, 107] のように，複数の識別器を組み合わせて coarse-to-fine にクラスラベルを推定するマルチクラス識別器をより効率的に構成できると期待している．

謝辞

本研究の遂行ならびに本論文をまとめるにあたり，多大なる御指導を賜りました九州大学大学院 システム情報科学研究院 情報知能工学部門 教授　内田誠一先生に謹んで感謝の意を表します．氏の適切な御指導により，本研究を遂行し，学位論文としてまとめることができました．時に応じて，厳しく指摘してくださったこと，優しく励ましてくださったことは，私が日々努力する糧となりました．ここに，心より深く感謝の意を表します．また，本論文をご精読頂き有用なコメントをいただきました本学，谷口倫一郎教授，諸岡健一准教授に深謝致します．グローリー株式会社 研究開発センターの黒岩素子さん，石田良介君には実験の実施や日常の議論を通じて多くの知識や示唆をいただきました．深く感謝します．また，ヒューマンインタフェース研究室諸氏にも御支援をいただきました．心より御礼申し上げます．

本研究を遂行するにあたり，グローリー株式会社の方々に多大な御理解と御支援をいただきました．尾上広和社長には，社会人博士課程を通して本研究に取り組む機会を与えていただきました．また，三和元純専務には，折に触れて御助言をいただきました．そして，研究開発センター 亀山博史センター長には，終始様々な御助言と御配慮をいただきました．さらに，研究開発センターの先輩，同僚，後輩の方々には，日々様々な御協力をいただいただけでなく，多くの励ましの言葉をいただきました．ここに記して心より感謝申し上げます．

最後に，これまで私をあたたかく応援し，支えてくれた家族に深く感謝します．

参考文献

[1] I. Jolliffe, Principal component analysis, Wiley Online Library, 2002.

[2] 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，2007–2008.

[3] D.J. Watts and S.H. Strogatz, “Collective dynamics of ‘ small-world ’ networks,” Nature，vol.393，no.6684，pp.440–442，1998.

[4] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” Science, vol.286, no.5439, pp.509–512, 1999.

[5] R. Albert and A.-L. Barabási, “Statistical mechanics of complex networks,” Reviews of Modern Physics, vol.74, no.1, pp.47–97, 2002.

[6] 増田直紀，今野紀雄，複雑ネットワーク—基礎から応用まで，近代科学社，2010.

[7] L.F. Costa, F.A. Rodrigues, G. Travieso, and P.R. Villas Boas, “Characterization of complex networks: A survey of measurements,” Advances in Physics, vol.56, no.1, pp.167–242, 2007.

[8] A. Torralba, R. Fergus, and W.T. Freeman, “80 million tiny images: A large data set for nonparametric object and scene recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.30, no.11, pp.1958–1970, 2008.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.248–255, 2009.

[10] J. Xiao, J. Hays, K.A. Ehinger, A. Oliva, and A. Torralba, “SUN database: Large-scale scene recognition from abbey to zoo,” Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3485–3492, 2010.

[11] F. Schroff, A. Criminisi, and A. Zisserman, “Harvesting image databases from the web,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.33, no.4, pp.754–766, 2011.

[12] D. Eppstein, M.S. Paterson, and F.F. Yao, “On nearest-neighbor graphs,” Discrete & Computational Geometry, vol.17, no.3, pp.263–282, 1997.

[13] G.T. Toussaint, “The relative neighbourhood graph of a finite planar set,” Pattern Recognition, vol.12, no.4, pp.261–268, 1980.

[14] B.V. Dasarathy, “Minimal Consistent Set (MCS) identification for optimal nearest neighbor decision systems design,” IEEE Transactions on Systems, Man, and Cybernetics, vol.24, no.3, pp.511–517, 1994.

[15] 和田俊和，“空間分割を用いた識別と非線形写像の学習:(1) 空間分割による最近傍識別の高速化,” 情報処理，vol.46，no.8，pp.912–918，2005.

[16] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol.20, no.3, pp.273–297, 1995.

[17] A. Krizhevsky, I. Sutskever, and G.E. Hinton, “ImageNet classification with deep convolutional neural networks,” Proceedings of Advances in Neural Information Processing Systems (NIPS), pp.1097–1105, 2012.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., "ImageNet large scale visual recognition challenge," International Journal of Computer Vision, vol.115, no.3, pp.211–252, 2015.

