半教師あり学習の考え方に基づく 空間の曖昧さを考慮した領域分割に関する研究

○瀧澤 重志*1

キーワード：領域分割，半教師あり学習，クラス分類，クラスタリング，矩形和楕円型領域，PBIL

1. はじめに

建築空間や街路のオープンスペースといった，都市スケールから見れば比較的小規模な空間を設計する際に，その中で発生する様々な人々の行為やイベントの発生を期待して，設計者は設計を行っていると考えられる．建築計画や都市計画の研究でも，例えば，滞留行動やオフィス内での行動分析など，行為やイベントに関する研究が盛んに行われている．こうした人々の行為やイベントには，設計者が意図するものだけでなく意図しないものも含まれる．空間が思わぬ使われ方をされることは，どのような空間でも多かれ少なかれ生じることである．こうした認識は，研究者や設計者の間で広く共有されている問題意識だと思われる．

しかし意図しない行為・イベントは必ずしも人々にとって歓迎されるものだけではない．その典型的な例が街頭犯罪である．犯罪発生と空間構成に関連があることは 1970 年頃に指摘され，今日では防犯環境設計として世界中で実践されている．その動きに追随するように，地理情報システムの発達により，犯罪と空間の関係を実証的に分析する地理的犯罪分析などの研究が広く行われている．申請者は，これまで主に京都市伏見区などを対象地域として，街頭犯罪と空間分析の関係を分析してきた [1]．

加えて近年の空間デザインは，コンピュテーショナル・デザインやアルゴリズミック・デザインなどの考え方も反映して，機能分割と空間構成が一見対応しない曖昧さを積極的に許容する空間となってきている．例えば，伊東豊雄らによるせんだいメディアテーク，石上純也による神奈川工科大学 KAIT 工房，藤本壮介の House O，平田晃久の杵屋本店など，明示的に分節化されていない空間が現代建築の世界で広まってきており，こうした空間を分析するための新たな方法が必要とされてきている．

空間分析を行う場合，実証性を高めるために統計解析手法を援用することは，今日では当然のように行われている．建築・都市計画の空間分析での統計解析の使われ方を概観すると，注目する行為やイベントが発生する場所の特徴を，クラスター分析によりタイプ分けする研究が多い．しかしクラスター分析は，判別分析などと異なり教師データ（目的変数）を使用しないので，行為・イベントが場所の違いによりなぜ発生したりしないのかを説明することはできない．この原因として，先に述べた空間の曖昧性が大きく関わっていると考えられる．分類問題では，まず空間的範囲を定め，各範囲にクラスラベルを付与して教師データを作成した上で分析を行うが，空間の曖昧性や領域形状の多様性の問題から，事前にどのように空間的範囲を定めたらよいのかについては，未解決の問題として残っている．

またこの問題は，主に計量地理学の分野で考慮されてきた空間的自己相関の問題と強く関連する．空間的自己相関とは，ある注目している空間の周辺も注目している空間と類似の空間的特徴を持っている性質である．空間データに対して，誤差項の相関を考慮しない通常の回帰分析手法を適用すると，モデルの精度が低下してしまうことが知られている．そのため，誤差項の共分散成分に空間的な制約を入れた地理的加重回帰分析などが発展してきた．しかしこの方法は，主にマクロな地理学的スケールの事象を扱うために開発されてきたため，領域を注目点からの距離だけで表現する．これは領域を単純な円（もしくは楕円）として想定することを意味しており，本研究が対象とする狭域の分析にそのまま応用するには問題がある．

以上の背景から本研究では，建築空間や街路レベルの比較的小さな空間スケールを対象として，本来曖昧な性質を有する空間上で意図せず発生する行為やイベントが，どのような空間的状況のもとで発生するかを統計的に高精度に分類・予測し，同時にそれら行為やイベントが発生する可能性が高い領域を，半教師学習 [2]という考え方を用いて自動的に決定する，新しい空間分析手法の枠組みを提示し，基礎的な分析モデルの提案・検証を行う．

既往研究 [1]において，ひったくり犯罪と街路の空間構成の関係を分析した際に，当該点だけでなく，ひったくりが潜在的に発生すると考えられる近隣部も考慮して分類を行う，本研究につながる新しい分析方法を構想した．しかしこの既往研究では，領域を線形なネットワーク空間上の距離で限定しており，さらに理論展開が十分でないことやイベント発生点の判別精度が低いなどの問題があり，本研究を行うに至っている．