[19] A. Krizhevsky and G. Hinton, Learning multiple layers of features from tiny images, Master's thesis, Department of Computer Science, University of Toronto, 2009.

[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of IEEE, vol.86, no.11, pp.2278–2324, 1998.

[21] G.F. Davis, M. Yoo, and W.E. Baker, "The small world of the american corporate elite, 1982-2001," Strategic Organization, vol.1, no.3, pp.301–326, 2003.

[22] H. Ebel, L.-I. Mielsch, and S. Bornholdt, "Scale-free topology of e-mail networks," Physical Review E of American Physical Society, vol.66, no.3, pp.035103:1–035103:4, 2002.

[23] R. Albert, H. Jeong, and A.-L. Barabási, "Internet: Diameter of the world-wide web," Nature, vol.401, no.6749, pp.130–131, 1999.

[24] E. Ravasz and A.-L. Barabási, "Hierarchical organization in complex networks," Physical Review E of American Physical Society, vol.67, no.2, pp.026112:1–026112:7, 2003.

[25] M.E. Newman, "Co-authorship networks and patterns of scientific collaboration," Proceedings of National Academy of Sciences, vol.101, no.S1, pp.5200–5205, 2004.

[26] A.-L. Barabási, et al., “Scale-free networks: A decade and beyond,” Science, vol.325, no.5939, pp.412–413, 2009.

[27] L.A.N. Amaral, A. Scala, M. Barthelemy, and H.E. Stanley, “Classes of small-world networks,” Proceedings of National Academy of Sciences, vol.97, no.21, pp.11149–11152, 2000.

[28] 地球観測データ統融合連携研究機構，“地球環境情報統融合プログラム (DIAS),” http://www.diasjp.net/.

[29] 日本 DNA データバンク，“DDBJ | DNA Data Bank of Japan,” http://www.ddbj.nig.ac.jp/.

[30] S. Hachul and M. Jünger, “Drawing large graphs with a potential-field-based multilevel algorithm,” Proceedings of International Symposium on Graph Drawing (GD), pp.285–295, 2004.

[31] Y. Koren, L. Carmel, and D. Harel, “ACE: A fast multiscale eigenvectors computation for drawing huge graphs,” Proceedings of IEEE Information Visualization Conference (INFOVIS), pp.137–144, 2002.

[32] 三末和男，“ネットワークの可視化技術: 大規模ネットワークと動的ネットワークへの挑戦,” 電子情報通信学会誌，vol.92，no.2，pp.112–117，2009.

[33] K. Boitmanis, U. Brandes, and C. Pich, “Visualizing internet evolution on the autonomous systems level,” Proceedings of International Symposium on Graph Drawing (GD), pp.365–376, 2007.

[34] E. Cerami, J. Gao, U. Dogrusoz, B.E. Gross, S.O. Sumer, B.A. Aksoy, A. Jacobsen, C.J. Byrne, M.L. Heuer, E. Larsson, et al., “The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data,” Cancer Discovery, vol.2, no.5, pp.401–404, 2012.

[35] T. Raeder and N.V. Chawla, “Market basket analysis with networks,” Social Network Analysis and Mining, vol.1, no.2, pp.97–113, 2011.

[36] G. Toussaint, “Geometric proximity graphs for improving nearest neighbor methods in instance-based learning and data mining,” International Journal of Computational Geometry & Applications, vol.15, no.02, pp.101–150, 2005.

[37] D.W. Matula and R.R. Sokal, “Properties of Gabriel graphs relevant to geographic variation research and the clustering of points in the plane,” Geographical Analysis, vol.12, no.3, pp.205–222, 1980.

[38] G.T. Toussaint, “Applications of the relative neighbourhood graph,” International Journal of Advances in Computer Science & Its Applications, vol.4, no.3, pp.77–85, 2014.

[39] R.C. Prim, “Shortest connection networks and some generalizations,” Bell System Technical Journal, vol.36, no.6, pp.1389–1401, 1957.

[40] R.L. Graham and P. Hell, “On the history of the minimum spanning tree problem,” Annals of the History of Computing, vol.7, no.1, pp.43–57, 1985.

[41] P.F. Felzenszwalb and D.P. Huttenlocher, “Efficient graph-based image segmentation,” International Journal of Computer Vision, vol.59, no.2, pp.167–181, 2004.