2. 半教師あり学習

機械学習のアルゴリズムは教師あり学習と教師無し学習の二つにわけられる．前者で用いられるデータは正解と

なる目的変数（ラベル）を有するが，後者はそれが無い場合に用いられる．前者の典型的なタスクは回帰や分類，後者はクラスタリングである．半教師有り学習は，ラベルあり・なしの混在データから学習することで，ラベルあり／無しデータだけを用いる場合と比べて，データ作成の効率性の確保や高精度な分類を目指した方法の総称である．

半教師あり学習はさらに，半教師ありクラスタリングと半教師あり分類の二つに分けられる．半教師ありクラスタリングの場合は，いくつかのデータに must/cannot リンクと呼ばれる所属するクラスターの制約などの教師情報が与えられて，クラスタリングのヒントとするものである．一方，半教師あり分類は，データの一部にだけラベルが着いているデータを対象として，ラベル無しの事例も用いて精度の高い分類モデルを構築するのが目標である．

本研究は，データに一部だけラベルが付いているような問題を扱うため，半教師分類の考え方で問題を定式化する．

3. 問題のフレームワーク

3.1 問題設定

本研究では 2 次元の平面上での分析を扱う．3 次元の空間属性の場合は，3 次元空間で属性を計測した後，それを平面に射影して用いる．本研究で扱う平面は，本来連続的な平面を，ビットマップにより離散化して考える．この理由は，特徴量の計測は離散的にサンプリングすることが多いことや，狭域空間では連続関数による近似が難しいこと，空間のオブジェクトが増えても，ビットマップの場合は計算量が一定でデータ構造が単純なため，計算効率の向上が図りやすいなどである．

考慮する平面領域を，1 辺の長さが l の正方形（セル）で格子状に分割し，セルとその集合（対象領域全体）を $c \in C$ と表す．各セルにはその中心で測定した説明変数 $\boldsymbol{v}_c = (v_c^1, \ldots, v_c^N)$と，所属クラスを示す目的変数が割り当てられる．分類問題はイベントの発生の有無に関する 2 クラス問題とし，それぞれのクラスを P, N と表す．目的変数は $o_c \in \{\mathrm{P}, \mathrm{N}\}$ である．各セルのデータは，$(\boldsymbol{v}_c, o_c)$ として表現される．

オリジナルなデータでは，イベントセルのみに $o_c = \mathrm{P}$ のラベルが与えられており，その他のセルは不定の状態である．いま，考慮するイベントが発生した点を含むセルとその集合を $e \in E \subset C$ と表す．また，e を中心とする半径が最大で r セルの範囲に含まれる近傍領域を $R_e \in \Re \subset C$ とし，その領域内のセルにもクラス P のラベルを付与することを許す．クラス N のラベルはその他のセル $c' \in C \setminus \Re$ に付与される（図 1）．

いま，C_P，C_Nをそれぞれクラス P，N のセルの集合とすると，対象領域は，$C = C_P \cup C_N, C_P \cap C_N = \emptyset$,として重複しない二つの領域に分割される．近傍領域の集合 $\Re$ によって，C を上記二つのデータセットに分割して返す関数を

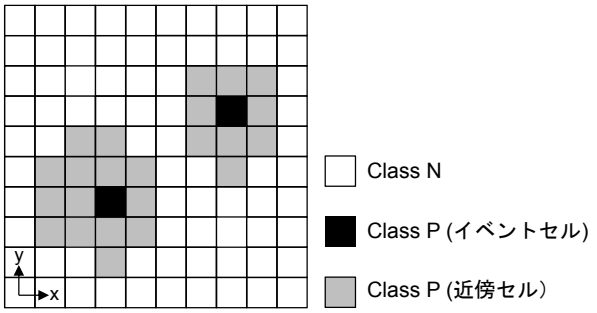

図 1　対象空間

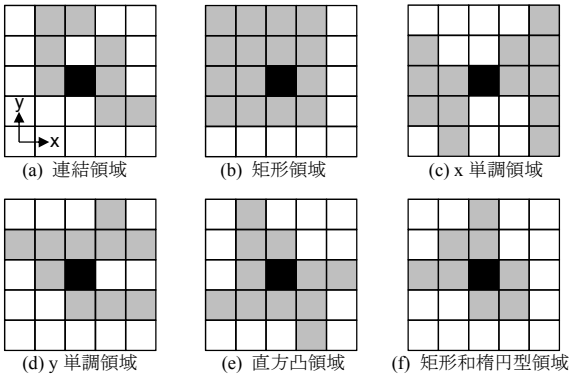

図 2　イベントセルを中心とする領域族の例

$dv(C, \Re)$ とする．また，データセットを入力して分類結果を返す分類器を，$f(\cdot)$と表す．さらに，分類結果の広い意味での誤差を返す関数を $err(\cdot)$とする．

以上の準備により，分類誤差を最小化する最適な領域 $\Re$ を発見する問題は，次式のような組み合わせ最適化として定式化できる．

$$\min_{\Re \subset C} err\left(f\left(dv(C, \Re)\right)\right), \quad \text{s.t. } R_e \in \Re \text{ が満たすべき領域制約など} \tag{1}$$

この近傍領域の制約条件を次に示す．

3.2 離散平面の領域族

ユークリッド空間 $\mathbb{R}^n$ の場合，$\mathbb{R}^n$内の開集合を U としたとき，U の任意の 2 点を U に含まれる連続曲線で結ぶことができるものを領域と呼ぶ．本研究では，当該セルの上下左右 4 近傍のセルで連結関係を定義する．図 2 は 2 次元の離散平面での領域族の例である．(a)は連結領域であり，上記の領域の定義を満たす最も一般的なものである．しかし，建築の領域としては，隙間があるようなものはまとまりが無く，加えて連結領域のパターンは，近傍の半径に対して $\mathbf{O}(2^r)$ で爆発的に増えるので現実的ではない．そこで，より限定した領域族を考える必要がある．例えば，円形領域や(b)矩形領域が考えられるが，これらは，逆に領域の形状の自由度が少なすぎる．(c)(d)の x(y)単調領域は，それぞれ x(y) 軸に平行な線との交わりが必ず連続か空となる領域である．x 単調かつ y 単調なのが(e)直方凸領域であり．領域のまとまりと境界部分のある程度の複雑さを兼ね

備えている．さらに，(f)矩形和楕円型領域は，イベントセルを端点とする矩形の集合で構成されており，その領域内はイベントセルからの可視性が保証されるので，狭域な空間分析での利用可能性が高い．

ちなみにこれらの領域の包含関係は，矩形領域⊂矩形和楕円型領域⊂直方凸領域⊂x(y)単調領域⊂連結領域となる．

3.3 目的関数

3.1 で述べたように，目的関数は分類誤差の最小化であるが，一口に分類誤差といっても様々なものがある．表 1 に分類性能を評価する際に用いられる混同行列を示す．混同行列は，実際のデータと予測されたクラスラベルで分類結果をクロス集計したものである．この情報から，最も基本的な指標である全体の正解率(*Accuracy*)が得られるが，その他に，各クラスそれぞれの正解率（*TPrate*，*TNrate*）なども定義することもできる．データセット中のクラスの分布が偏っている場合は，全体の正解率よりも，各クラスの正解率を個別に評価した方が，少数クラスの分類精度が向上する．

また，既往研究においては，イベントセルの精度を独立で評価していなかったので，場合に応じてその点も考慮する必要がある．

表 1　混同行列

		予測値	
		Positive	Negative
実測値	Positive	*TP*(件数)	*FN*
	Negative	*FP*	*TN*

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$TPrate\ (Sinsitivity) = \frac{TP}{TP + FN}$$

$$TNrate\ (Specificity) = \frac{TN}{TN + FP}$$

3.4 分類手法

線形判別関数，ロジスティック回帰，サポートベクターマシン，決定木，CAEP などの様々な分類モデルの適用が考えられる．線形性を仮定した単純なモデルの場合は，そのモデルでの誤差を最小化する領域形状を数値的に求めることができる可能性があるが，分類モデルが線形なため高い分類精度は期待できない．現代的な分類モデルの場合はモデルが複雑になるため，真の最適領域を発見するのは困難になるが，その反面で高い分類精度が期待できる．