[42] 角田夏貴，渡辺俊典，菅原　研，“最小全域木の適応しきい値切断による画像分割方式,” 電子情報通信学会論文誌 D，vol.87，no.2，pp.586–594，2004.

[43] O. Grygorash, Y. Zhou, and Z. Jorgensen, “Minimum spanning tree based clustering algorithms,” Proceedings of IEEE International Conference on Tools with Artificial Intelligence (ICTAI), pp.73–81, 2006.

[44] N. Päivinen, “Clustering with a minimum spanning tree of scale-free-like structure,” Pattern Recognition Letters, vol.26, no.7, pp.921–930, 2005.

[45] 井上光平，原　健二，浦浜喜一，“最小木フィルタによるエッジ保存インパルス雑音除去,” 映像情報メディア学会誌，vol.66，no.8，pp.J287–J289，2012.

[46] M.J. Berger, “Data structures for adaptive grid generation,” SIAM Journal on Scientific and Statistical Computing, vol.7, no.3, pp.904–916, 1986.

[47] C.T. Zahn, “Graph-theoretical methods for detecting and describing gestalt clusters,” IEEE Transactions on Computers, vol.100, no.1, pp.68–86, 1971.

[48] 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,” Proceedings of National Academy of Sciences, vol.104, no.22, pp.9404–9409, 2007.

[49] 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, 2002.

[50] R. Urquhart, “Graph theoretical clustering based on limited neighbourhood sets,” Pattern Recognition, vol.15, no.3, pp.173–187, 1982.

[51] F. Rayar, S. Barrat, F. Bouali, and G. Venturini, “An approximate proximity graph incremental construction for large image collections indexing,” Proceedings of International Symposium on Methodologies for Intelligent Systems (ISMIS), pp.59–68, 2015.

[52] C.D. Correa and P. Lindstrom, “Locally-scaled spectral clustering using empty region graphs,” Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp.1330–1338, 2012.

[53] J.S. Sánchez, F. Pla, and F.J. Ferri, "On the use of neighbourhood-based non-parametric classifiers," Pattern Recognition Letters, vol.18, no.11, pp.1179–1186, 1997.

[54] D.A. Zighed, S. Lallich, and F. Muhlenbach, "A statistical approach to class separability," Applied Stochastic Models in Business and Industry, vol.21, no.2, pp.187–197, 2005.

[55] H. Ueda and T. Takeuchi, "The relative neighborhood graph for estimating two-dimensional voids in the cold dark matter universe," Publications of the Astronomical Society of Japan, vol.58, no.2, pp.283–290, 2006.

[56] M. Ichino and H. Yaguchi, "An apparent simplicity appearing in pattern classification problems," Pattern Recognition, vol.33, no.9, pp.1467–1474, 2000.

[57] D. Watanabe, "A study on analyzing the grid road network patterns using relative neighborhood graph," Proceedings of International Symposium on Operations Research and Its Applications (ICORES), pp.112–119, 2010.

[58] A. Adamatzky, "Developing proximity graphs by Physarum polycephalum: Does the plasmodium follow the toussaint hierarchy?," Parallel Processing Letters, vol.19, no.01, pp.105–127, 2009.

[59] J.W. Jaromczyk and G.T. Toussaint, "Relative neighborhood graphs and their relatives," Proceedings of IEEE, vol.80, no.9, pp.1502–1517, 1992.

[60] B.K. Bhattacharya, R.S. Poulsen, and G.T. Toussaint, Application of proximity graphs to editing nearest neighbor decision rule, Proceedings of International Symposium on Information Theory, 1981.

[61] C.B. Barber, D.P. Dobkin, and H. Huhdanpaa, “The quickhull algorithm for convex hulls,” ACM Transactions on Mathematical Software (TOMS), vol.22, no.4, pp.469–483, 1996.

[62] J. Katajainen, O. Nevalainen, and J. Teuhola, “A linear expected-time algorithm for computing planar relative neighbourhood graphs,” Information Processing Letters, vol.25, no.2, pp.77–86, 1987.

[63] J. Katajainen and O. Nevalainen, “An almost naive algorithm for finding relative neighbourhood graphs in L_p metrics,” Informatique Théorique et Applications, vol.21, no.2, pp.199–215, 1987.

[64] K.J. Supowit, “The relative neighborhood graph, with an application to minimum spanning trees,” Journal of ACM, vol.30, no.3, pp.428–448, 1983.

[65] P.K. Agarwal and J. Mataušek, “Relative neighborhood graphs in three dimensions,” Proceedings of ACM-SIAM Symposium on Discrete Algorithms, pp.58–65, 1992.

[66] 坂垣内洵也，加藤丈和，和田俊和，“修正相対近傍グラフを用いたターゲット追跡と認識,” 情報処理学会研究報告, CVIM，vol.2004，no.91，pp.121–128，2004.

[67] H. Hacid and A.D. Zighed, “An effective method for locally neighborhood graphs updating,” Proceedings of International Conference on Database and Expert Systems Applications (DEXA), pp.930–939, 2005.

[68] 吉田哲也，H. Hacid，“高次元データへのアクセス回数を削減する近傍グラフ更新手法,” データ工学ワークショップ論文集，pp.1–9，2007.

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

[70] T.F. Cox and M.A. Cox, Multidimensional scaling, CRC press, 2000.

[71] 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, 2000.

[72] J. Shi and J. Malik, “Normalized cuts and image segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.22, no.8, pp.888–905, 2000.

[73] A.Y. Ng, M.I. Jordan, Y. Weiss, et al., “On spectral clustering: Analysis and an algorithm,” Proceedings of Advances in Neural Information Processing Systems (NIPS), pp.849–856, 2002.

[74] A. Najafi, A. Joudaki, and E. Fatemizadeh, “Nonlinear dimensionality reduction via path-based isometric mapping,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.38, no.7, pp.1452–1464, 2016.

[75] 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，2003.

[76] 栗田多喜夫，“パターン認識とニューラルネットワーク，” "http://home.hiroshima-u.ac.jp/tkurita/lecture/prnn.pdf，2002.

[77] 岩田具治，斉藤和巳，上田修功，“パラメトリック埋め込み法によるクラス構造の可視化，” 情報処理学会論文誌，vol.46，no.9，pp.2337–2346，2005.

[78] G.E. Hinton and S.T. Roweis, “Stochastic neighbor embedding,” Proceedings of Advances in Neural Information Processing Systems (NIPS), pp.833–840, 2002.

[79] 山田武士，斉藤和己，上田修功，“クロスエントロピー最小化に基づくネットワークデータの埋め込み,” 情報処理学会論文誌，vol.44，no.9，pp.2401–2408，2003.

[80] 岩田具治，斉藤和巳，“パラメトリック埋め込み法を用いた分類器の視覚的解析,” 情報処理学会論文誌，vol.48，no.12，pp.4012–4022，2007.

[81] I.S. Dhillon, D.S. Modha, and W.S. Spangler, “Class visualization of high-dimensional data with applications,” Computational Statistics & Data Analysis, vol.41, no.1, pp.59–90, 2002.

[82] R.A. Fisher, “The use of multiple measurements in taxonomic problems,” Annals of Eugenics, vol.7, no.2, pp.179–188, 1936.

[83] B.E. Boser, I.M. Guyon, and V.N. Vapnik, “A training algorithm for optimal margin classifiers,” Proceedings of Workshop on Computational Learning Theory, pp.144–152, 1992.

[84] J.C. Platt, “12 fast training of support vector machines using sequential minimal optimization,” Advances in Kernel Methods, pp.185–208, 1999.

[85] A.K. Menon, “Large-scale support vector machines: Algorithms and theory,” Research Exam, University of California, pp.1–17, 2009.

[86] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, A practical guide to support vector classification, Technical Report, National Taiwan University, 2003.

[87] K. Ogawa, Y. Suzuki, and I. Takeuchi, “Safe screening of non-support vectors in pathwise SVM computation.,” Proceedings of International Conference on Machine Learning (ICML), pp.1382–1390, 2013.

[88] L.E. Ghaoui, V. Viallon, and T. Rabbani, Safe feature elimination for the lasso and sparse supervised learning problems, arXiv preprint arXiv:1009.4219, 2010.

[89] R. Koggalage and S. Halgamuge, "Reducing the number of training samples for fast support vector machine classification," Neural Information Processing-Letters and Reviews, vol.2, no.3, pp.57–65, 2004.

[90] J. Chen, C. Zhang, X. Xue, and C.-L. Liu, "Fast instance selection for speeding up support vector machines," Knowledge-Based Systems, vol.45, pp.1–7, 2013.

[91] J. Wang, P. Neskovic, and L.N. Cooper, "Selecting data for fast support vector machines training," Trends in Neural Computation, pp.61–84, Springer, 2007.