3.5 最適化手法

問題(1)は組み合わせ最適化問題になる．上記の分類モデルと関連するが，線形判別関数などの比較的単純な分類モデルを選択した場合，最適な領域を求める問題は混合整数計画問題などで定式化できる可能性がある．しかし，現代的な分類モデルでの厳密解法は難しいと考えられ，メタヒューリスティクスなどの近似解法を用いることが基本的な方針になると思われる．

4. 実装

本研究では以下のような設定で，3 章の問題設定を実装する．

4.1 領域族

本研究では，領域族として，イベント点からの可視性が保証される矩形和楕円型領域を用いる．

4.2 目的関数

既往研究では，精度の悪いクラスの誤差だけを評価していたが，この方法では最適化に時間がかかるなどの問題が予想されるので，本実装では誤差関数として Balanced Error Rate (BER)を用いて目的関数を構成する．

$$BER = 1 - \frac{TPrate + TNrate}{2}$$

BER は各クラスの誤差率を等しい重みで評価するので，バランスのとれた誤差評価が期待できる．

4.3 分類手法

本研究は，筆者のこれまでの既往研究で用いてきた，コントラストパタンによる分類モデルの CAEP[3]を用いる．CAEP は，少数クラスを含めた分類性能の高さと，結果の可読性に優れた手法である．

4.4 最適化手法

CAEP はパターンの集合で分類を行う複雑な分類モデルなので，領域形状の組み合わせ最適化問題は近似解法となる．本研究では，メタヒューリスティクスの分布推定アルゴリズムで最もシンプルな Population-Based Incremental Learning（PBIL）[4]を用いる．PBIL は遺伝的アルゴリズム（GA）に似ているが，確率分布を進化させる点が GA と異なっている．コード化が単純なことや，多くの問題で標準的な GA よりもよい精度の解を発見できる．

PBIL を本問題に適用するために，遺伝子コードの定義を行う．図 3 は，イベントセルが一つの場合の $r = 3$ の遺伝子コードの対応関係の模式図である．セルの外周上下についている番号が，確率/遺伝子ベクトルのインデックスを示している．このコード化は上下別に y 方向のセルの高さを，x 座標の昇順で定義したものである．例えば，図 4 に示すようにインデックスが 1 のセルの場合は，領域が無い場合も含めて 4 つの高さが考えられる．遺伝子情報によってセルの高さが決定されると，そのセルとイベントセルを対角に持つ一つの矩形領域が確定する．

5. ケーススタディ

5.1 設定

4 章で提案した実装のケーススタディを行う．図 5 に示すような xy 方向のサイズがそれぞれ 21 セルの平面の中央

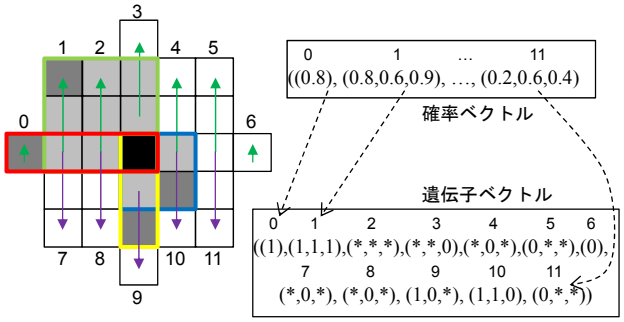

図 3　r=3 の場合の近傍領域と遺伝子コードの対応関係．*は{0,1}のワイルドカードを示す．

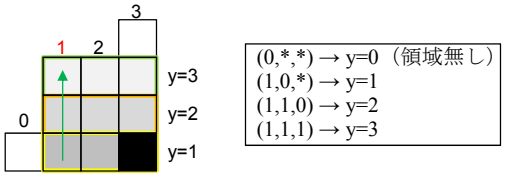

図 4　遺伝子ベクトルのインデックスが 1 のセルにおける y 方向の高さと遺伝子コードの対応関係

にイベントセルが一つだけ存在する状況を例とする．近傍の最大半径は r =6 とする．この平面に，a1～a3 までの 3 種類の属性が，図 5 のように分布しているとする．a1n と a2n はもとの a1 と a2 の分布にノイズを入れたものである．これらの属性を表 2 のように組み合わせた 2 種類の説明変数のセットを作成し，分類を行う．PBIL のパラメータは，論文[3]のデフォルト値である *learnRate* = 0.1，*negLearnRate* = 0.075，*mutProb* = 0.02，mutShift = 0.05，N = 100 を用い，*ITER_COUNT* のみ 100 世代に短くしている．また，CAEP のパラメータは，*min_sup* = 0.1, *max_dim* = 3, *min_gr* = 3.0 とした．