[92] N. Jankowski and M. Grochowski, "Comparison of instances seletion algorithms I. Algorithms survey," Proceedings of Artificial Intelligence and Soft Computing (ICAISC), pp.598–603, 2004.

[93] M. Grochowski and N. Jankowski, "Comparison of instance selection algorithms II. Results and comments," Proceedings of Artificial Intelligence and Soft Computing (ICAISC), pp.580–585, 2004.

[94] W. Zhang and I. King, "A study of the relationship between support vector machine and gabriel graph," Proceedings of International Joint Conference on Neural Networks (IJCNN), vol.1, pp.239–244, 2002.

[95] W. Zhang and I. King, "Locating support vectors via β-skeleton technique," Proceedings of International Conference on Neural Information Processing (ICONIP), vol.3, pp.1423–1427, 2002.

[96] X. Liu, J.F. Beltran, N. Mohanchandra, and G.T. Toussaint, "On speeding up support vector machines: Proximity graphs versus random sampling for pre-

selection condensation," Proceedings of International Conference on Computer Science and Mathematics, vol.73, pp.1037–1044, 2013.

[97] 水上嘉樹，古賀和利，"線素方向特徴を用いた変位抽出を行う手書き漢字認識システム," 電子情報通信学会技術研究報告, PRMU，vol.96，no.598，pp.57–63，1997.

[98] 斎藤泰一，山田博三，山本和彦，"手書漢字の方向パターン・マッチング法による解析," 電子情報通信学会論文誌 D，vol.65，no.5，pp.550–557，1982.

[99] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," ACM Transactions on Intelligent Systems and Technology, vol.2, no.3, pp.27:1–27:27, 2011.

[100] D. Michie, "Memo functions and machine learning," Nature, vol.218, no.5136, pp.19–22, 1968.

[101] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.886–893, 2005.

[102] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," Proceedings of European Conference on Computer Vision (ECCV), pp.778–792, 2010.

[103] S. Leutenegger, M. Chli, and R.Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," Proceedings of International Conference on Computer Vision (ICCV), pp.2548–2555, 2011.

[104] 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," 電子情報通信学会論文誌 D，vol.68，no.4，pp.781–788，1985.

[105] 小川英光，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，1994.

[106] S. Cheong, S.H. Oh, and S.-Y. Lee, “Support vector machines with binary tree architecture for multi-class classification,” Neural Information Processing-Letters and Reviews, vol.2, no.3, pp.47–51, 2004.

[107] B. Fei and J. Liu, “Binary tree of SVM: A new fast multiclass training and classification algorithm,” IEEE Transactions on Neural Networks, vol.17, no.3, pp.696–704, 2006.

付 録A
SVM識別器の評価実験結果

本付録では，本文中に示せなかったクラスごとのSVM識別器の評価実験結果について示す．具体的には，表 6.1 に示したLS-Handの方向特徴を用いた実験結果および，表 6.2 に示したLS-Handの輝度特徴を用いた実験結果，表 6.3 に示したMNISTの方向特徴を用いた実験結果でのクラスごとのSVM識別器の認識率について示す．

A.1　大規模データセットでのクラスごとのSVM識別器の認識率

まず，表 6.1 に示したLS-Handの方向特徴を用いた実験結果において，クラスごとのSVM識別器の認識率を評価した結果を図 A.1 に示す．同図から，平均認識率だけではなく，クラスごとの認識率についても，事前選択の有無による認識率の低下がないことが確認できる．また，提案手法で事前選択されるパターン数と同数のパターンを学習パターンとした場合，SVM識別器の認識率が低下していることも確認できる．

次に，表 6.2 に示したLS-Handの輝度特徴を用いた実験結果において，クラスごとのSVM識別器の認識率を評価した結果を図 A.2 に示す．同図からも，平均認識率だけではなく，クラスごとの認識率についても，事前選択の有無による認識率の低下がないことが確認できる．

以上のように，平均の認識率だけでなくSVM識別器の認識率をクラスごとに解析しても提案手法を用いた学習パターンの事前選択によるSVM識別器の認識率の変化はほとんどない．よって，大規模数字画像データセットであるLS-Handを学習パターンとした場合，提案手法による学習パターンの事前選択により，各クラスのSVM識別器の認識率を低下させず学習処理を高速化できる．

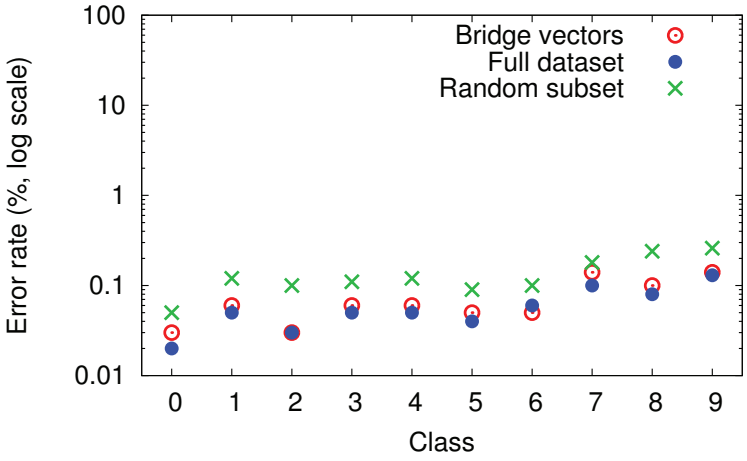


図 A.1: LS-Hand の方向特徴におけるクラスごとの認識率

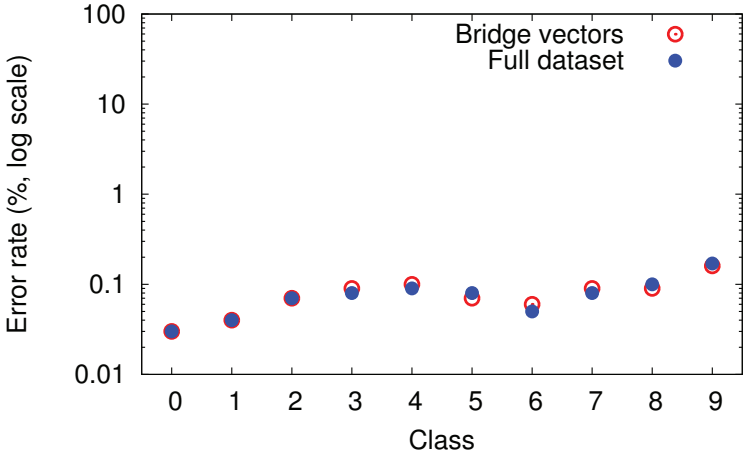


図 A.2: LS-Hand の輝度特徴におけるクラスごとの認識率

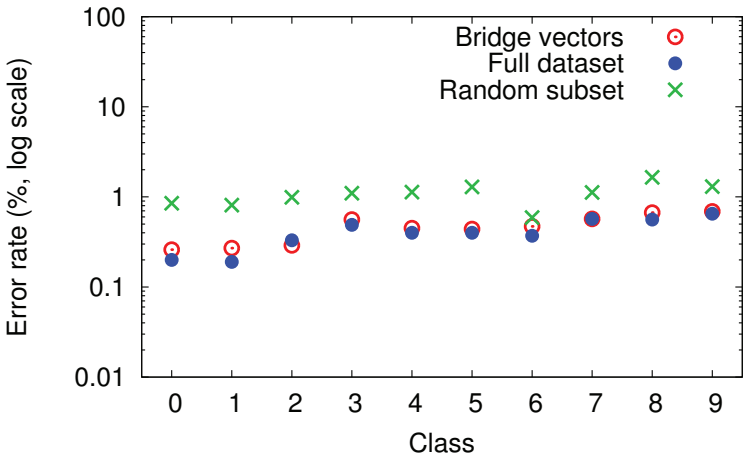


図 A.3: MNIST の方向特徴におけるクラスごとの認識率

A.2 公開データセットでのクラスごとのSVM識別器の認識率

表 6.3 に示した MNIST の方向特徴を用いた実験結果において，クラスごとの SVM 識別器の認識率を評価した結果を図 A.3 に示す．同図から，平均認識率だけではなく，クラスごとの認識率についても，事前選択の有無による認識率の低下がないことが確認できる．また，提案手法で事前選択されるパターン数と同数のパターンを学習パターンとした場合，SVM 識別器の認識率が低下していることも確認できる．よって，公開手書き数字画像データセットである MIST を学習パターンとした場合でも，提案手法による学習パターンの事前選択により，各クラスの SVM 識別器の認識率を低下させず学習処理を高速化できる．