5.2 結果

4 章の方法により得られた最適近傍領域とそれらの分類精度を図 6 に示す．説明変数にノイズを含まない Case 1 の場合は，完全な分類に成功している．Case2 は完全な分類はできないが，精度は 9 割を超えている．Case 2 の領域形状は複雑であり，ある程度複雑な領域形状を考慮することの利点が示されている．

6. まとめ

本研究では，建築や街路レベルの比較的小さな空間上で発生するイベントを，高精度に分類・予測するための近傍領域の抽出手法の枠組みを提案し，実装と人工データでの検証を行い，よい結果を得た．今後は実データでの検証に進む予定である．

謝辞

本研究は，科学研究費補助金基盤研究(C)（25420633）の補助を得て行われました．

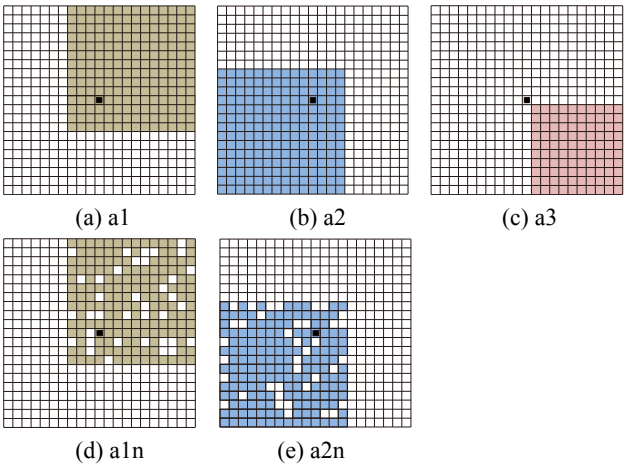

図 5　説明変数の分布

表 2　説明変数の組み合わせ

	a1	a1n	a2	a2n	a3
Case 1	+		+		+
Case 2		+		+	+

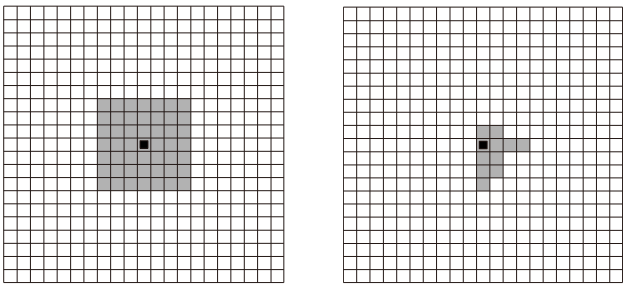

図 6　得られた最適近傍領域

表 3　最適近傍領域での分類精度

	BER	*Accuracy*	*TPrate*	*TNrate*
Case 1	0.0	1.0	1.0	1.0
Case 2	0.066	0.957	0.909	0.958

［参考文献］

1) A. Takizawa, Emerging Pattern Based Street Crime Analysis - Street Level Spatial Analysis of Crime Location Associated with Built Environment in Fushimi Ward, Kyoto City -, Journal of Architecture and Planning (Transactions of AIJ), 78(686), pp.957-967, 2013.
2) X. Zhu and A. Goldberg, Introduction to Semi-supervised Learning, Morgan & Claypool Publishers, 2009.
3) G. Dong, X. Zhang X, L. Wong and J. Li, CAEP: Classification by aggregating emerging patterns, Proc. of the 2nd International Conference on Discovery Science, pp.30-42, 1999.
4) S. Baluja, Population-Based Incremental Learning: A Method for Integrating Genetic Search Based Function Optimization and Competitive Learning, Technical Report, Pittsburgh, PA: Carnegie Mellon University, 1994.

＊1　大阪市立大学大学院工学研究科共通分野 准教授 博士（工）